

Program and Rules

Redes de Comunicações II

**Licenciatura em
Engenharia de Computadores e Informática
DETI-UA**



Professors

- Prof. Paulo Salvador (theory classes)
 - ◆ Email: salvador@ua.pt
 - ◆ Web: <http://www.av.it.pt/salvador>
- Prof. Daniel Corujo (practice classes)
 - ◆ Email: dcorujo@ua.pt
- Prof. Amaro de Sousa (practice classes)
 - ◆ Email: asou@ua.pt

Prof. Ayman Radwan (practice classes)

- ◆ Email: aradwan@ua.pt



UC Informations

- All materials, documents and software will be available on eLearning.ua.pt (Moodle).
 - ◆ Subjected to weekly updates.

Flexible office hours

- ◆ Email (to discuss any topic or schedule a meeting).
- ◆ Discord: Invite <https://discord.gg/bPPpKy5>
 - ◆ **Change your nick to your real name** (First and Last names).
 - ◆ **Ask RC2 student role.**
 - ◆ Only after you will have access to the course contents.



Objectives and Outcomes

- The objective of the course is to present students with:
 - ◆ Essential concepts in computer networks;
 - ◆ Identifying the fundamentals applied to the control and transport of data.
- It is intended that at the end the students should:
 - ◆ Have an understanding of the underlying fundamentals of communication networks;
 - ◆ Understand new technologies and concepts of communication networks;
 - ◆ Be able to use their knowledge to respond to current changes in communication networks.



Program

- Local Area Networks (LAN)
 - ◆ Virtual LAN: purpose, implementation, segmentation models, Layer 2 interconnection (802.1Q and VXLAN) and Layer 3 interconnection.
 - ◆ Spanning Tree-protocol(s).
- Network Design Models
 - ◆ Types of topology. Redundancy and resilience requirements. Hierarchical design model.
- IP Routing
 - ◆ Unicast Routing: static, dynamic and police based routing.
 - ◆ Internal Routing protocols (RIPv1, RIPv2, OSPF, ISIS).
 - ◆ Internet general AS architecture and core networks. Inter-AS routing (MP-BGP).
 - ◆ Multicast Routing protocols (IGMP, MLD, PIM-DM, PIM-SM, PIM-SSM).
- Overlay Networks: IP-IP and GRE IP tunnels.
- Core Networks: SDH and DWDM. MPLS.
- (Other) Access Networks
 - ◆ CATV/HFC (DOCSIS), SDH/SONET/GPON,
 - ◆ Celular networks (4G/5G).
- Communication models: client-server and P2P.
- VoIP Service: SIP and WebRTC.
- Sensor Networks: BT, Zigbee, LoRA, NB-IoT.



Evaluation

- Final Grade = 50% * Theory Grade + 50% * Practice Grade
 - ◆ There are no minimum grade for any component.
 - ◆ Theory grade
 - ✚ 1 Final Exam (100%) in the exam season;
 - ✚ and/or 1 Exam in “repeat exam” season;
 - ✚ The best grade is considered.
 - ◆ Practice Grade
 - ✚ 2 multiple choice tests (25%+25%)
 - During practice classes;
 - First test – May 9th, 11th or 13th;
 - Second test – June 17th, 20th or 22nd.
 - ✚ “Repeat exam” season
 - One single test with all topics.
 - The best grade is considered.



Planning (tentative)

Semana	Teórica (2F-1.5h)	Prática (3h)	Prática (2F-2h)	Prática (4F-2h)	Prática (6F-2h)
07/Mar	Program and rules. Local Area Networks (LAN): Virtual LAN: purpose, implementation, segmentation models, Layer 2 interconnection (802.1Q and VXLAN) and Layer 3 interconnection.	TP1: Trabalho GNS3 com SWL3.	TP1	TP1	TP1
14/Mar	Spanning Tree-protocol(s).	TP2: VLAN and Spanning-Tree Protocol	TP2	TP2	TP2
21/Mar	Network Design Models: Types of topology. Redundancy and resilience requirements. Hierarchical design model.	TP2: VLAN and Spanning-Tree Protocol	TP2	TP2	TP2
28/Mar	IP Unicast Routing: static, dynamic and police based routing. Internal Routing protocols (RIPv1, RIPv2, OSPF, ISIS).	TP3: Dynamic Routing	TP2/TP3	TP2	TP2
04/Apr	IP Unicast Routing: static, dynamic and police based routing. Internal Routing protocols (RIPv1, RIPv2, OSPF, ISIS).	TP3: Dynamic Routing	TP3	TP3	TP3
11/Apr	Quinta-Feira	TP3: Dynamic Routing + (Optional) Policy Based Routing	Quinta-Feira	Páscoa	TP3 (terça-feira)
18/Apr	Páscoa	TP4: IPv4 tunnels. IPv6 over IPv4 tunneling.	Páscoa	TP3	TP3
25/Apr	Feriado/Sem.Académica	Sem.Académica	Feriado/Sem.Académica	Sem.Académica	Sem.Académica
02/May	Overlay Networks: IP-IP and GRE IP tunnels.	TP4: IPv4 tunnels. IPv6 over IPv4 tunneling.	TP3	TP3	TP4
09/May	IP Multicast Routing: protocols (IGMP, MLD, PIM-DM, PIM-SM, PIM-SSM).	TESTE PRÁTICO	TESTE PRÁTICO	TESTE PRÁTICO	TESTE PRÁTICO
16/May	Internet general AS architecture and core networks. Inter-AS routing. MP-BGP.	TP5: MPBGP	TP4	TP4	TP5
23/May	Core Networks: SDH and DWDM. MPLS.	TP5: MPBGP	TP5	TP5	TP5
30/May	Communication models: client-server and P2P. VoIP Service: SIP and WebRTC.	TP6: Voip (SIP)	TP5	TP5	TP6
06/Jun	Sensor Networks: BT, Zigbee, LoRA, NB-IoT.	TP6: Voip (SIP)	TP6	TP6	Feriado
13/Jun	(Other) Access Networks: CATV/HFC (DOCSIS), SDH/SONET/GPON, Cellular networks (4G/5G).	TP7: (optional) MPLS	TP6/TP7	TP6/TP7	TESTE PRÁTICO
20/Jun	Revisions.	TESTE PRÁTICO	TESTE PRÁTICO	TESTE PRÁTICO	

TP1+TP2+TP3

TP4+TP5+TP6

TP4+TP5+TP6



Bibliography

- Theoretical classes slides.
- A Practical Approach to Corporate Networks Engineering, António Nogueira, Paulo Salvador, River Publishers, ISBN-13: 978-8792982094, 2013.
- Computer Networks: A Systems Approach, Larry Peterson, Bruce Davie, ISBN-13: 978-0128182000, 6th Edition, 2021.
- Computer Networking: a Top-Down Approach, Kurose J., Ross K., 7th edition, Addison Wesley, ISBN-13: 978-9332585492, 2017
- Designing for Cisco Network Service Architectures (ARCH), Marwan Al-shawi, Andre Laurent, Cisco Press, 4th edition, ISBN-13: 978-1587144622, 2016.
- MPLS in the SDN Era: Interoperable Scenarios to Make Networks Scale to New Services, Antonio Sanchez Monge, Krzysztof Grzegorz Szarkowicz, O'Reilly Media; 1st edition, ISBN-13: 978-1491905456, 2016.
- Packet Guide to Voice over IP: A system administrator's guide to VoIP technologies, Bruce Hartpence, O'Reilly Media; 1st edition, ISBN-13: 78-1449339678, 2013.
- Guide to Wireless Communications, 3rd Edition, Jorge Olenewa, 4th edition, ISBN-13: 978-1305958531, 2016.
- TCP/IP Teoria e Prática, Fernandes B., Bernardes M., FCA, 2012 (em português).
- Engenharia de Redes Informáticas, Edmundo Monteiro, Fernando Boavida, FCA, ISBN-13: 978-972-722-694-8, 10^a Edição Atualizada e Aumentada, 2011 (em português).



Local Area Networks (LAN)

Redes de Comunicações II

**Licenciatura em
Engenharia de Computadores e Informática
DETI-UA**

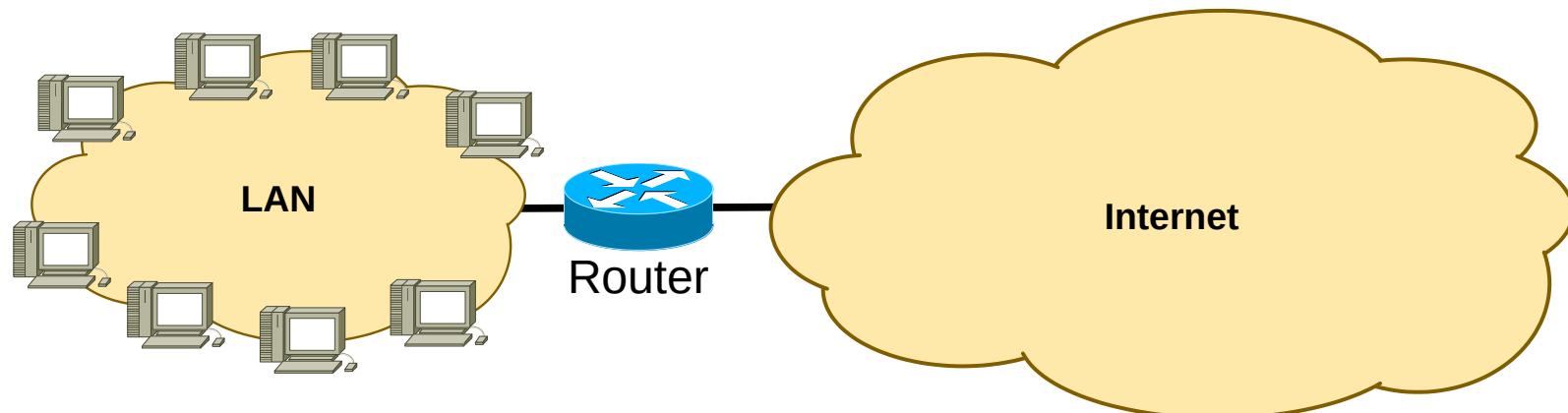


universidade de aveiro

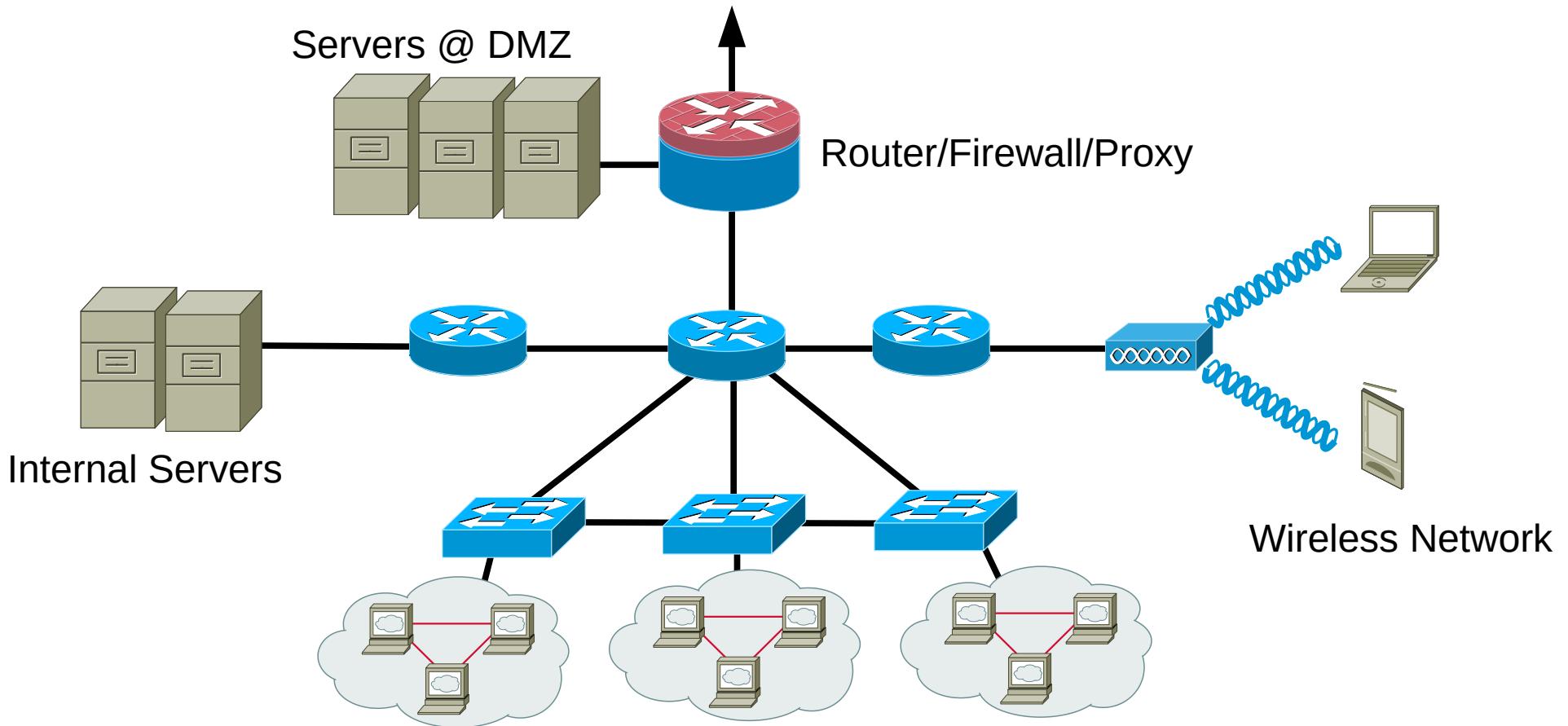
deti.ua.pt

Local Area Network (LAN)

- Is a computer network within a small geographical area.
 - ◆ Home, school, room, office building or group of buildings.
- Is composed of inter-connected hosts capable of accessing and sharing data, network resources and Internet access.
 - ◆ Host refers generically to a PC, server, or any other terminal.
- Technologies
 - ◆ Current: Ethernet, 802.11 (Wi-Fi)
 - Legacy: Token Ring, FDDI, ...

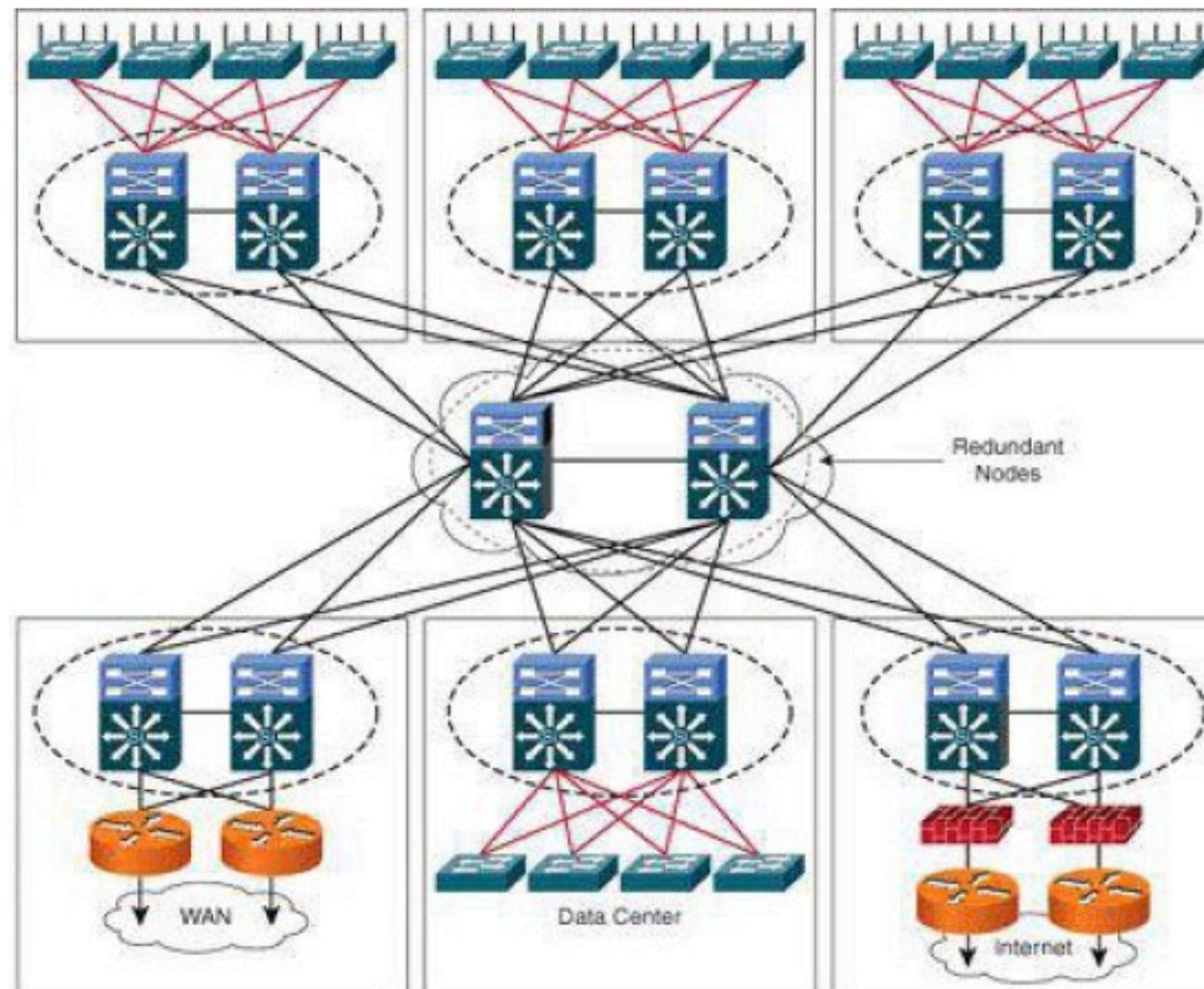


Corporate Access Networks: Small LAN



Corporate Access Networks: Medium/Large LAN

- Hierarchical architecture

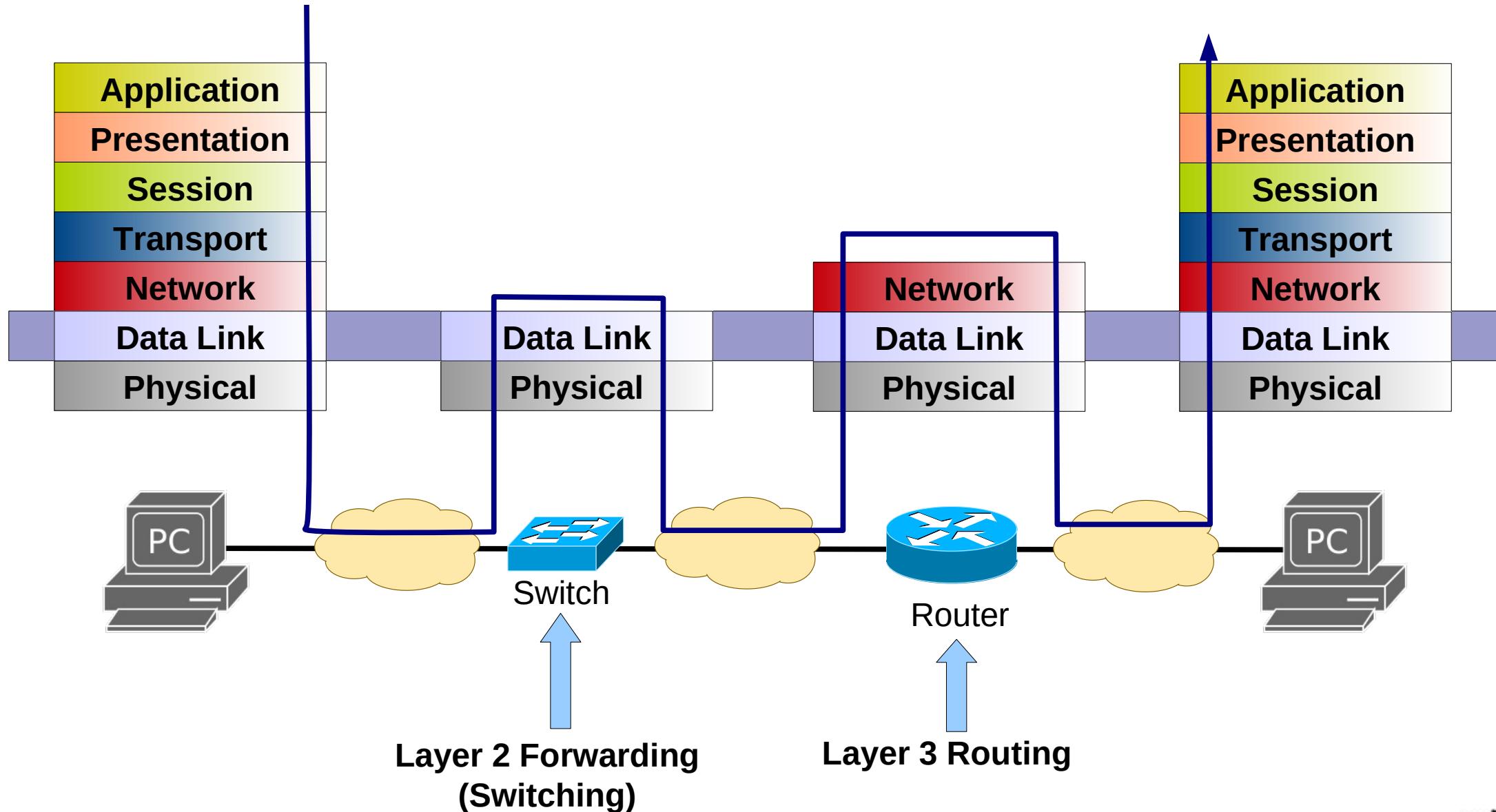


Ethernet (802.3)

- Most successful LAN technology.
- Invented at Xerox Palo Alto Research Center (PARC).
- Xerox, DEC and Intel defined in 1978 the standard for Ethernet 10Mbps.
- Uses “Carrier Sense/Multiple Access” with “Collision Detect” (CSMA/CD)
 - ◆ Carrier Sense: hosts can perceive if the communication channel is being used.
 - ◆ Multiple Access: multiple hosts can access simultaneously
 - ◆ Collision Detect: host “listen” the communication channel while transmitting to detect transmission collisions.
 - ◆ Collision: multiple physical signals overlapping and interfering with each other.



Ethernet based LAN



Network Devices

- Switch

- ◆ OSI Layer 2 inter-connection
- ◆ Implements VLAN
- ◆ Spanning-tree based routing
 - ◆ STP, RSTP, MSTP
- ◆ Wireless Access Points

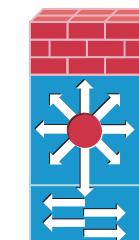
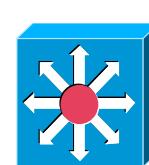


- Router

- ◆ OSI Layer 3 inter-connection
- ◆ Have extra functionalities like QoS, Security, VPN gateway, network monitoring, etc...

- L3 Switch

- ◆ Switch+Router
- ◆ Low-end and mid-end range routing functionalities are limited
- ◆ High-end have full routing functionalities
- ◆ Many have dedicated L2 routing hardware



- Router with switching modules

- ◆ L3 Switch with full routing capabilities

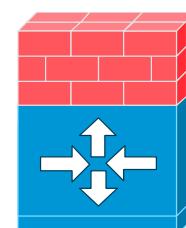
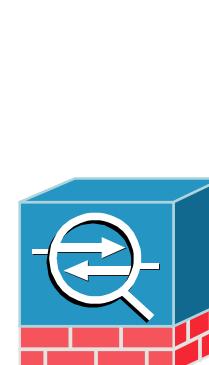
- Load-Balancer

- Firewall

- IDS/IPS (Intrusion Detection/Prevention System)

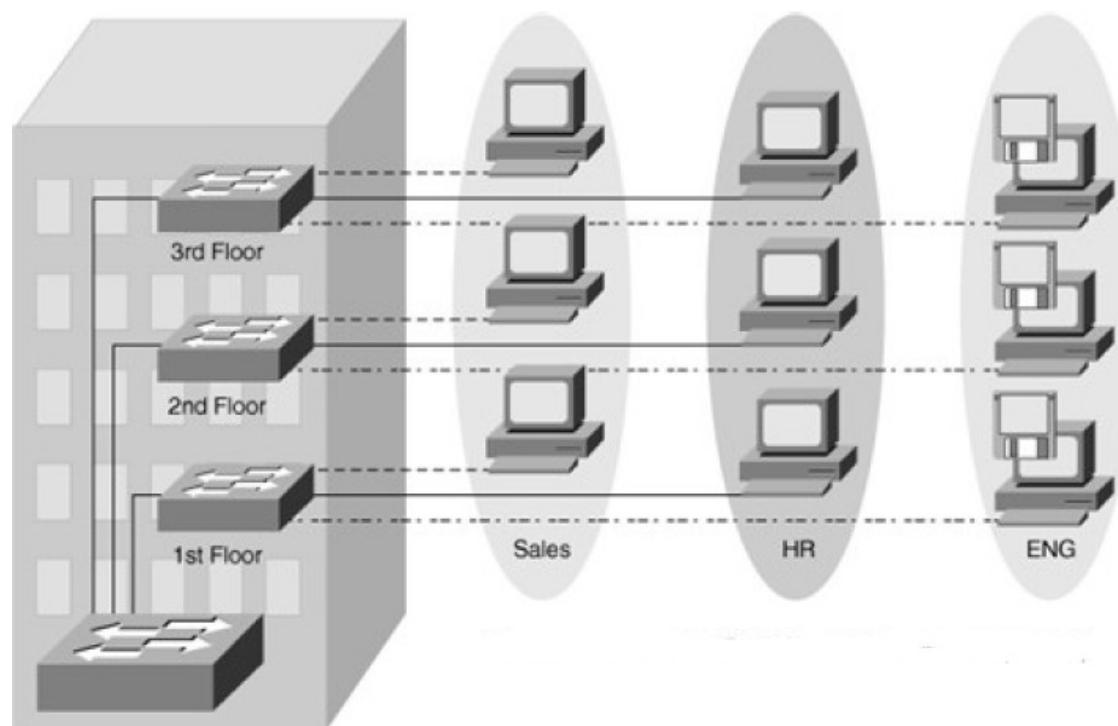
- VPN Gateway/Server

- Services proxy



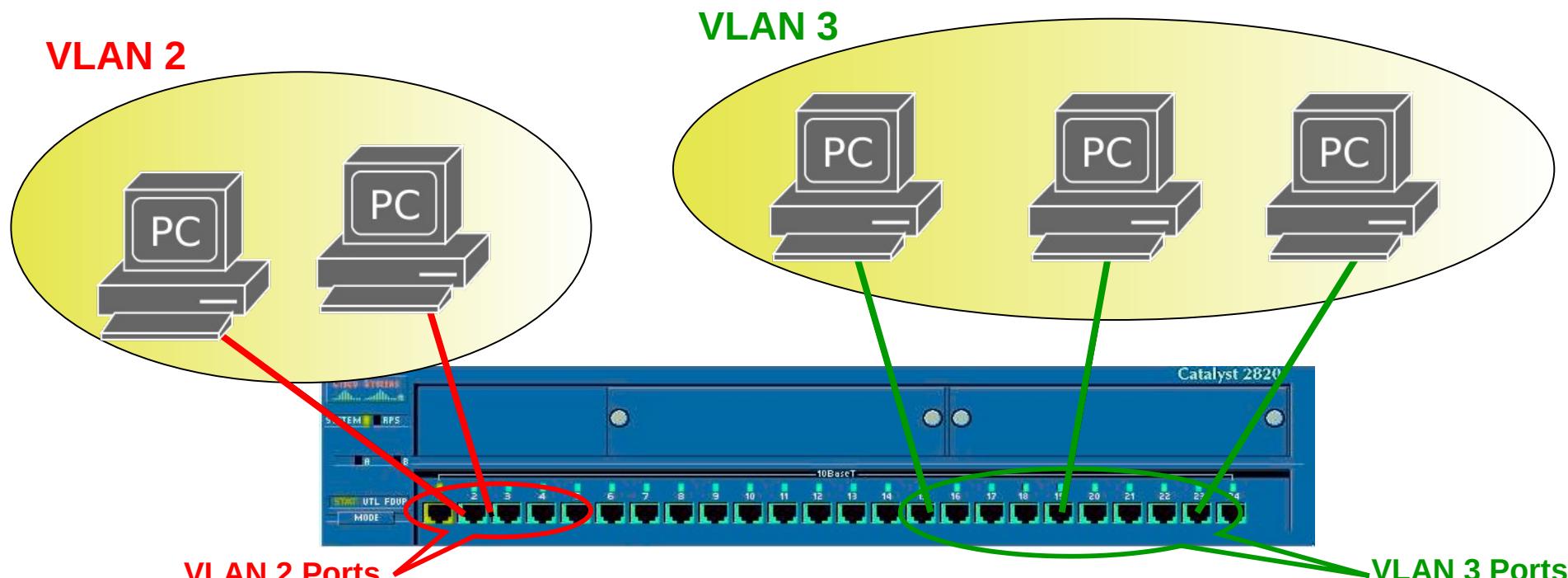
Virtual LAN (VLAN)

- A Virtual LAN (VLAN) is a group of hosts/users with a common set of requirements or characteristics in the same broadcast domain.
 - ◆ Independent of their physical location.
- Solves the scalability problems of large networks.
 - ◆ By breaking a single broadcast domain into several smaller broadcast domains.
 - ◆ Allows better/simpler network administration and security deployment.
- Hosts in different VLAN do not communicate by Layer 2.
 - ◆ Its communications are done at Layer 3 (with IP routing).



Defining Host VLAN

- The VLAN to which a host belongs depends only on the port of the switch.
 - ◆ Configured only in the switch.
 - ◆ Example: If port 1 is configured as VLAN 2, and port 20 is configured as VLAN 3:
 - ◆ If host is connected to port 1 it is on VLAN 2,
 - ◆ If host is connected to port 20 it is on VLAN 3.
- VLAN 1 is usually reserved to network administration.
 - ◆ Used to access configurations remotely via IP.



Example – VLAN

Pings sent by 10.0.0.1



```
# ping 10.0.0.2
```

```
Pinging 10.0.0.2 with 32 bytes of data:
```

```
Reply from 10.0.0.2: bytes=32 time<10ms TTL=128
```

```
Ping statistics for 10.0.0.2:
```

```
  Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
  Approximate round trip times in milli-seconds:
    Minimum = 0ms, Maximum = 0ms, Average = 0ms
```

```
# ping 10.0.0.5
```

```
Pinging 10.0.0.5 with 32 bytes of data:
```

```
Request timed out.
Request timed out.
Request timed out.
Request timed out.
```

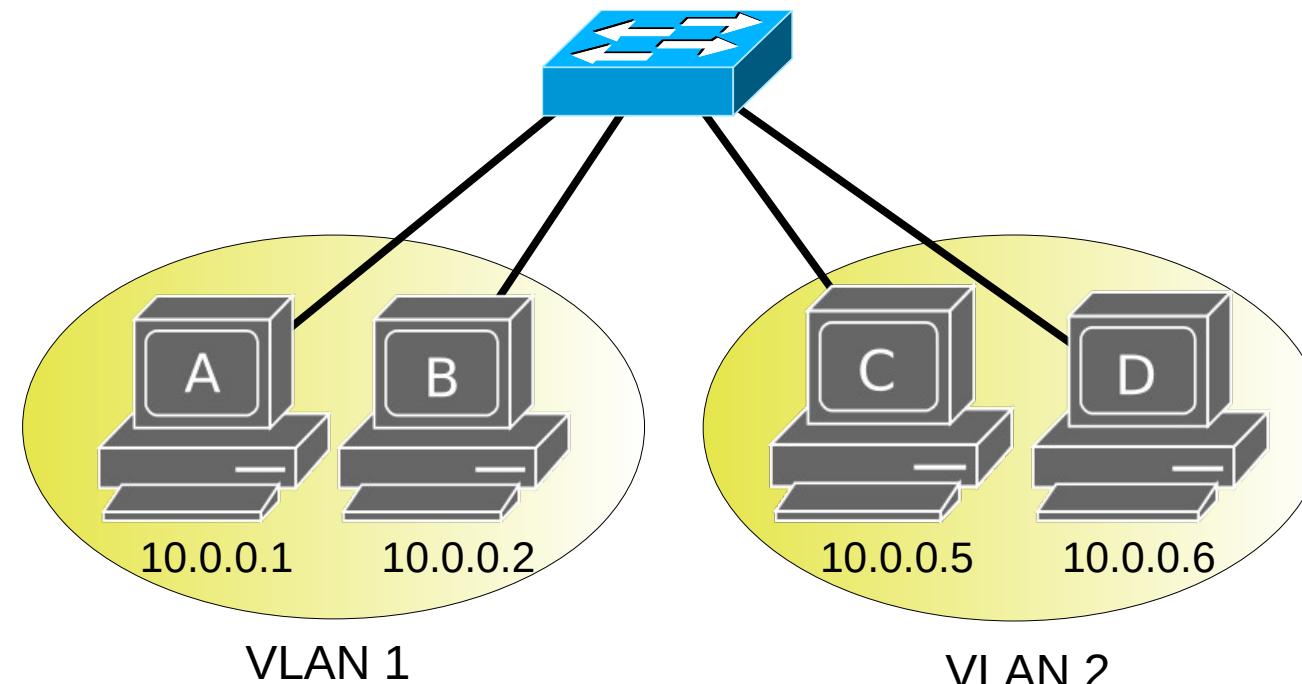
```
Ping statistics for 10.0.0.5:
```

```
  Packets: Sent = 4, Received = 0, Lost = 4 (100% loss),
  Approximate round trip times in milli-seconds:
    Minimum = 0ms, Maximum = 0ms, Average = 0ms
```

```
# ping 10.0.0.6
```

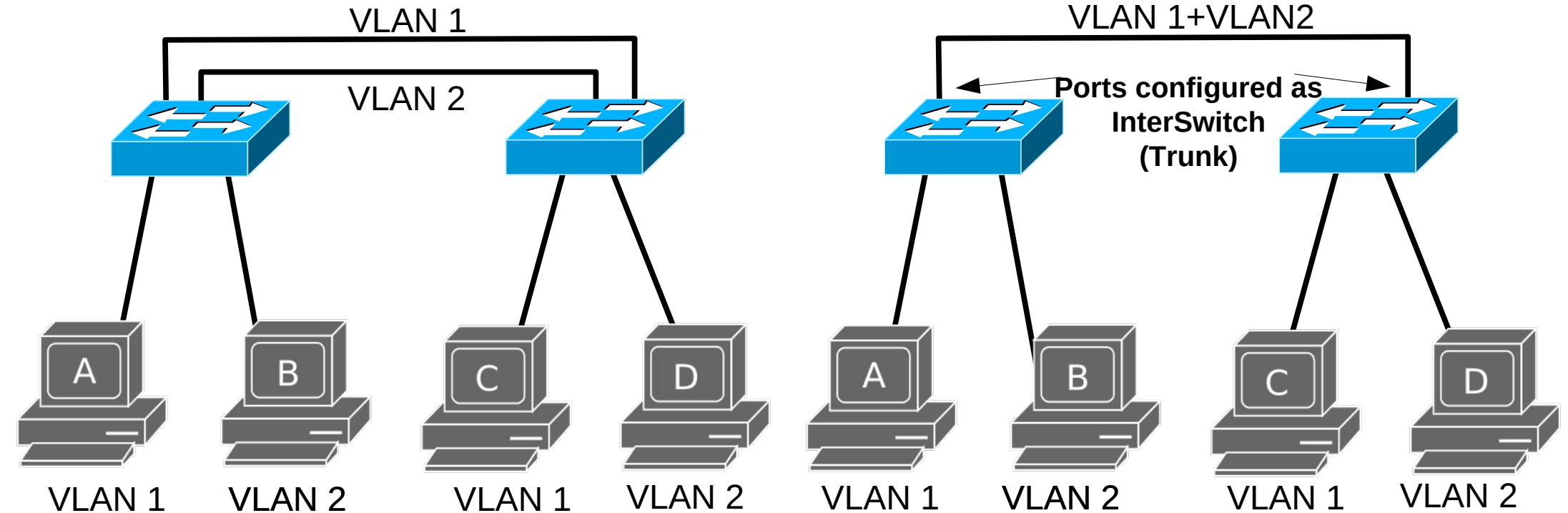
```
Pinging 10.0.0.6 with 32 bytes of data:
```

```
Request timed out.
Request timed out.
Request timed out.
Request timed out.
```



Interconnection of Switches

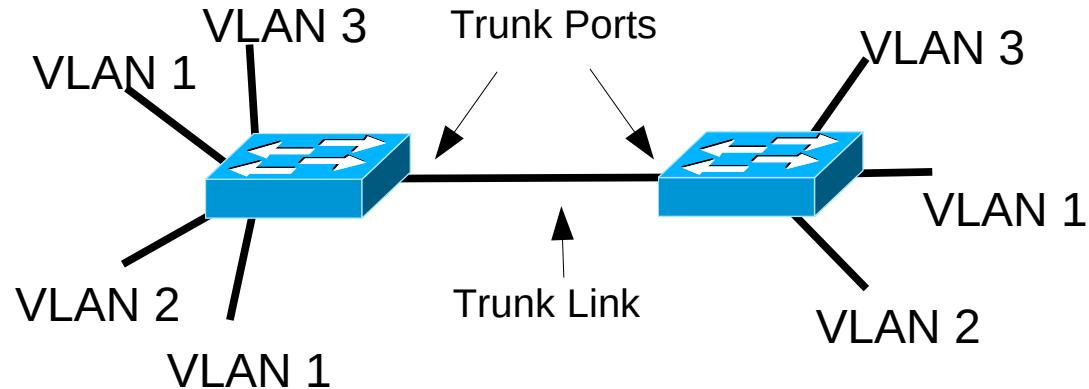
- Physical link per VLAN
 - With a single physical link.
 - Using InterSwitch/Trunk port(s).



- Using a single physical link requires a mechanism to differentiate frames from different VLAN.
 - ◆ Frames must have a tagged
 - ✚ Added when forwarding to a trunk port.
 - ✚ Read and removed when receiving a frame from a trunk port



IEEE802.1Q Standard



Ethernet frame without a VLAN tag

6	6	2		
destination	source	type	data	

Ethernet frame with a VLAN tag

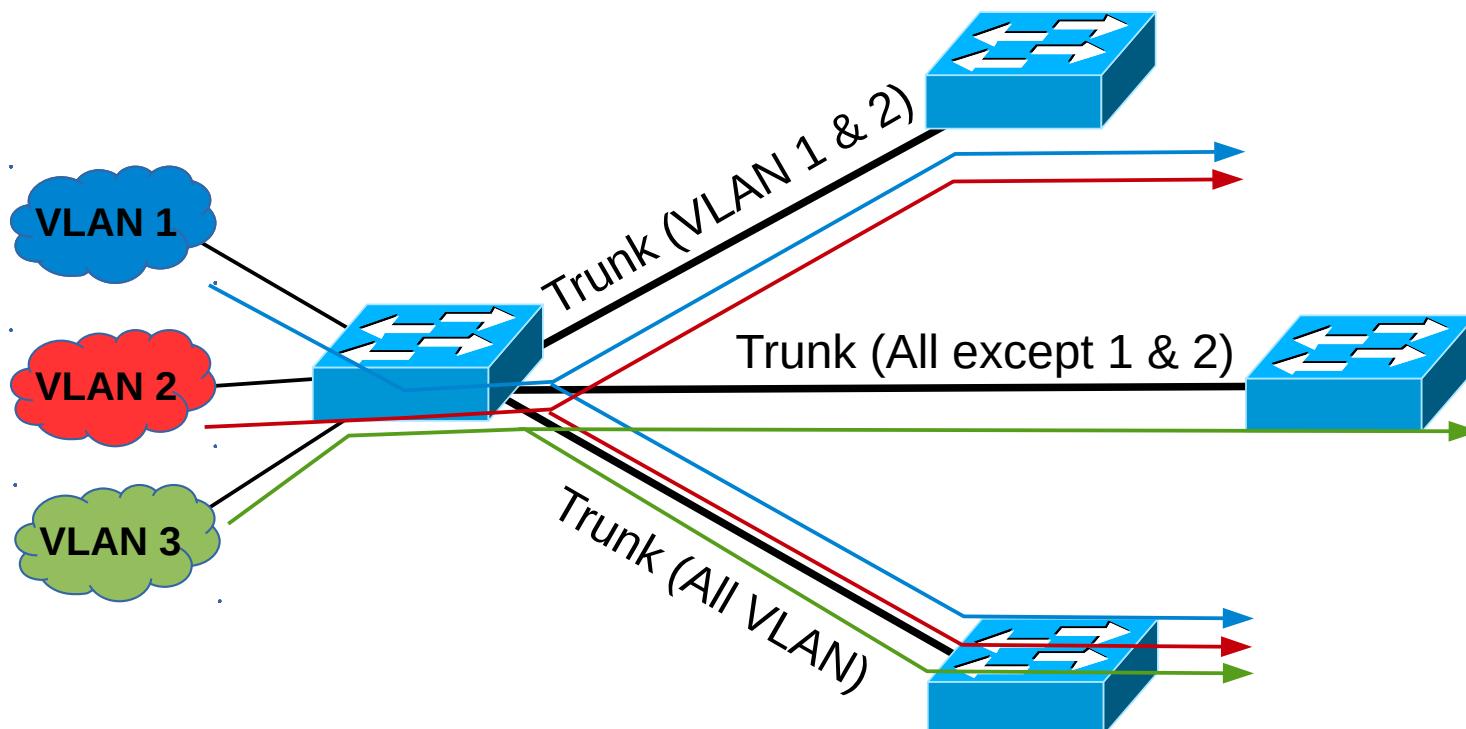
6	6	4	2		
destination	source	TAG	type	data	
		16bits	3bits	1bit	12bits
		8100h	priority	CFI	VLAN ID

- Priority: Traffic relative priority according to standard 802.1q (0 to 7 values).
- CFI: Used to guarantee compatibility with older technologies (always zero in Ethernet).
- VLAN ID: VLAN identifier.



Trunk Links

- The physical link between two Trunk ports is called a Trunk link.
- A trunk carries traffic for multiple VLANs using IEEE 802.1Q.
 - ◆ Inter-Switch Link (ISL) encapsulation is an alternative but it getting obsolete.
- Trunks may transport all VLAN or only some!

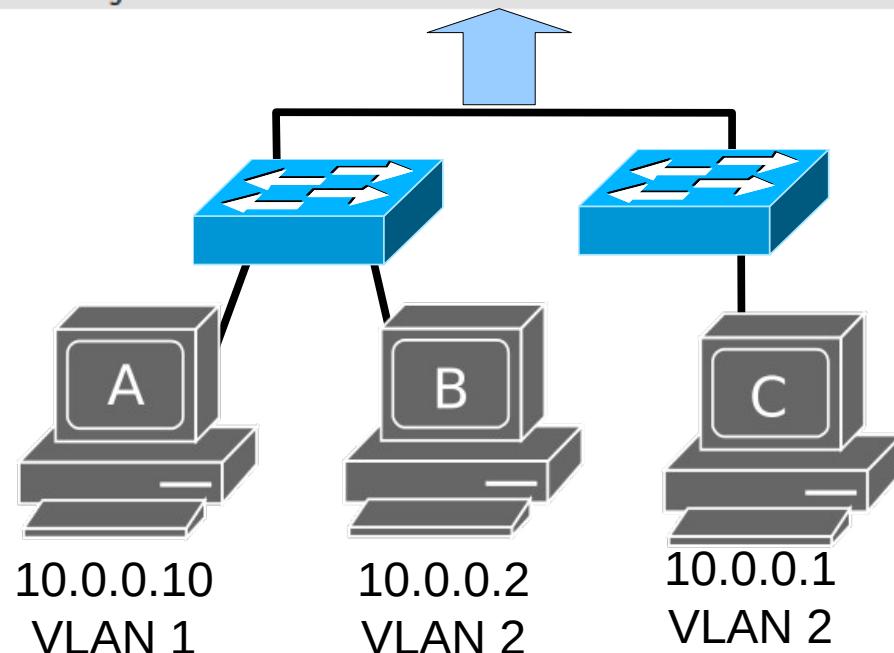


Example – InterSwitch/Trunk Ports

Filter: icmp				Expression...	Clear	Apply
No..	Time	Source	Destination	Protocol	Info	
23	11.535990	10.0.0.2	10.0.0.1	ICMP	Echo (ping) request	
24	11.536995	10.0.0.1	10.0.0.2	ICMP	Echo (ping) reply	
27	12.538443	10.0.0.2	10.0.0.1	ICMP	Echo (ping) request	
28	12.539186	10.0.0.1	10.0.0.2	ICMP	Echo (ping) reply	

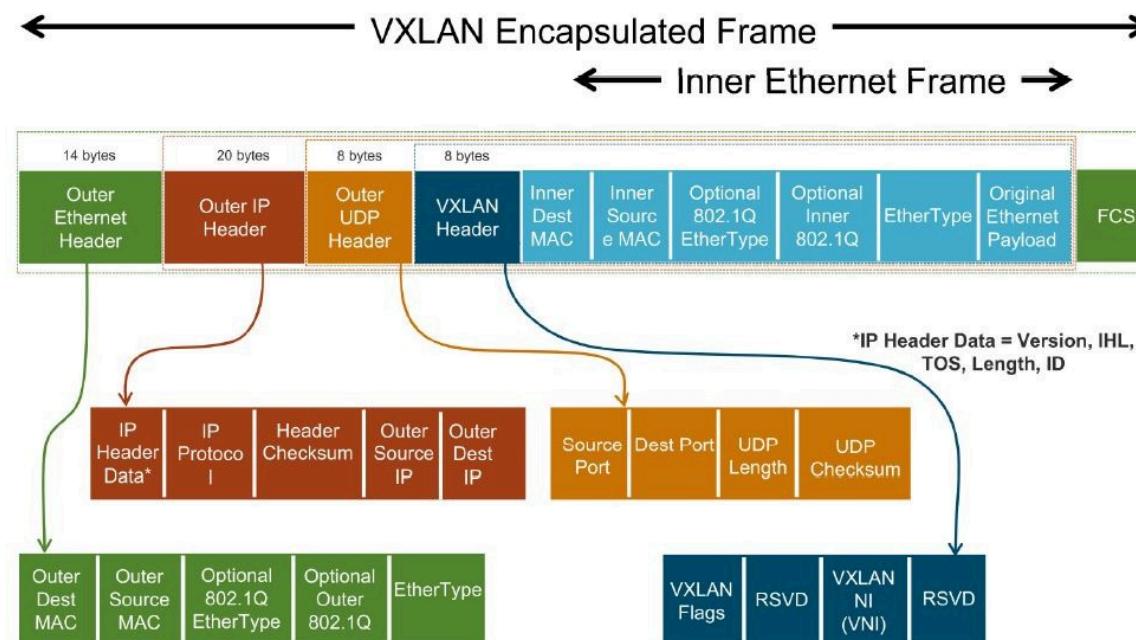
Frame 23 (102 bytes on wire, 102 bytes captured)
Ethernet II, Src: 00:aa:00:53:7c:00 (00:aa:00:53:7c:00), Dst: 00:aa:00:fa:67:00 (00:aa:00:fa:67:00)
802.1Q Virtual LAN, PRI: 0, CFI: 0, ID: 2
000. = Priority: 0
...0 = CFI: 0
.... 0000 0000 0010 = ID: 2
Type: IP (0x0800)
Internet Protocol, Src: 10.0.0.2 (10.0.0.2), Dst: 10.0.0.1 (10.0.0.1)
Internet Control Message Protocol

ID:2 == VLAN 2



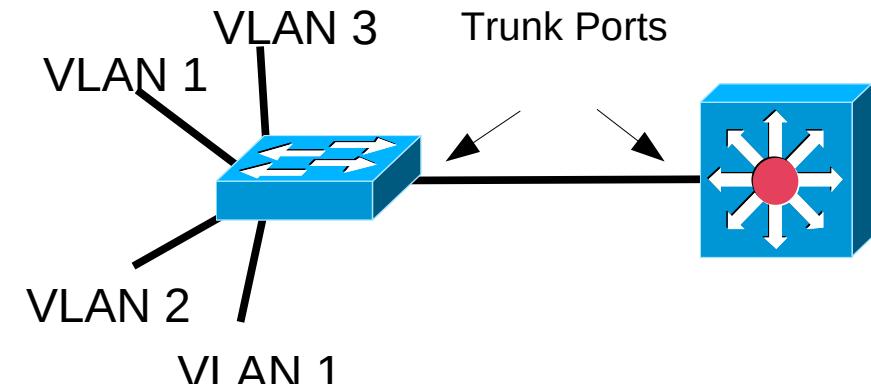
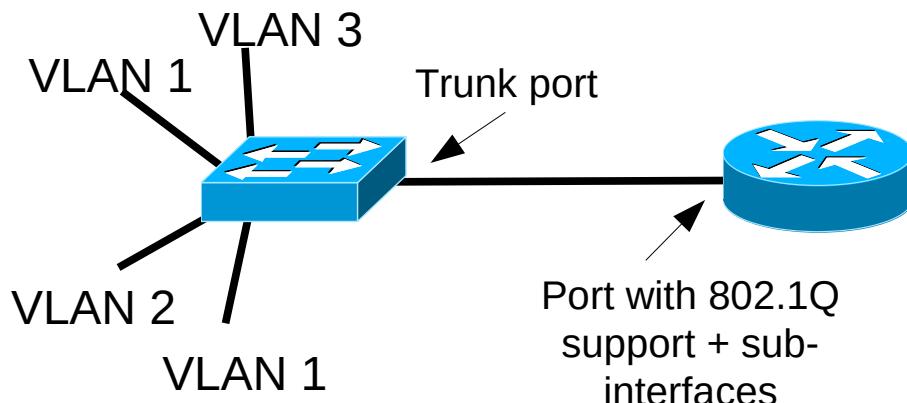
Virtual Extensible LAN (VXLAN)

- Alternative/Complement to 802.1Q in Layer3 Switches.
- Encapsulates OSI Layer 2 Ethernet frames within Layer 4 UDP/IP datagrams .
 - ◆ Default port 4789.
- VLAN may be additionally identified by a VNI field with 24 bits.
 - ◆ 802.1Q tag only as 12 bits.
 - ◆ Allows for a very large number of VLAN.
- Usually used when connecting remote VLAN (connected only via IP) in Datacenter and Cloud scenarios.

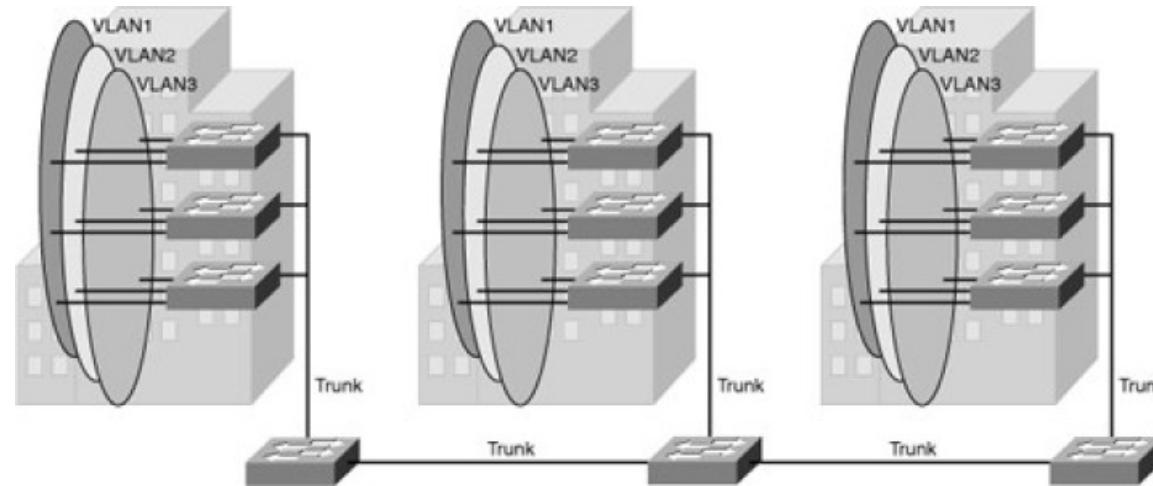


IP Connection between VLANs

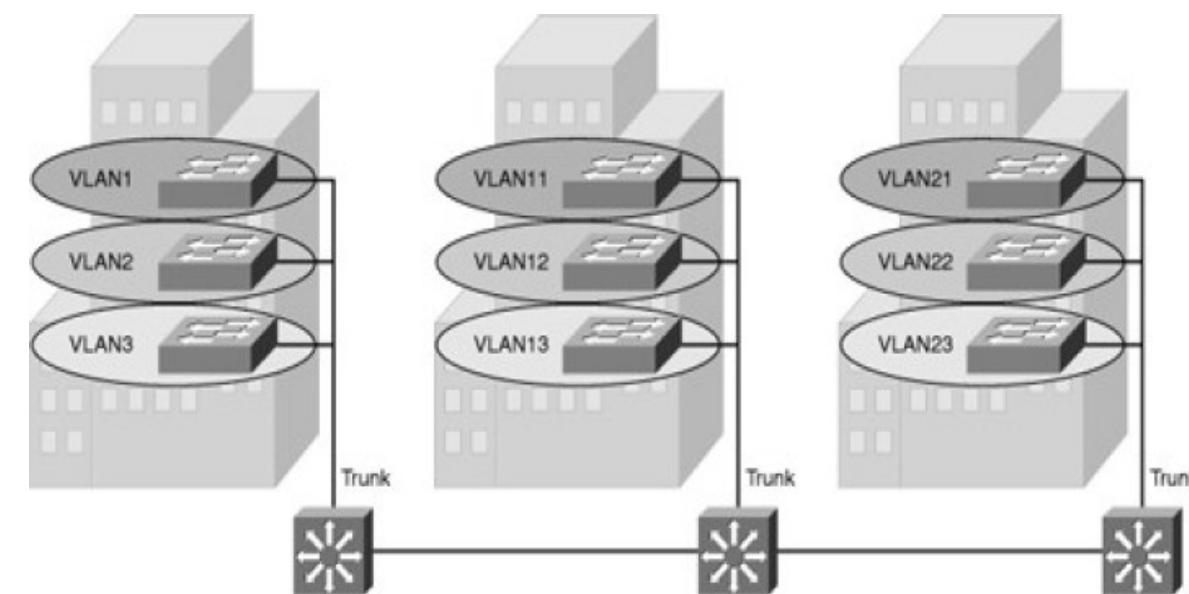
- To communicate between different VLAN it is required to use Layer 3 (IP Routing).
- Common solutions:
 - ◆ A router with support to 802.1Q,
 - ◆ Connecting the physical router interface to a Trunk port.
 - ◆ The router's physical interface is sub-divided in sub-interfaces (one for each VLAN).
 - ◆ The IP gateway for a VLAN host is the IP address of the respective sub-interface in the Router.
 - ◆ A Layer 3 switch,
 - ◆ Connecting both switches (L3 and L2) using Trunk ports.
 - ◆ Each VLAN is mapped to a virtual Layer 3 interface.
 - ◆ The IP gateway for a VLAN host is the IP address of the respective virtual interface in the L3 switch.



VLAN Segmentation Models



- **End-to-End VLAN**
 - ◆ VLANs are associated with switch ports widely dispersed over the network



- **Local VLAN**
 - ◆ Local VLANs are generally confined to a wiring closet.



VLAN Segmentation Purpose

- Joint in the same logical network services/terminals/users with same traffic/security/QoS policies.
 - ◆ Each VLAN must have an unique IP (sub-)network.
 - ◆ May have more than one IP (sub-)network.
 - ✚ Including IPv4 public and IPv4 private networks.
 - ✚ And, IPv6 networks.
- Neighbor (local) VLANs with similar traffic/security/QoS policies should have IP (sub-)networks that can be summarized/aggregated.
 - ◆ E.g.: VLAN of VoIP phones in Building 1 (VLAN 21: 200.0.0.0/24)
 - ◆ VLAN of VoIP phones in Building 2 (VLAN 22: 200.0.1.0/24)
 - ◆ Summarized/aggregated address of VLAN21+VLAN22: 200.0.0.0/23.



VLAN Segmentation (examples)

- Local VLANs

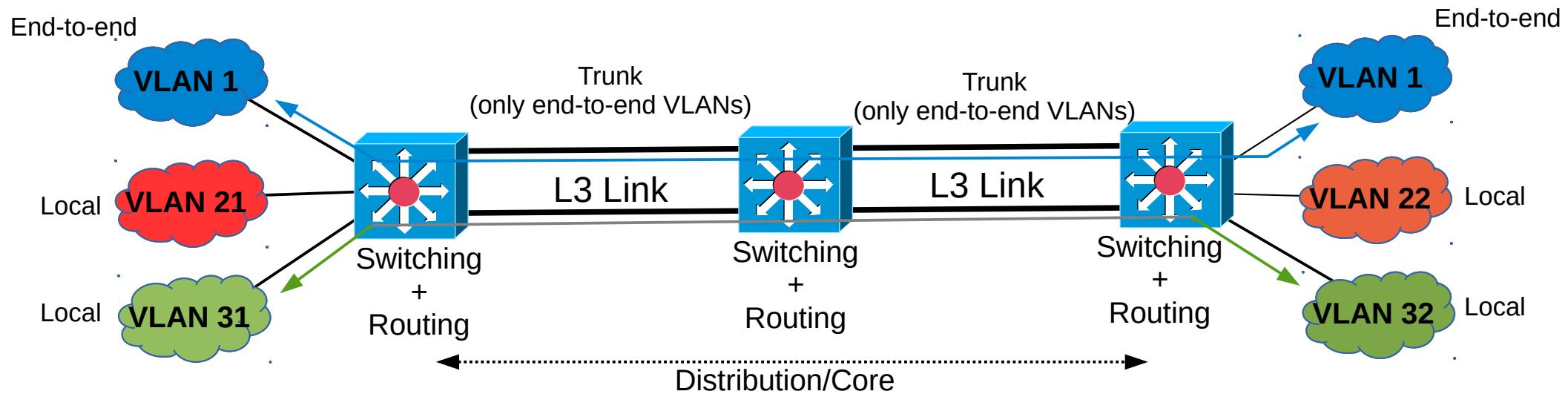
- Per service/function
 - VoIP phones, Video conference, printers, cameras, PCs, servers, ...
- Per user role
 - Engineers I, engineers II, technicians, administrators, ...
- Per location
 - Building I, floor 4, right wing, etc...
- Mixture of service/function, role, location
 - e.g.: VLAN of VoIP phones, of the Engineers in Building I.

- End-to-end VLANs

- Services/roles that have a global scope within the network.
- Wireless network
 - Same IP network (same IP address) independently of location.
 - To avoid IP changes when moving from location to location.
- Administration VLAN (optional)
 - VLAN used by the network administrator to remotely access network equipments.
 - Same administrator of (all) equipments independent of location.



Inter-(V)LAN Traffic

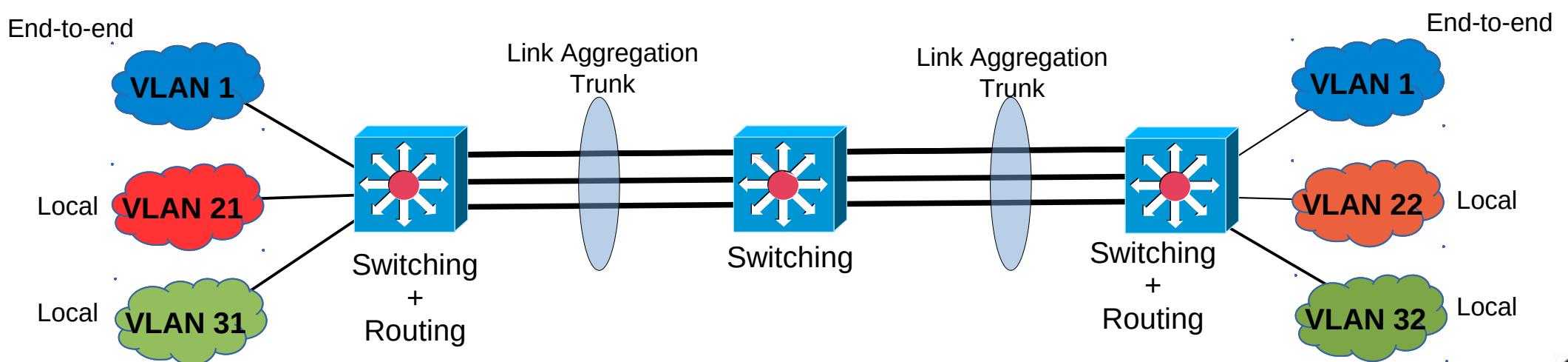


- End-to-end VLANs traffic **should be switched** over the Distribution/Core layers
 - ◆ Using a trunk (for end-to-end VLANs only).
- Local VLANs traffic **should be routed** over the Distribution/Core layers
 - ◆ Using standard layer 3 Links.
 - ◆ Using IP routing.
 - ◆ Exchange the routing information only through the L3 links
 - ◆ End-to-end VLAN should be passive interfaces for the routing processes.
 - Routes are not exchanged → Traffic is not routed!

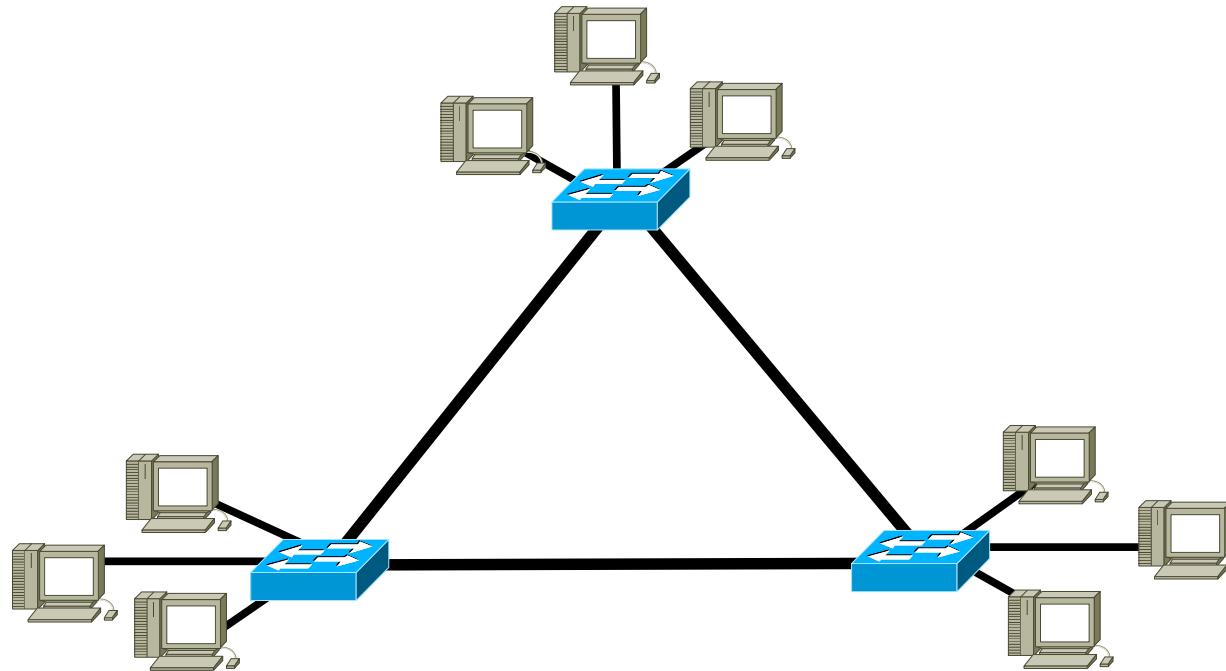


Ethernet Link Aggregation

- The throughput/speed of one connection link may not be enough to fulfill the requirements.
- Multiple Ethernet links may aggregated, provide a seamless trunk connection with N times the single throughput/speed of one link.
- Ethernet frames are “load-balanced” between all available physical links.



Redundant Layer 2 Network



- Objective: Allow the network for dynamically recover from network failures.
- Problem: Link redundancy creates Layer 2 loops. Causes the collapse of communications when MAC frames with broadcast address are sent by any host due to infinite frame flooding.

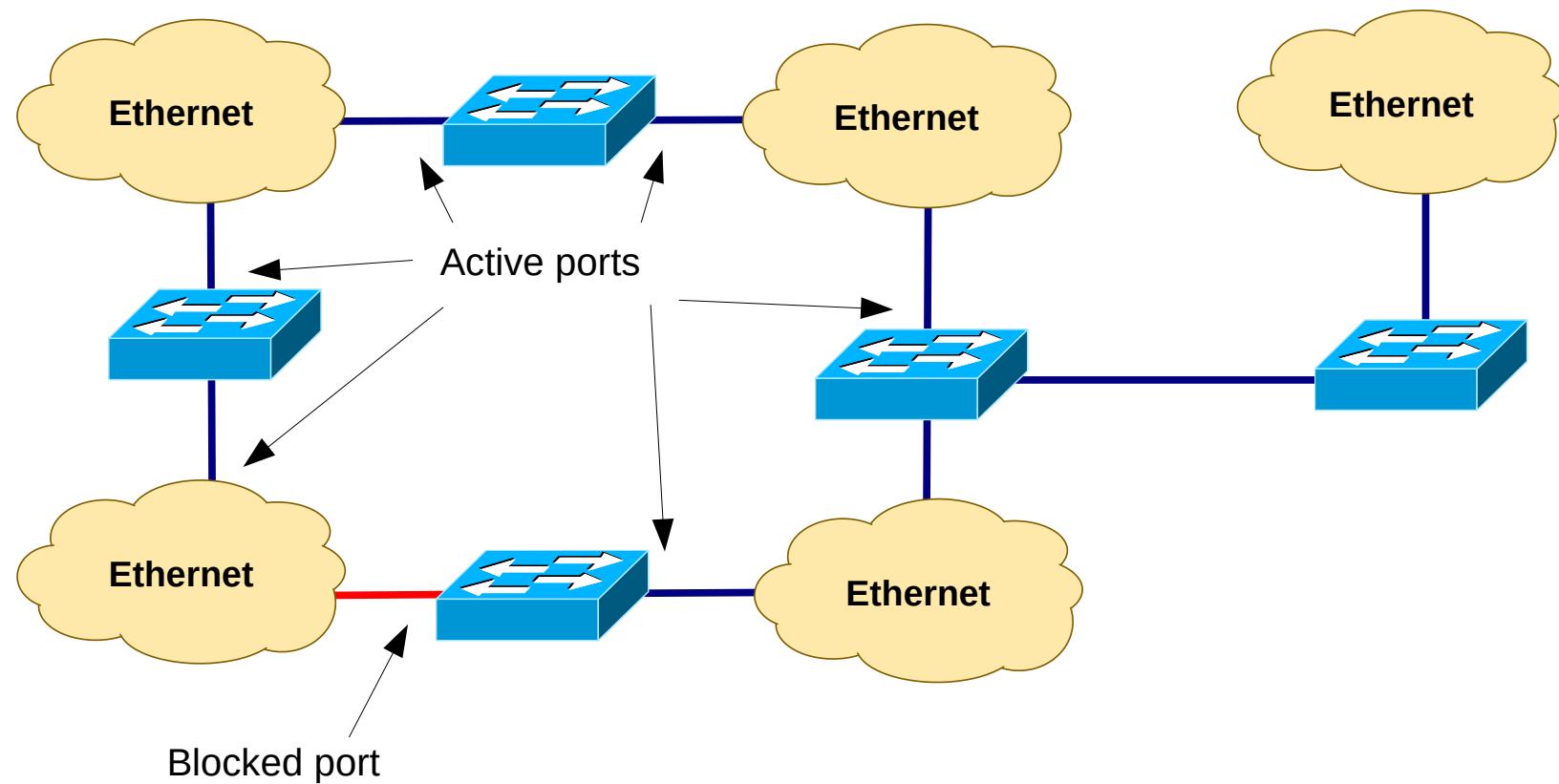


Spanning Tree Protocol (SPT)

- STP enables the network to deterministically block ports and provide a loop-free topology in a network with redundant links.
- There are several STP Standards and Features:
 - ◆ STP is the original IEEE 802.1D version (802.1D-1998) that provides a loop-free topology in a network with redundant links.
 - ◆ RSTP, or IEEE 802.1W, is an evolution of STP that provides faster convergence of STP.
 - ◆ Multiple Spanning Tree (MST) is an IEEE standard. MST maps multiple VLANs into the same spanning-tree instance.
 - ◆ Per VLAN Spanning Tree Plus (PVST+) is a Cisco enhancement of STP that provides a separate 802.1D spanning-tree instance for each VLAN configured in the network.
 - ◆ RPVST+ is a Cisco enhancement of RSTP that uses PVST+. It provides a separate instance of 802.1W per VLAN.

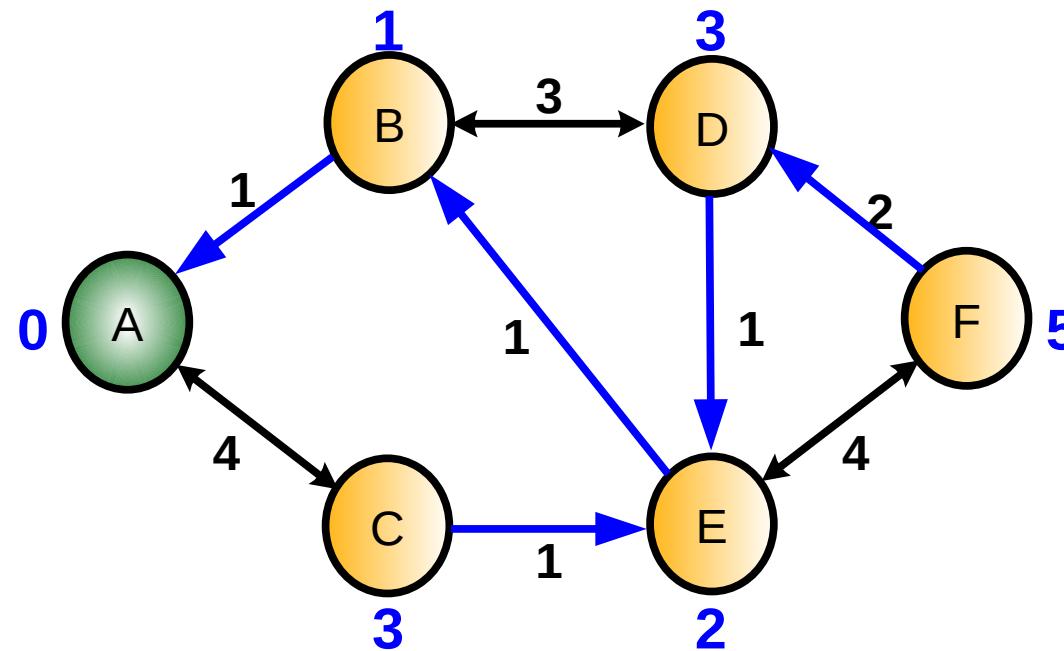


Spanning-Tree

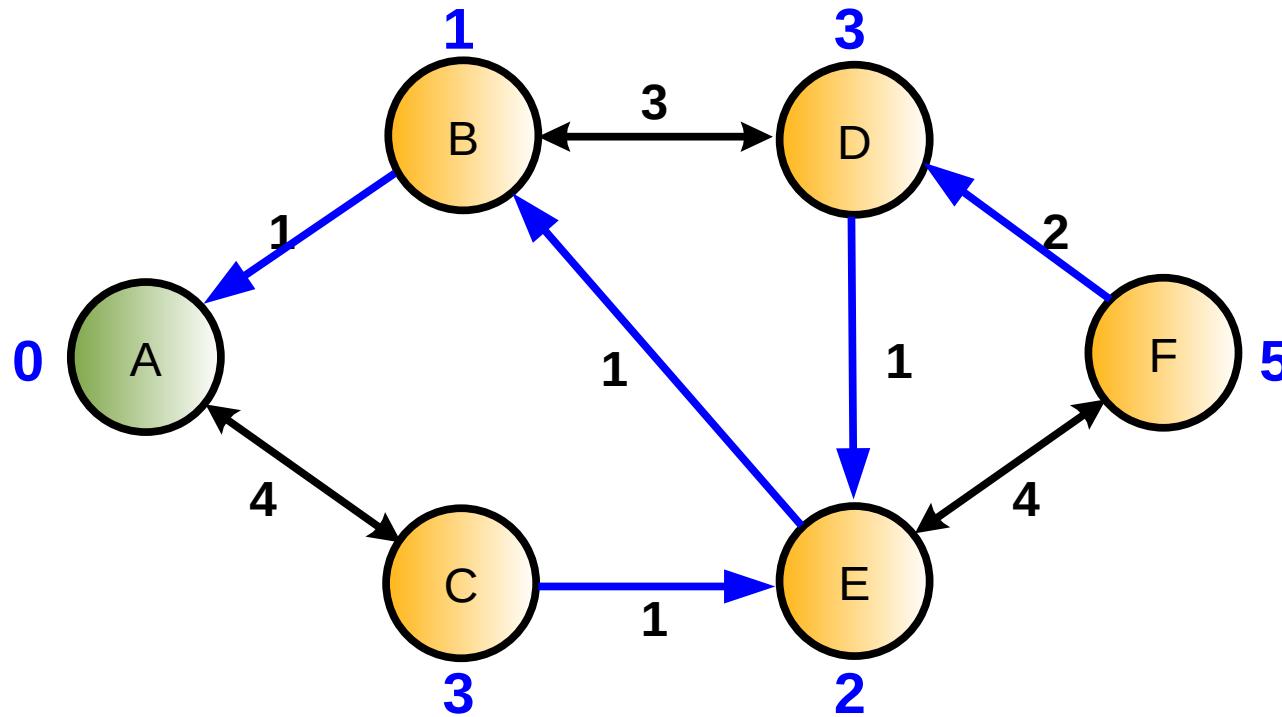


Routing based on Spanning Trees

- It is chosen an origin/root node.
- All nodes use the **Bellman-Ford Distributed and Asynchronous Algorithm** to calculate the neighbored node (and respective path cost) that provide the smallest cost to the origin/root node.
- The set of links used by all nodes to provide the shortest paths to the origin/root node is called the **Spanning Tree**.
- It is required a criteria to solve ties.



Bellman Equations



- When link cost are not negative, then:

Shortest path from one node X to node A

=

Cost of the link from that node X to the node that follows it in the shortest path to A

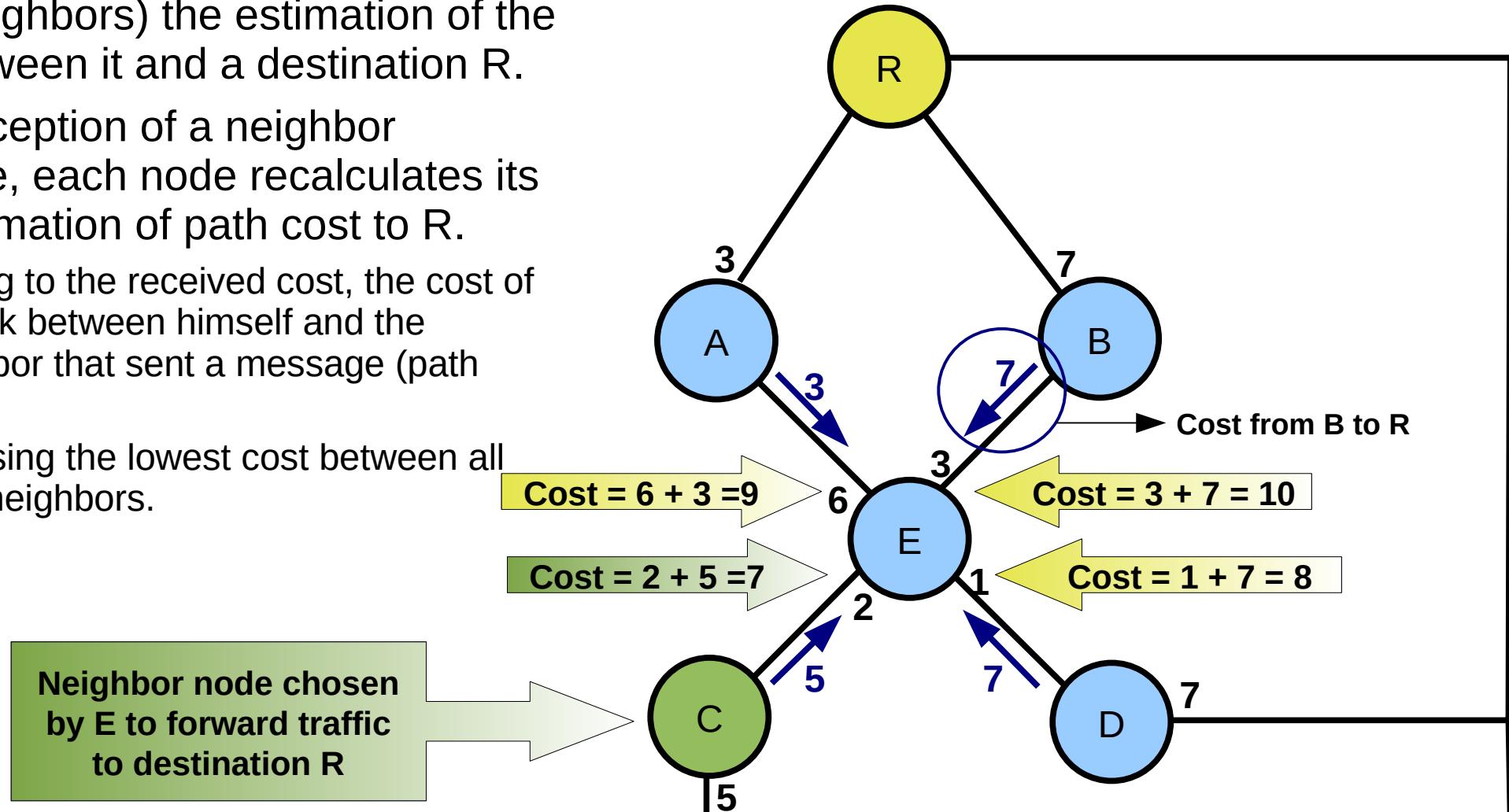
+

Shortest path from that node to node A

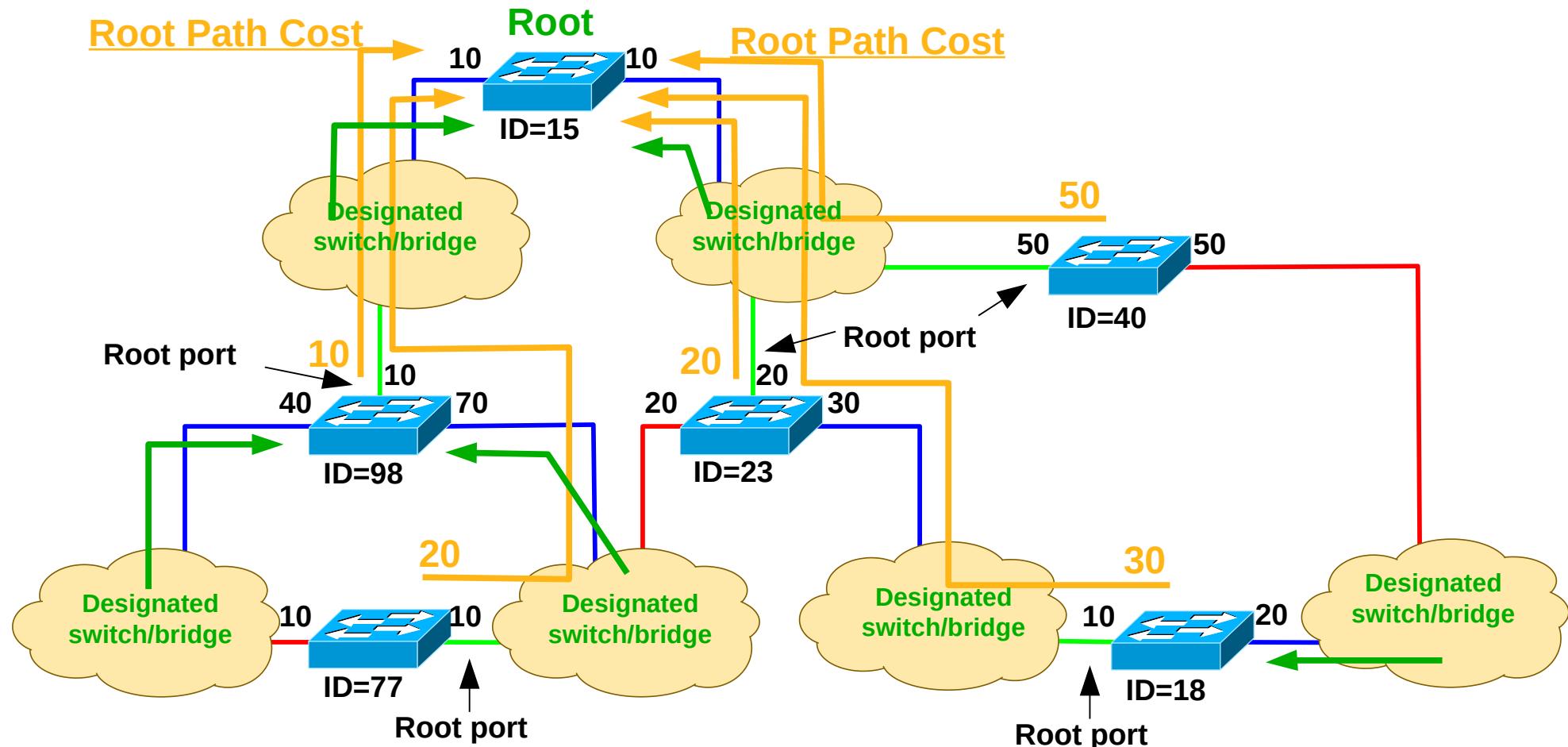


Bellman-Ford Distributed and Asynchronous Algorithm

- Each node transmits periodically (to all its neighbors) the estimation of the cost between it and a destination R.
- Upon reception of a neighbor message, each node recalculates its own estimation of path cost to R.
 - ◆ Adding to the received cost, the cost of the link between himself and the neighbor that sent a message (path cost).
 - ◆ Choosing the lowest cost between all links/neighbors.

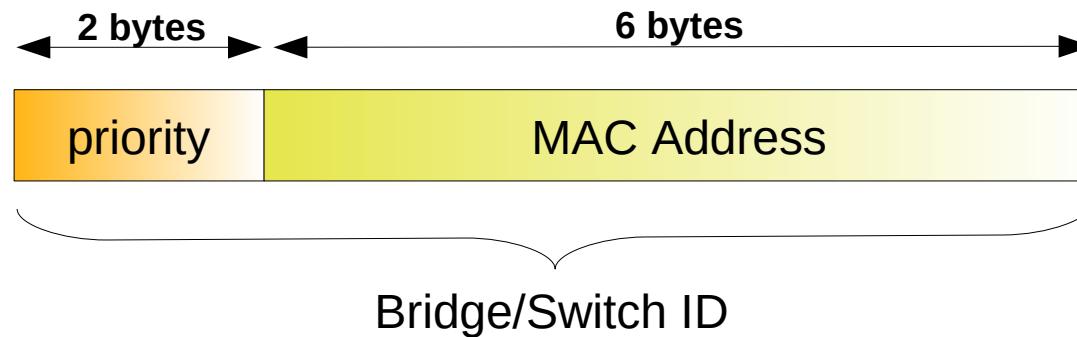


Spanning Tree Basic Concepts (1)



Spanning Tree Basic Concepts (2)

- Bridge/Switch ID – each switch is identified by an 8 bytes identifier based on:
 - ◆ 2 **Priority** bytes, defined by configuration.
 - ◆ 6 bytes (one of the **MAC Address** of the switch, or any other unique 48 bit sequence).
 - ◆ Priority has precedence over the 6 bytes sequence (usually MAC address).

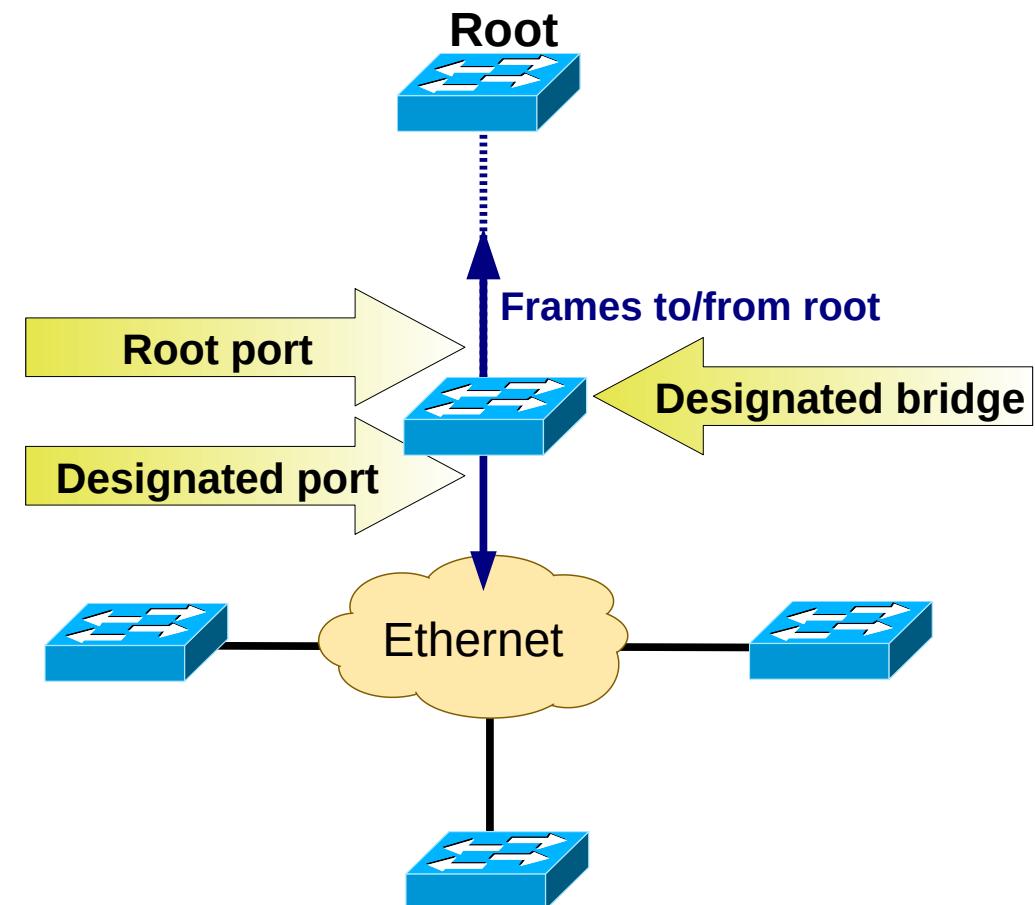


- Root Switch/bridge – Switch chosen as origin/root of the spanning tree.
 - ◆ Switch com **lowest ID**.
- Path cost – Cost associated with each port.
 - ◆ Has a default value, but can be changed by configuration.



Spanning Tree Basic Concepts (3)

- Designated Bridge – Switch responsible to forward the packets from an Ethernet segment to and from the root.
 - ◆ The root bridge is the designated_bridge to all Ethernet segments connected to it.
- Designated Port – Port of the designated bridge that connects an Ethernet segment (to which is designated).
- Root Port – Port of the designated bridge that provides the path to the root.



Spanning Tree Basic Concepts (4)

- Possible Port States

- ◆ **Blocking state:**

- MAC address learning and packet forwarding are disabled;
 - Receives and processes BPDU.
 - After *MaxAge* time without receiving BPDU, it transitions to Listening state.

- ◆ **Listening state:**

- MAC address learning and packet forwarding are disabled;
 - Receives and processes BPDU.
 - When *ForwardDelay* timer expires the port transitions to Learning state.

- ◆ **Learning state:**

- Learns MAC address;
 - Packet forwarding are disabled;
 - Receives and processes BPDU.
 - When *ForwardDelay* timer expires the port transitions to Forwarding state.

- ◆ **Forwarding state:**

- MAC address learning and packet forwarding are enabled;
 - Receives and processes BPDU.

- ◆ **Disabled state:**

- MAC address learning and packet forwarding are disabled;
 - Does not receive BPDU.

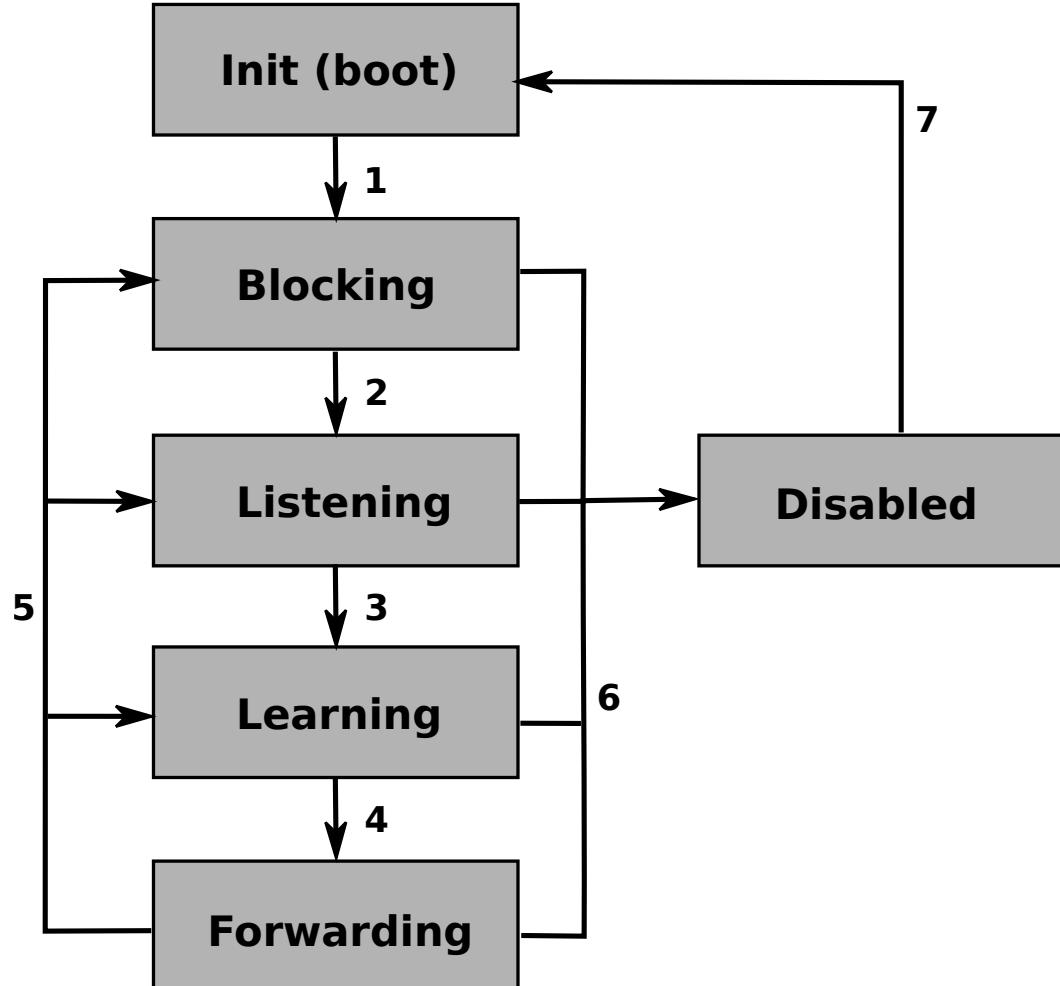


Spanning Tree Basic Concepts (5)

- Each switch has an associated cost of the shortest path to the root (Root Path Cost), given by the sum of the costs of all root ports along the path to the root.
- The Root Port, in each switch, is the port that provides the best path to the root (**lowest Root Path Cost**).
 - ◆ If more than one have the lowest cost, it is chosen the one with the neighbor with the lowest ID.
 - ◆ If more than one link is used to connect to the “best” neighbor it is used the one with the lowest (neighbor) port identifier.
- The Designated Bridge, from each Ethernet segment, is the switch with the **lowest Root Path Cost** from all connected to that segment.
 - ◆ If more than one have the lowest cost, it is chosen the one with the lowest ID.
- The Designated Port, from each Ethernet segment, is the port that connects it to its Designated Bridge.
- The root and designated ports will be in Forwarding state.
- All remaining ports will be in Blocking state.



Port States Diagram



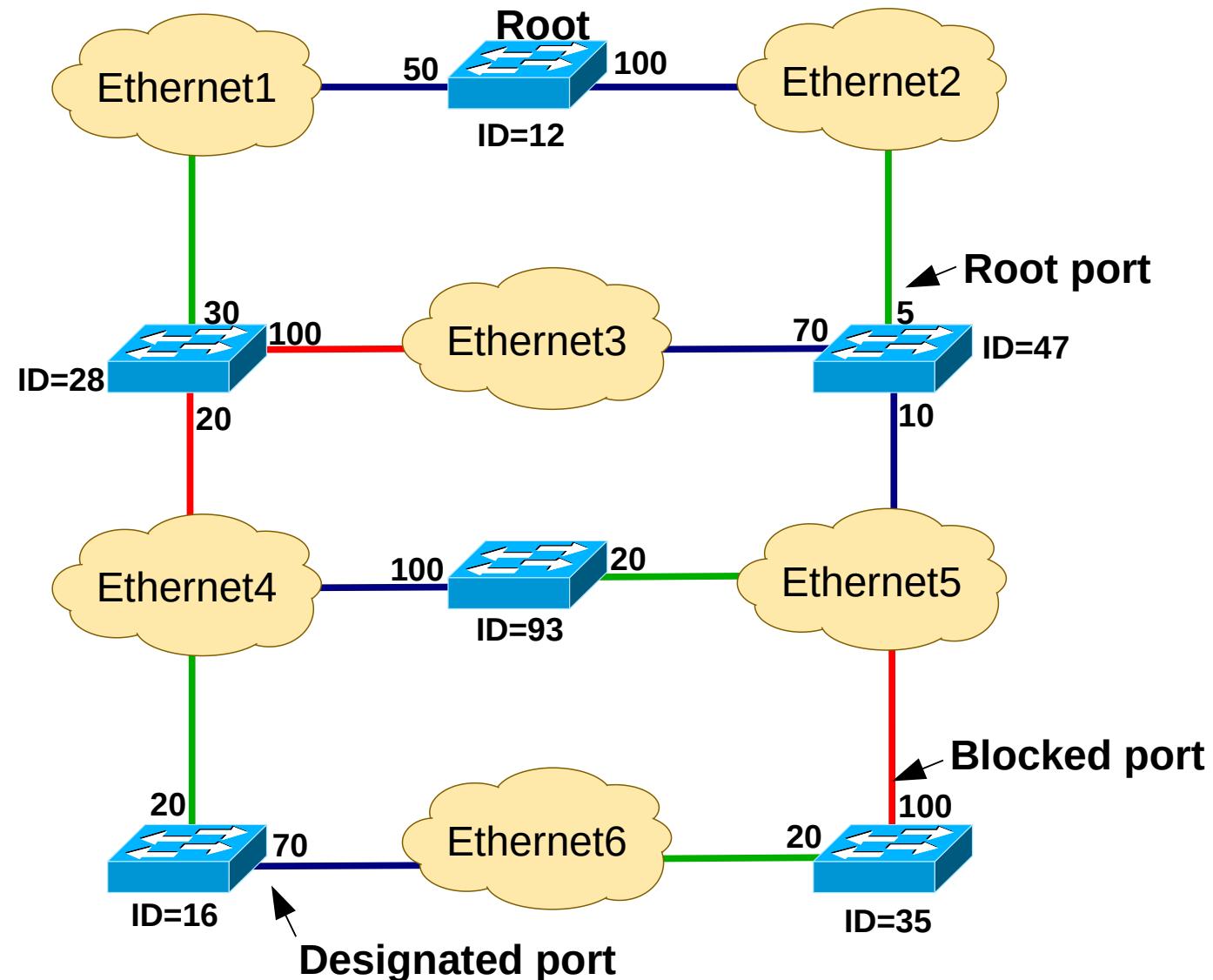
- 1) A port boots up and transitions to **Blocking** state.
- 2) When *MaxAge* timer expires the port transitions to **Listening** state.
- 3) When *ForwardDelay* timer expires the port transitions to **Learning** state.
- 4) When *ForwardDelay* timer expires the port transitions to **Forwarding** state.
- 5) After a topology change the port transitions immediately to **Blocking** state.
- 6) and 7) Administrative actions.



Example – Spanning Tree (1)

Designated bridges

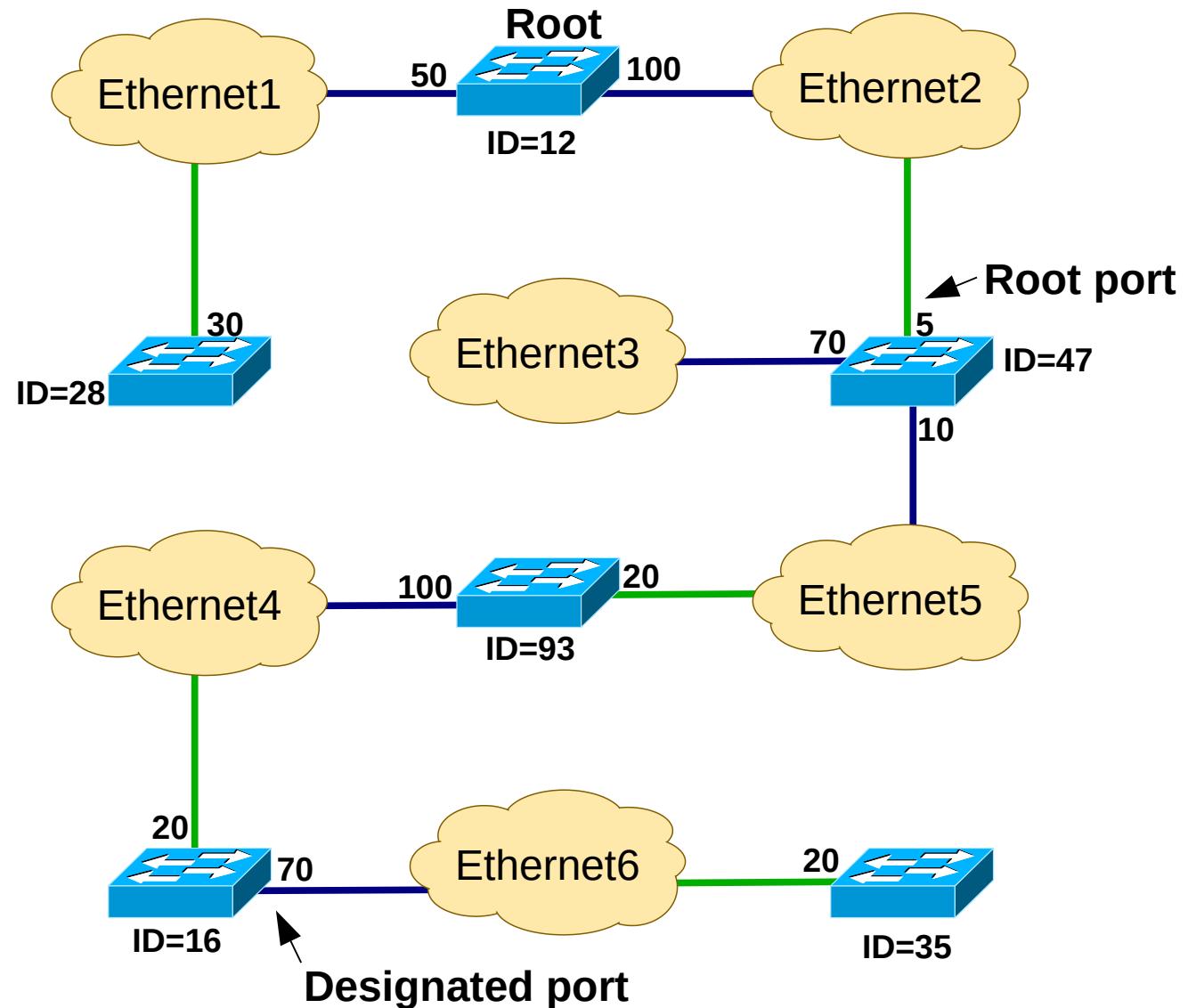
Eth1	12
Eth 2	12
Eth 3	47
Eth 4	93
Eth 5	47
Eth 6	16



Example – Spanning Tree (2)

Designated bridges

Eth1	12
Eth 2	12
Eth 3	47
Eth 4	93
Eth 5	47
Eth 6	16



Protocolo IEEE 802.1D

BPDUs (Bridge Protocol Data Units)

- To build the spanning tree, switches exchange special messages between them called Bridge Protocol Data Units (BPDU).
- There are two types: *Configuration e Topology Change Notification.*

IEEE 802.3 Ethernet

Destination: 01:80:c2:00:00:00 (01:80:c2:00:00:00)

Source: 00:16:e0:9a:c3:92 (00:16:e0:9a:c3:92)

Length: 39

Logical-Link Control

DSAP: Spanning Tree BPDU (0x42)

SSAP: Spanning Tree BPDU (0x42)

Control field: U, func=UI (0x03)

Spanning Tree Protocol

Protocol Identifier: Spanning Tree Protocol (0x0000)

Protocol Version Identifier: Spanning Tree (0)

BPDU Type: Configuration (0x00)

Root ID: 32768 / 00:05:1a:4e:fd:58

Root Path Cost: 200004

Bridge ID: 32768 / 00:16:e0:9a:c3:80

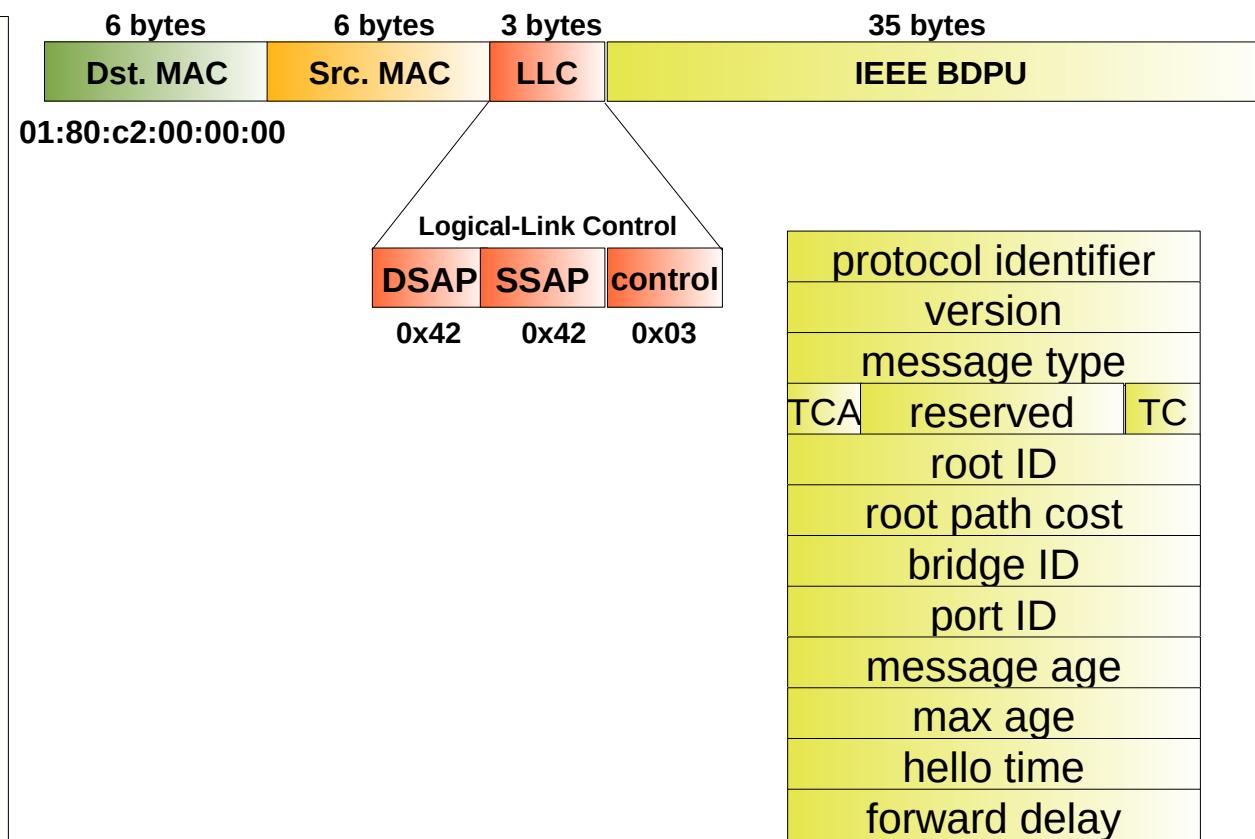
Port ID: 0x8012

Message Age: 1

Max Age: 20

Hello Time: 2

Forward Delay: 15



Configuration BPDU

- The setup of the Spanning Tree id done using Conf - BPDU (configuration messages).

IEEE 802.3 Ethernet

Destination: 01:80:c2:00:00:00 (01:80:c2:00:00:00)
Source: 00:16:e0:9a:c3:92 (00:16:e0:9a:c3:92)
Length: 39

Logical-Link Control

DSAP: Spanning Tree BPDU (0x42)
SSAP: Spanning Tree BPDU (0x42)
Control field: U, func=UI (0x03)

Spanning Tree Protocol

Protocol Identifier: Spanning Tree Protocol (0x0000)
Protocol Version Identifier: Spanning Tree (0)
BPDU Type: Configuration (0x00)

Root ID: 32768 / 00:05:1a:4e:fd:58

Root Path Cost: 200004

Bridge ID: 32768 / 00:16:e0:9a:c3:80

Port ID: 0x8012

Message Age: 1

Max Age: 20

Hello Time: 2

Forward Delay: 15

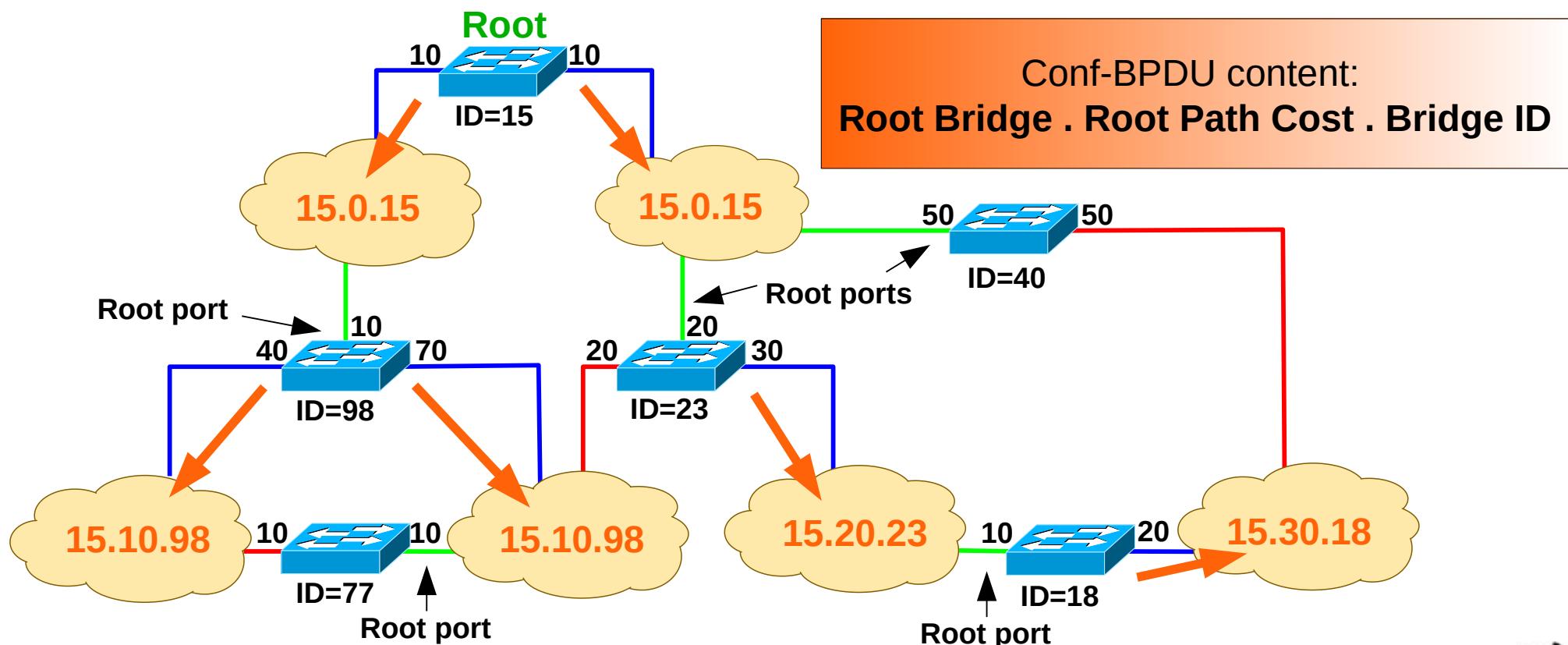
- More relevant fields:

- Root ID: ID of the current root bridge.
- Root Path Cost: estimation of the cost to the root.
- Bridge ID: own bridge identifier.
- Port ID: identifier of the port by which the BPDU was sent.
 - Port priority (1 byte) + Port number



Spanning Tree Maintenance

- Periodically switches sent Conf-BPDUs by its Designated Ports.
 - Periodicity of Conf-BPDU messages = hello time
 - Recommended Hello time: 2 seconds.
 - Defined at the root bridge.



Sorting of Best BPDU

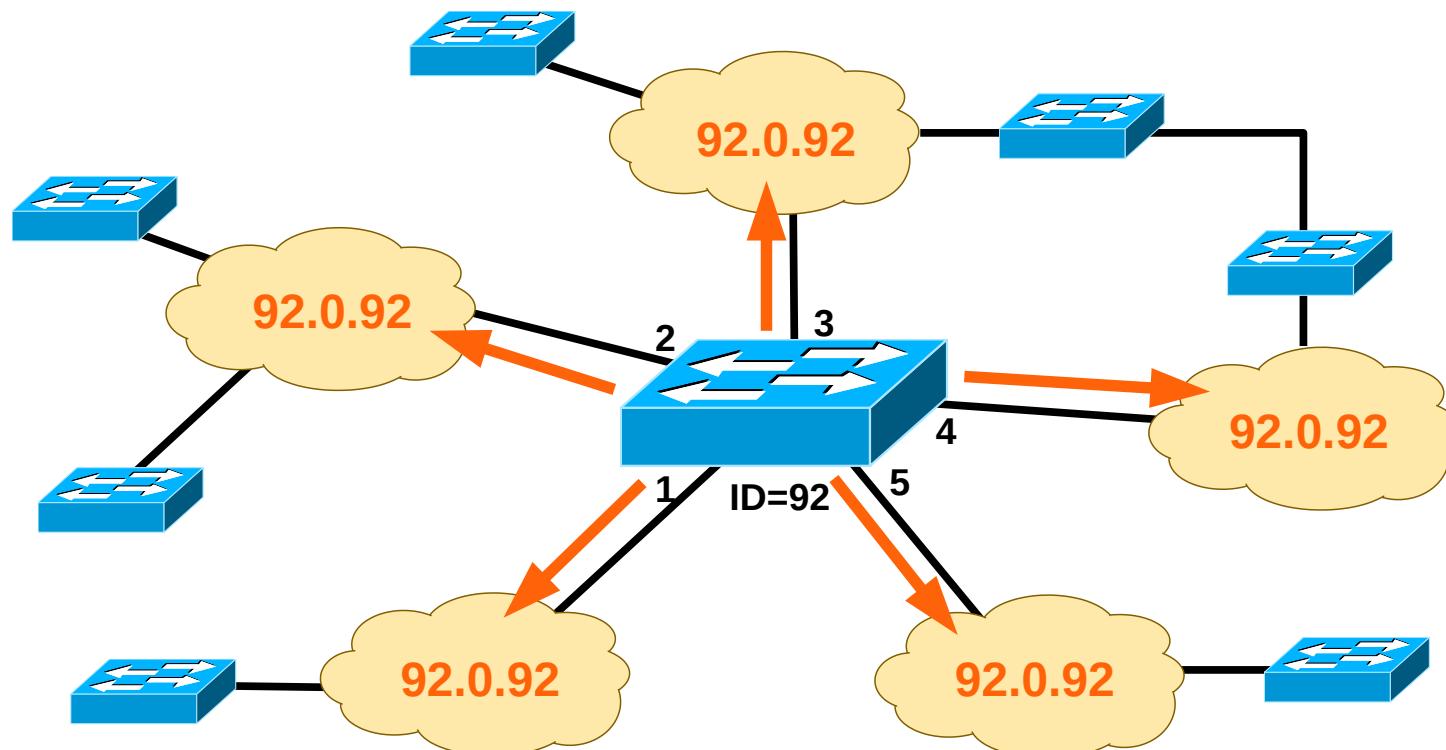
- A Conf-BPDU C1 is considered better than a Conf-BPDU C2 if:
 - ◆ The Root ID of C1 is lower than the one in C2,
 - ◆ With equal Root ID, if Root Path Cost of C1 is lower than the one in C2,
 - ◆ With equal Root ID and Root Path Cost, if the Bridge ID of C1 is lower than the one in C2,
 - ◆ With equal Root ID, Root Path Cost and Bridge ID, if the Port ID of C1 is lower than the one in C2.

Root ID	Root Path Cost	Bridge ID	Port ID
18	27	32	2
18	27	32	4
18	27	43	1
18	35	23	3
23	31	45	2

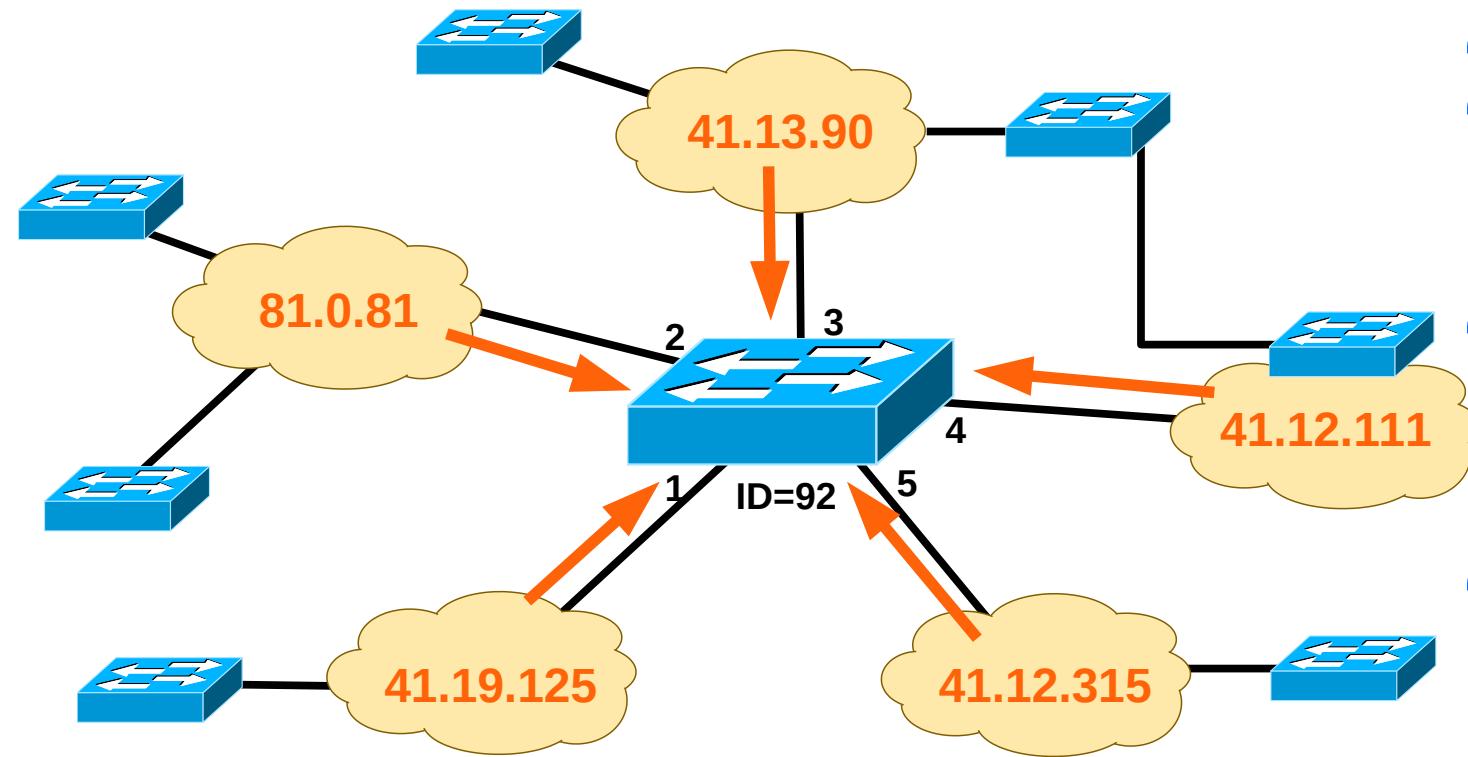


Building the Spanning Tree (1)

- Each switch initially assumes to be the Root Bridge.
 - ◆ Assumes Root Path Cost = 0,
 - ◆ Sends Conf-BPDU to all its ports.



Building the Spanning Tree (2)



Best Conf-BPDU received by Bridge 92 (until now)

Estimations of Bridge 92 (assuming port costs equal to 1).

- Bridge92 is not root (BridgeID 92>41)
- Bridge 92 Root Port is 4.
 - Lowest RootID (41).
 - Lowest Root Path Cost ($12+1=13$).
 - Lowest Neighbor BridgeID ($111 < 315$)
- Bridge 92 is Designated Bridge via ports 1 and 2
 - Port 2, Lowest RootID (41).
 - Port 1, Same RootID (41) and Lowest Root Path Cost ($13 < 19$).
- Bridge 92 ports 3 and 5 are blocked.
 - Neighbors have the same RootID (41).
 - Via port 3, Neighbor has the same Root Path Cost (13), but lower BridgeID ($90 < 92$).
 - Via port 5, Neighbor has lower Root Path Cost (12).

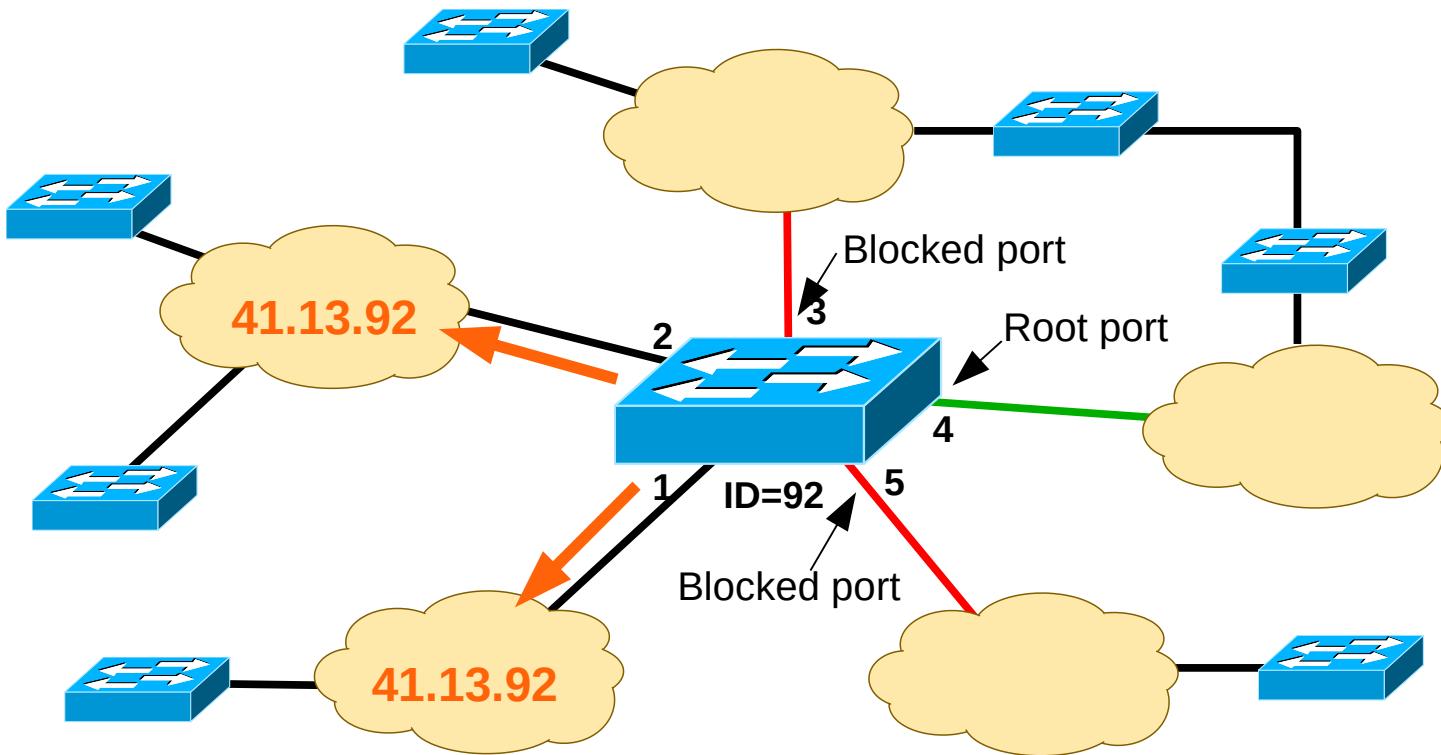
Root Bridge = 41

Root port = 4

Root Path Cost = $12 + 1 = 13$



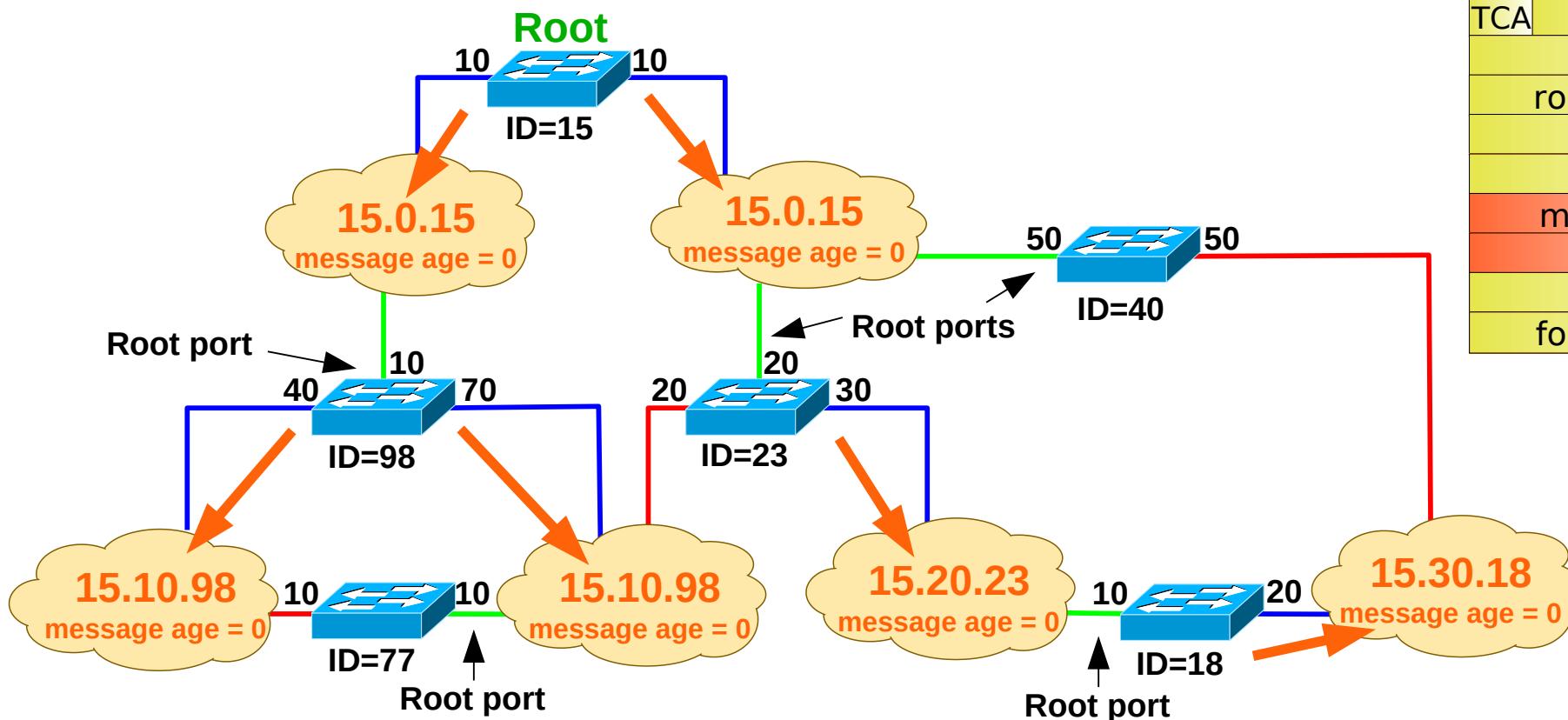
Building the Spanning Tree (3)



Conf-BPDU sent by Bridge 92 - **41.13.92**



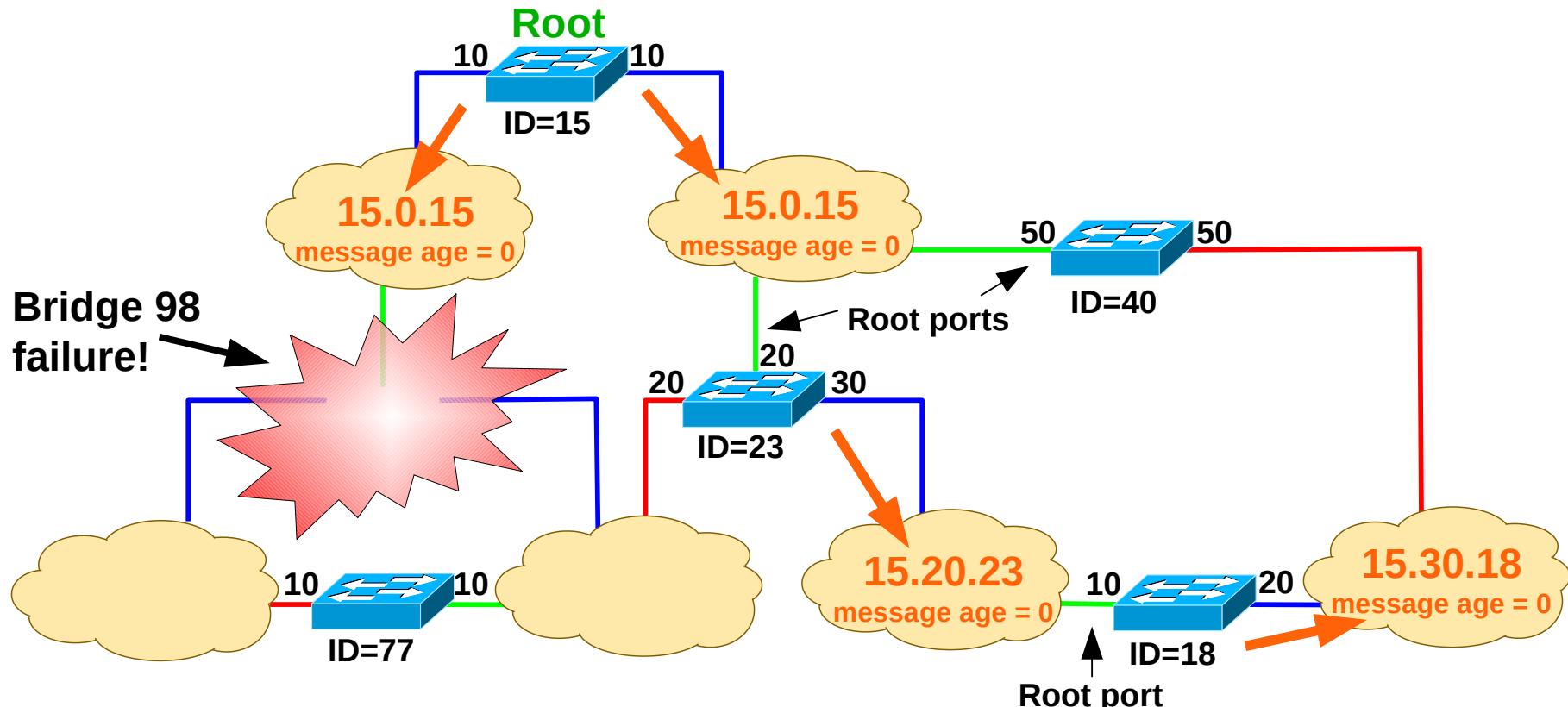
Network Failures (1)



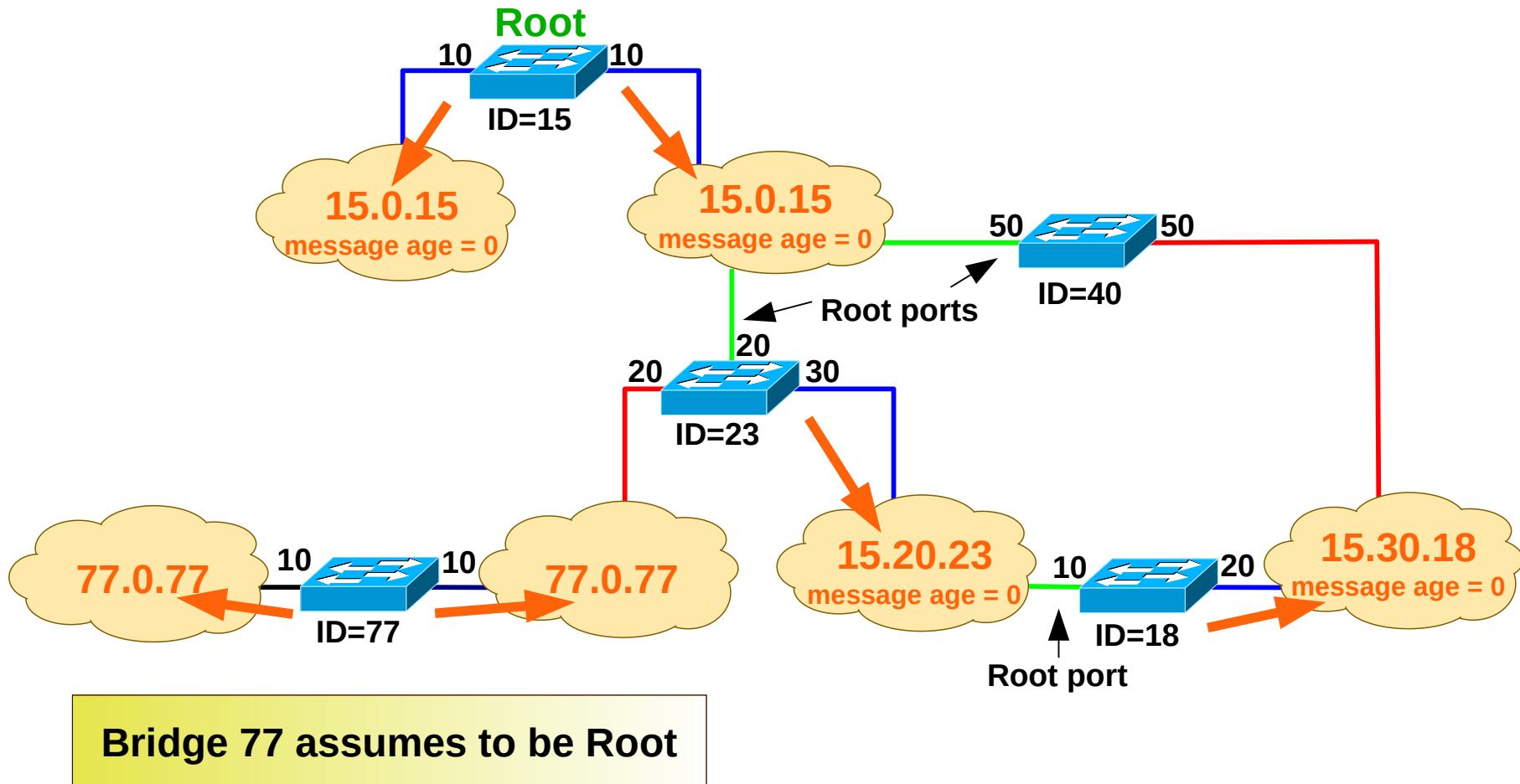
Conf-BPDU		
protocol identifier		
version		
message type		
TCA	reserved	TC
root ID		
root path cost		
bridge ID		
port ID		
message age		
max age		
hello time		
forward delay		



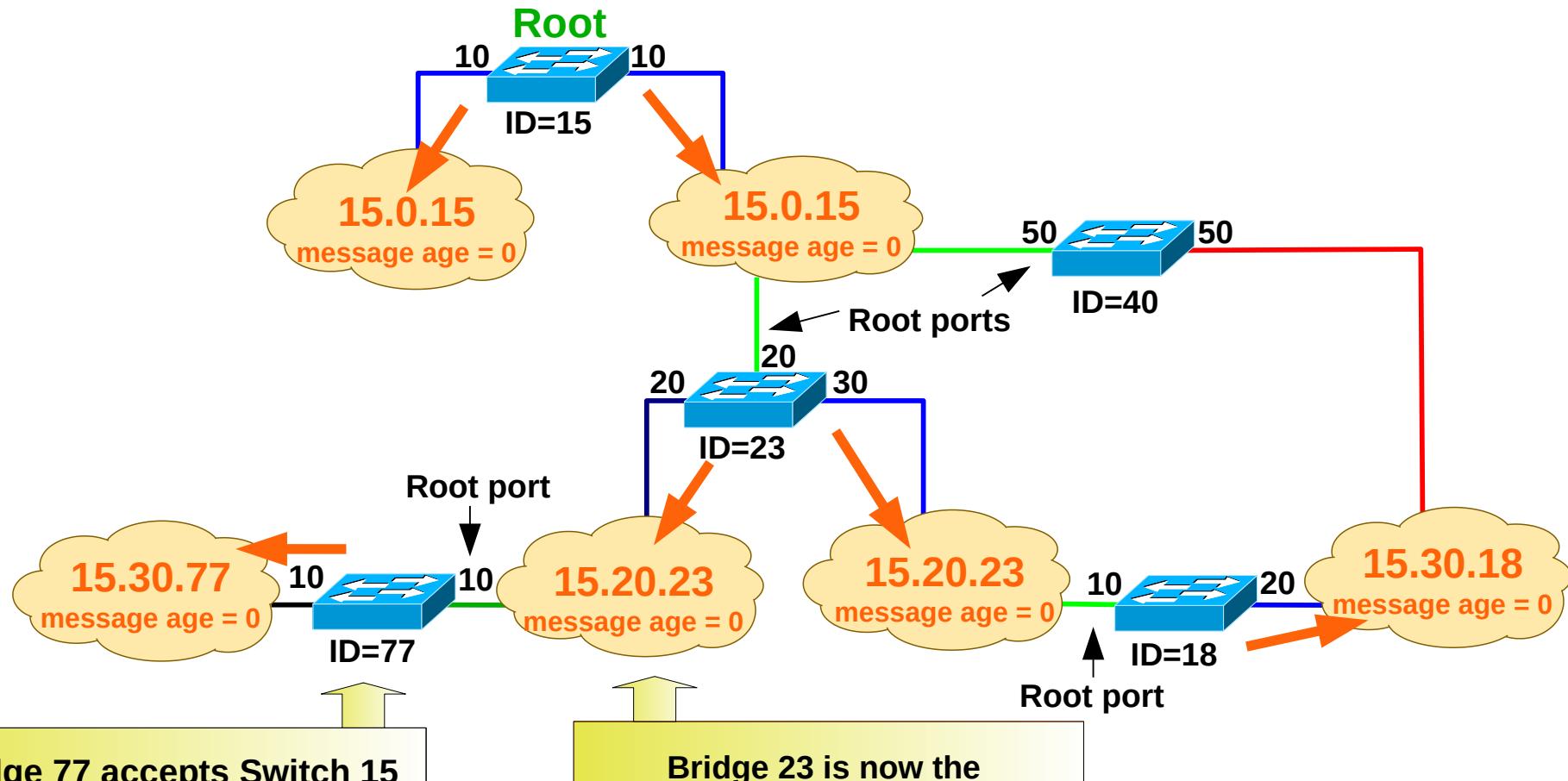
Network Failures (2)



Network Failures (3)



Network Failures (4)



Forwarding Tables Entries Lifetimes

- Forwarding Tables Long Lifetime – Many frames will be lost when network is changing topology.
- Forwarding Tables Short Lifetime – Creates too much traffic due to frequent flooding.
- There are two forwarding tables lifetimes:
 - ◆ **Long**: used by default (recommended value = 300 seconds)
 - ◆ **Short**: used when SPT is re-configuring (recommended value = 15 seconds)



Topology Change Notification

Conf (Configuration) BPDU

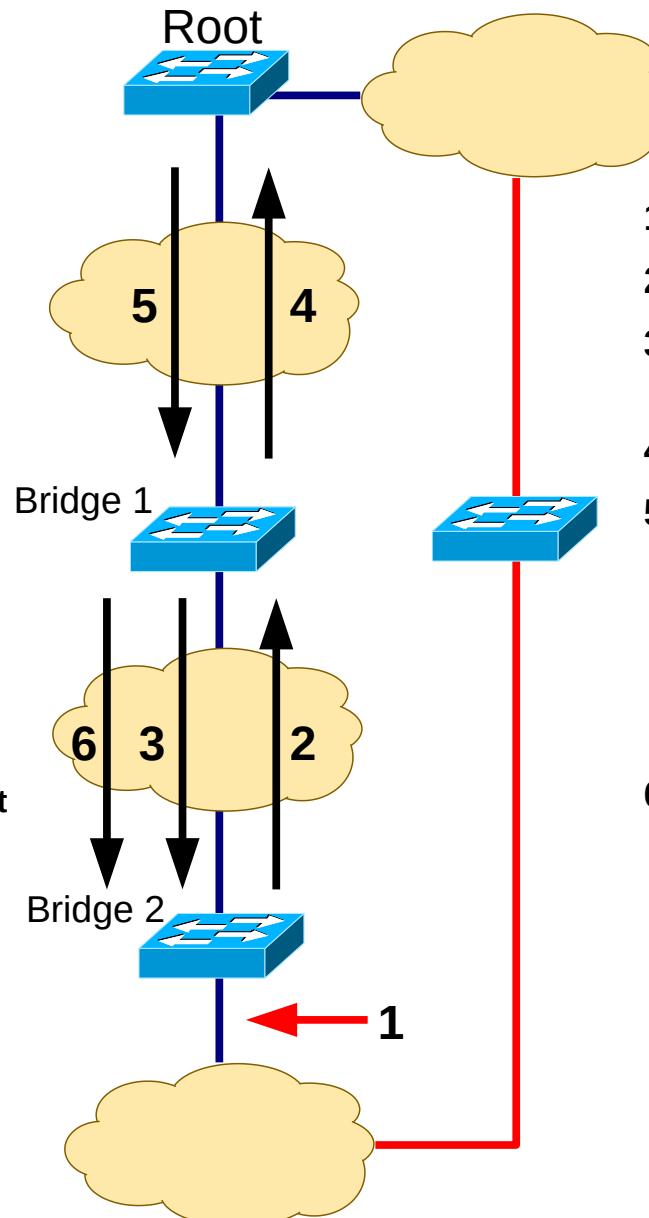
protocol identifier		
version		
message type = 0		
TCA	reserved	TC
root ID		
root path cost		
bridge ID		
port ID		
message age		
max age		
hello time		
forward delay		

TCA - flag Topology Change Acknowledgment

TC - flag Topology Change

TCN (Topology Change Notification)
BPDU

protocol identifier
version
message type = 1



1. Port changes state to disabled or blocking
2. Sends TCN-BPDU (periodicity = hello time)
3. Sends Conf-BPDU with TCA = 1 while receiving TCN-BPDU
4. Sends TCN-BPDU (periodicity = hello time)
5. Sends Conf-BPDU with TCA = 1 while receiving TCN-BPDU and with TC=1 for a period of time equal to *ForwardDelay* + *MaxAge*

Root bridge uses the forwarding table short lifetime during this period

6. Sends Conf-BPDU with TC=1

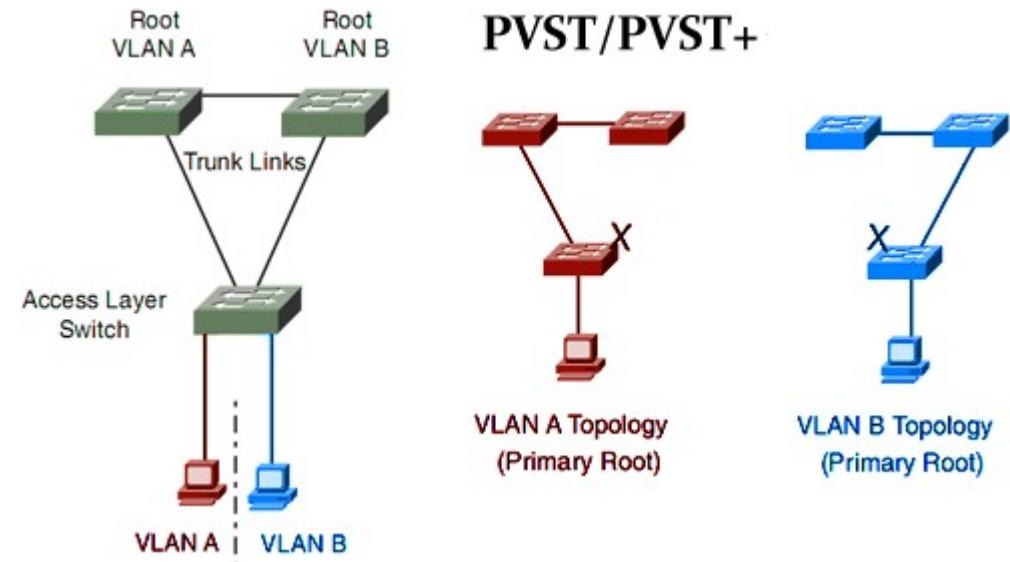
Bridge 1 uses the forwarding table short lifetime while receiving Conf-BPDU with TC=1

Bridge 2 uses the forwarding table short lifetime while receiving Conf-BPDU with TC=1



Other Protocols (1)

- Cisco's proprietary versions of SPT are:
 - ↳ Per-VLAN Spanning Tree (PVST).
 - ↳ Per-VLAN Spanning Tree Plus (PVST+).
- ↳ Create a different spanning tree for each VLAN.
 - ↳ Different roots, costs, blocked ports, etc...
 - ↳ In a complex switching network some switches may not have ports of all VLAN.



```
Ethernet II, Src: c2:00:05:7f:f1:01 (c2:00:05:7f:f1:01), Dst: PVST+ (01:00:0c:cc:cc:cd)
802.1Q Virtual LAN, PRI: 0, CFI: 0, ID: 1
    000. .... .... = Priority: 0
    ...0 .... .... = CFI: 0
    .... 0000 0000 0001 = ID: 1
Length: 50
Logical-Link Control
Spanning Tree Protocol
    Protocol Identifier: Spanning Tree Protocol (0x0000)
    Protocol Version Identifier: Spanning Tree (0)
    BPDU Type: Configuration (0x00)
    BPDU flags: 0x00
    Root Identifier: 32768 / 0 / c2:00:05:7f:00:00
    Root Path Cost: 0
    Bridge Identifier: 32768 / 0 / c2:00:05:7f:00:00
    Port identifier: 0x802a
    Message Age: 0
    Max Age: 20
    Hello Time: 2
```

Identificador da VLAN



Other Protocols (2)

- IEEE 802.1p
 - ◆ Extension of IEEE 802.1Q.
 - ◆ Provides QoS based on relative priorities.
 - ◆ Defines the field *User Priority* (3 bits) that allows 8 levels of priority.
 - ◆ The standard recommends:
 - ✚ Priority 7 : Critical traffic,
 - ✚ Priorities 5–6 : Delay sensitive traffic (voice and live video),
 - ✚ Priorities 1–4 : Delay variation sensitive traffic (*streaming*),
 - ✚ Priority 0 : Other traffic.



Other Protocols (3)

- IEEE 802.1w Rapid Spanning Tree Protocol

- Extension of IEEE 802.1D.
- Speeds up the convergence time of the Spanning Tree in case of topology changes
 - There are only three port states in RSTP that correspond to the three possible operational states.
 - Adds two additional port roles to a port when in blocking state
 - Alternate port: possible alternative Root port.
 - Backup port: possible alternative Designated port.
- Adds a negotiated mechanism between switches.
 - Uses the reserved bits in the Conf-BPDU.

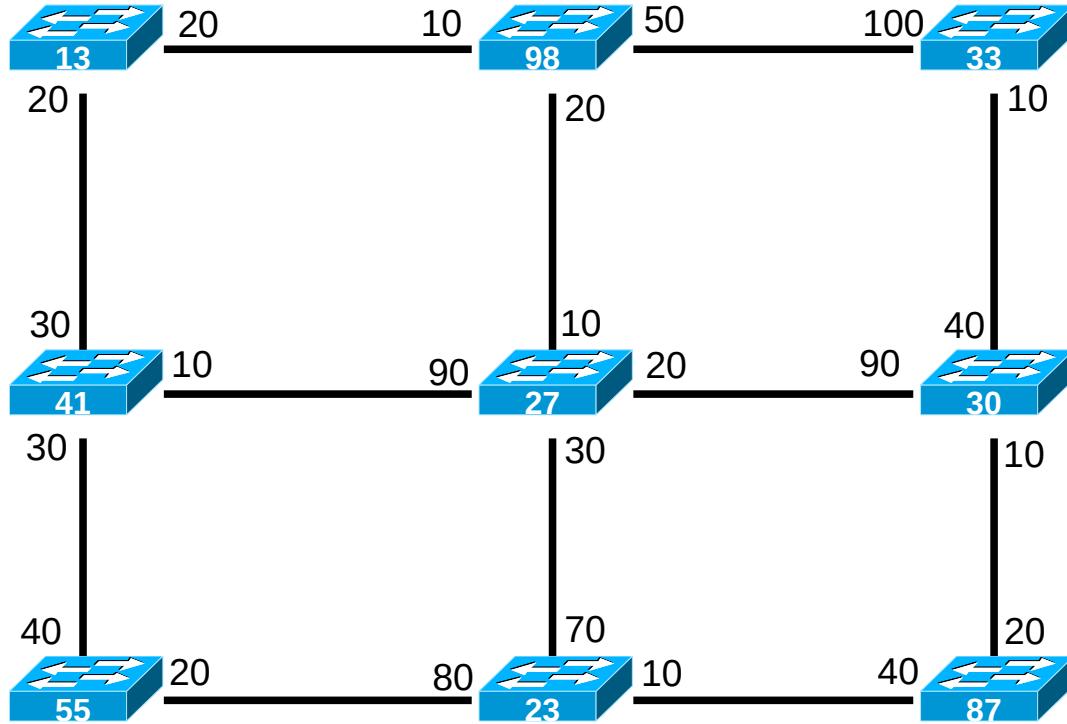
STP (802.1D) Port State	RSTP (802.1w) Port State	Is Port Included in Active Topology?	Is Port Learning MAC Addresses?
Disabled	Discarding	No	No
Blocking	Discarding	No	No
Listening	Discarding	Yes	No
Learning	Learning	Yes	Yes
Forwarding	Forwarding	Yes	Yes

Conf (Configuration) BPDU

protocol identifier		
version		
message type = 0		
TCA	reserved	TC
root ID		
root path cost		
bridge ID		
port ID		
message age		
max age		
hello time		
forward delay		

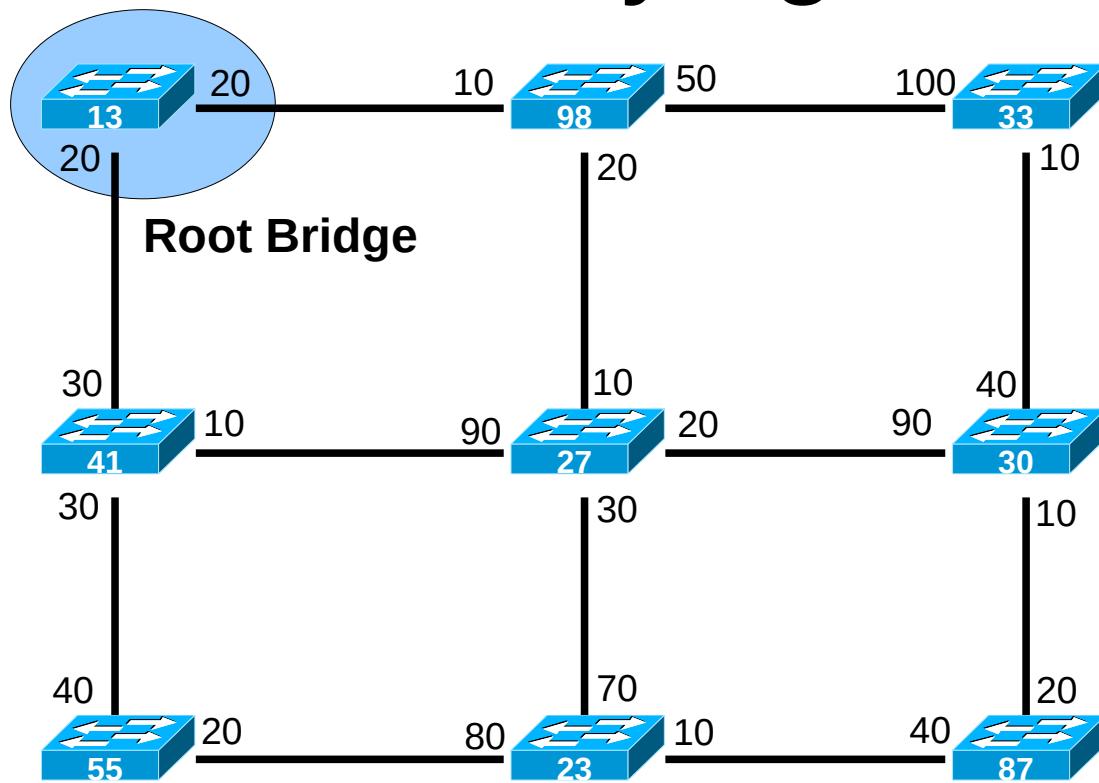


How to determine the Spanning-tree



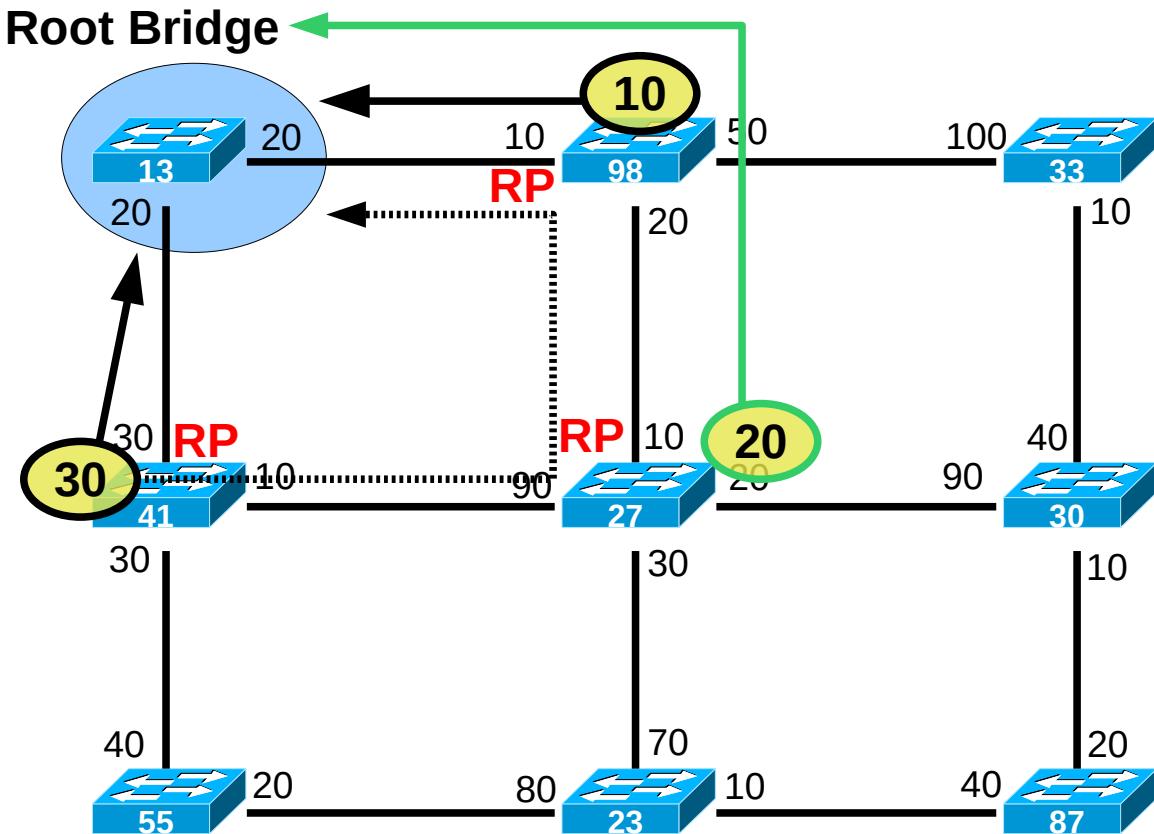
1. Identify the **root bridge**
2. Identify “**root path costs**” and **root ports**
3. Identify **designated bridges** and **designated ports**

Identifying the root bridge



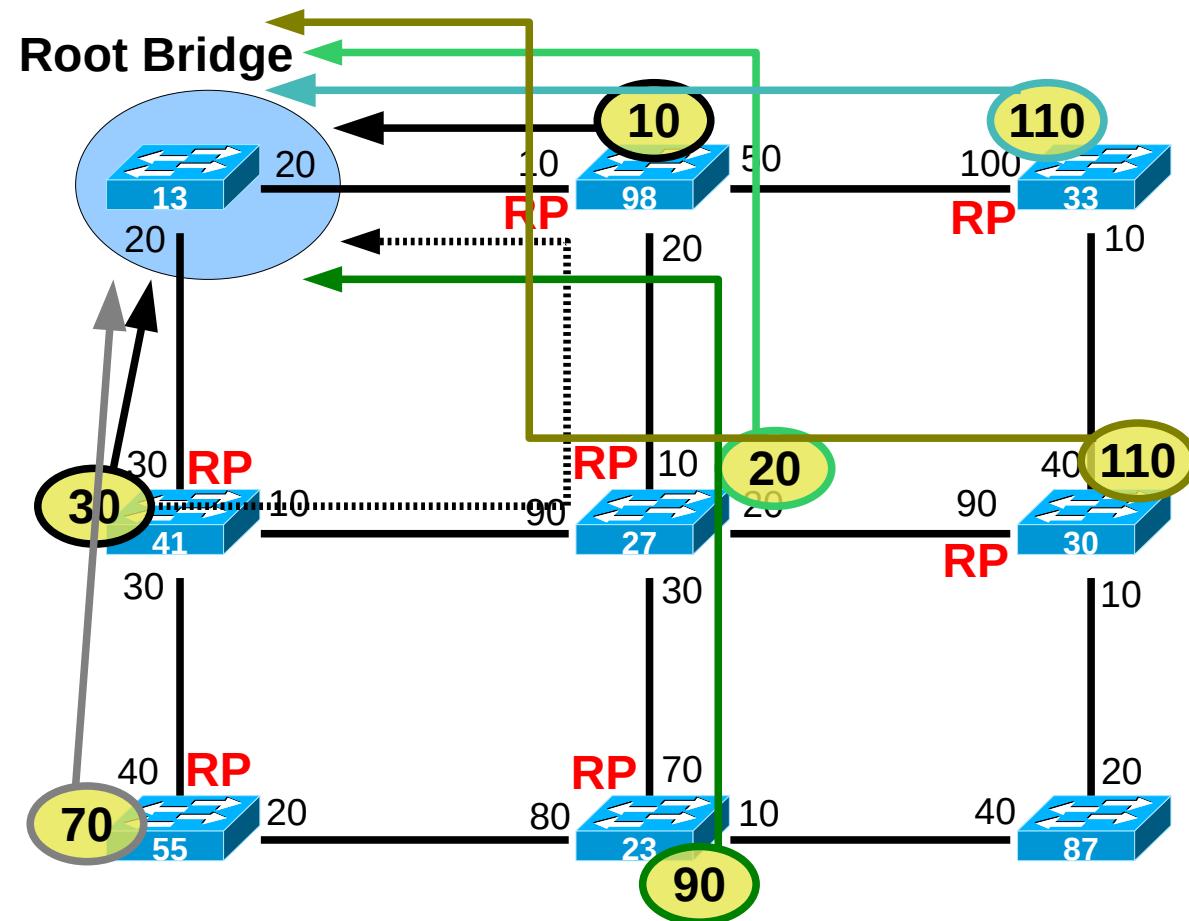
- The root bridge is the one with the lowest ID
 - ID = priority + MAC
 - The bridge with the lowest priority will be the root
 - For equal priorities it's necessary to analyze the bridge's MAC address

“Root Path Costs” and root ports



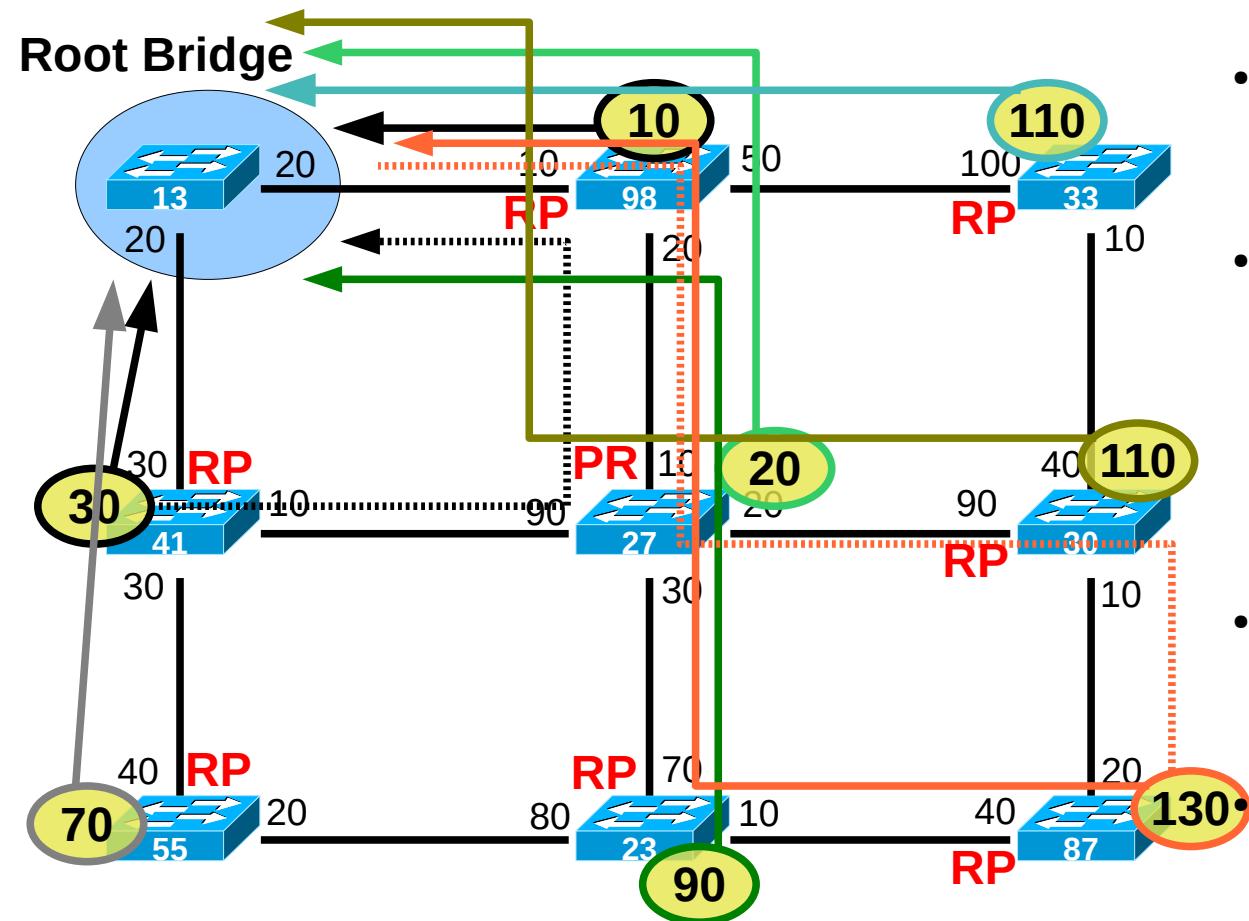
- “Root Path Cost” (RPC) is the cost of the path between a bridge and the root.
- The cost is given by the sum of all “output” ports' costs in the path to the root.
 - In each bridge, it's given by the sum of the RPC of the neighbor bridge plus the cost of the port that connects to that neighbor bridge.
- For paths with the same cost, it's chosen the one announced by the bridge with the lowest ID.
- Tip: start the RPC calculations from the bridges “closer” to the root.

“Root Path Costs” and root ports



- “Root Path Cost” (RPC) is the cost of the path between a bridge and the root.
- The cost is given by the sum of all “output” ports' costs in the path to the root.
 - In each bridge, it's given by the sum of the RPC of the neighbor bridge plus the cost of the port that connects to that neighbor bridge.
- For paths with the same cost, it's chosen the one announced by the bridge with the lowest ID.
- Tip: start the RPC calculations from the bridges “closer” to the root.

“Root Path Costs” and root ports

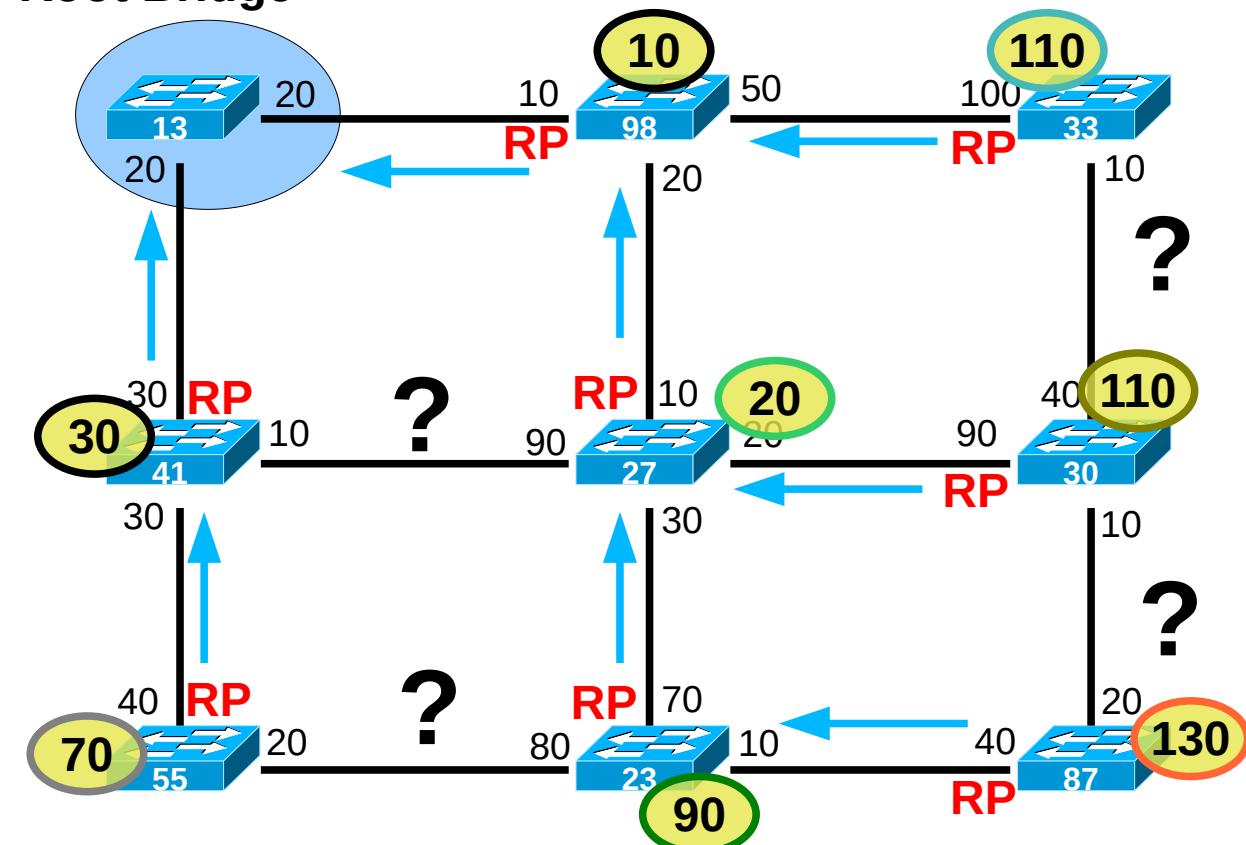


- “Root Path Cost” (RPC) is the cost of the path between a bridge and the root.
- The cost is given by the sum of all “output” ports' costs in the path to the root.
 - In each bridge, it's given by the sum of the RPC of the neighbor bridge plus the cost of the port that connects to that neighbor bridge.
- For paths with the same cost, it's chosen the one announced by the bridge with the lowest ID.

Tip: start the RPC calculations from the bridges “closer” to the root.

Designated bridges and ports

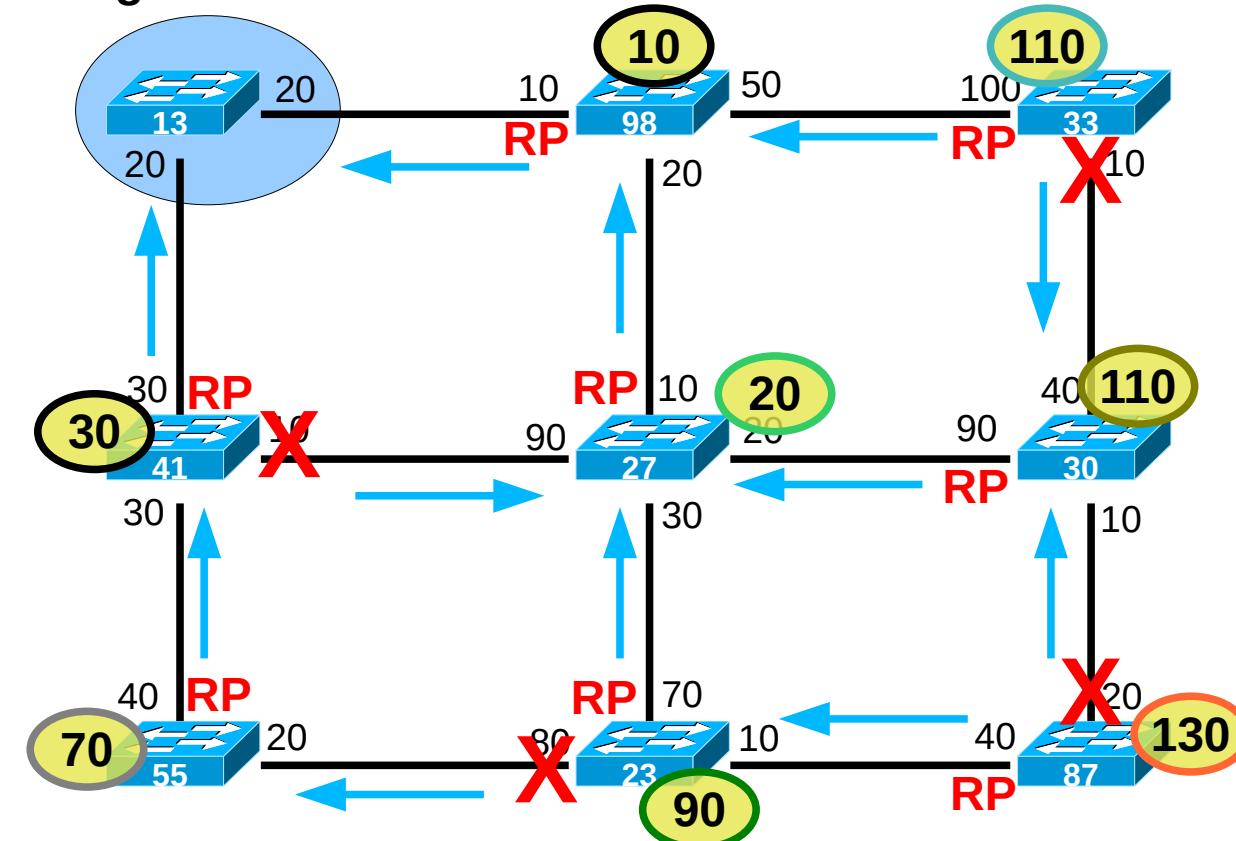
Root Bridge



- A Ethernet segment's designated bridge is the one with:
 - The lowest RPC
 - For equal costs, the one with the lowest ID
- The root bridge is always the designated bridge of all Ethernet segments connected to it.
- In a Ethernet segment that belongs to the minimum cost path, the designated bridge is always the one that provides that path to the root.

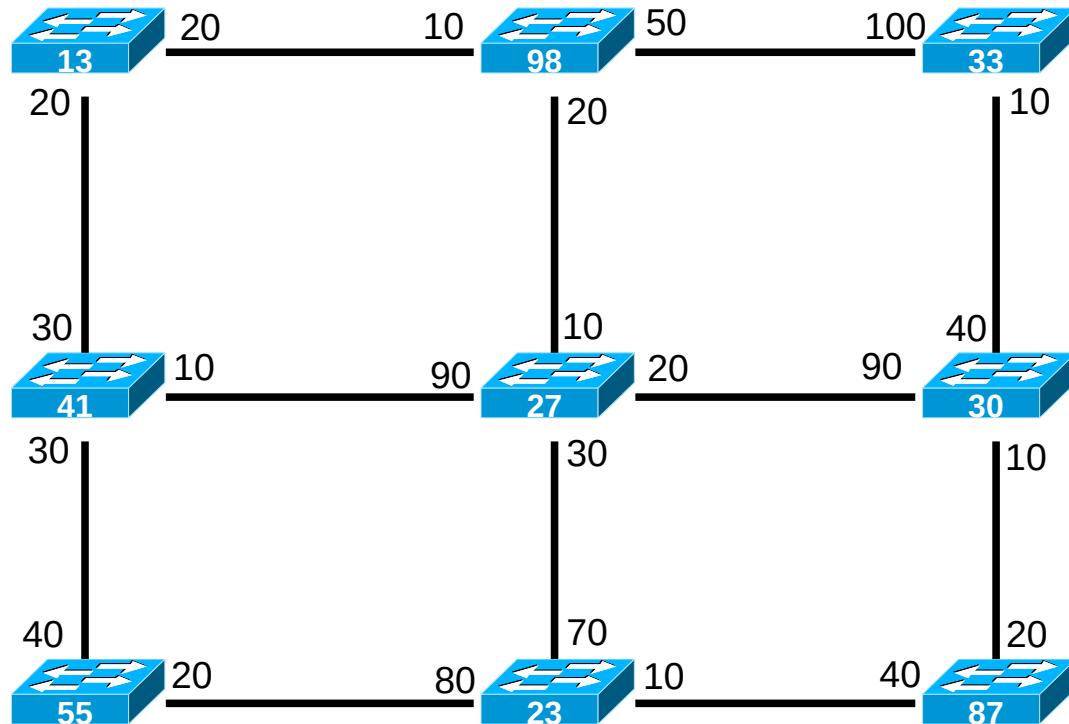
Designated bridges and ports

Bridge raíz



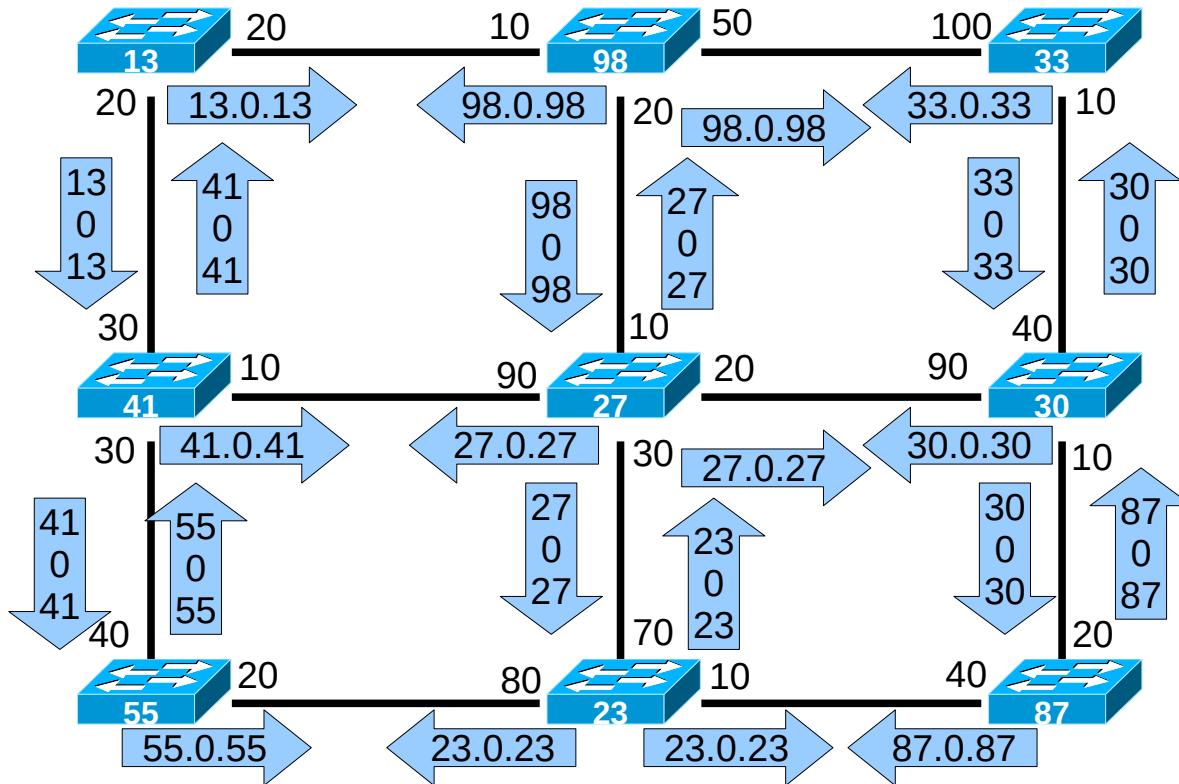
- A Ethernet segment's designated bridge is the one that has:
 - The lowest Root Path Cost
 - For equal costs, the lowest ID
- Ethernet segment 41-27: Designated bridge 27
 - Lowest cost
- Ethernet segment 30-33: Designated bridge 30
 - Same cost, lowest ID
- Ethernet segment 23-55: Designated bridge 55
 - Lowest cost
- Ethernet segment 30-87: Designated bridge 30
 - Lowest cost

ST Construction - Messages Exchange



- At start, all bridges assume to be the root bridge.
- Send Conf-BPDUs to all connected Ethernet segments.

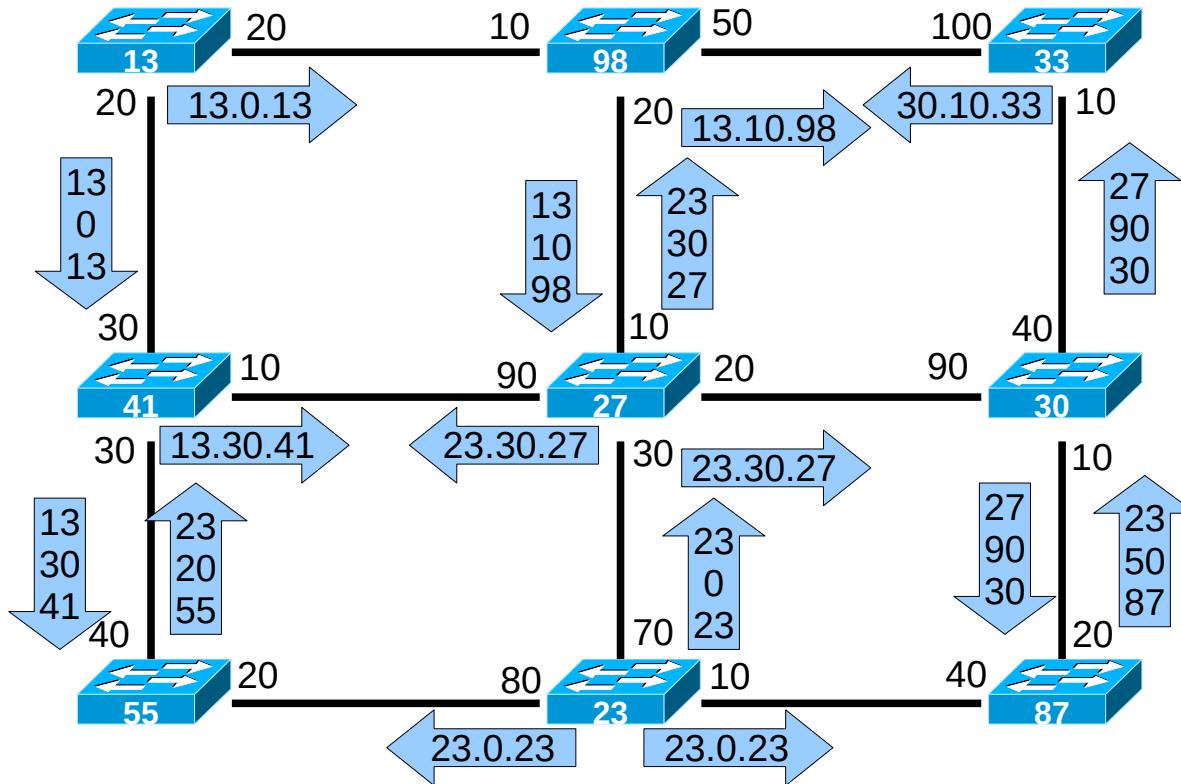
ST Construction - Messages Exchange



- At start, all bridges assume to be the root bridge.
- Send Conf-BPDUs to all connected Ethernet segments.
 - 13 remains root
 - 98 accepts 13 as root (cost 10)
 - 33 accepts 30 as root (cost 10)
 - 41 accepts 13 as root (cost 30)
 - 27 accepts 23 as root (cost 30)
 - 30 accepts 27 as root (cost 90)
 - 55 accepts 23 as root (cost 20)
 - 23 remains root
 - 87 accepts 23 as root (cost 50)

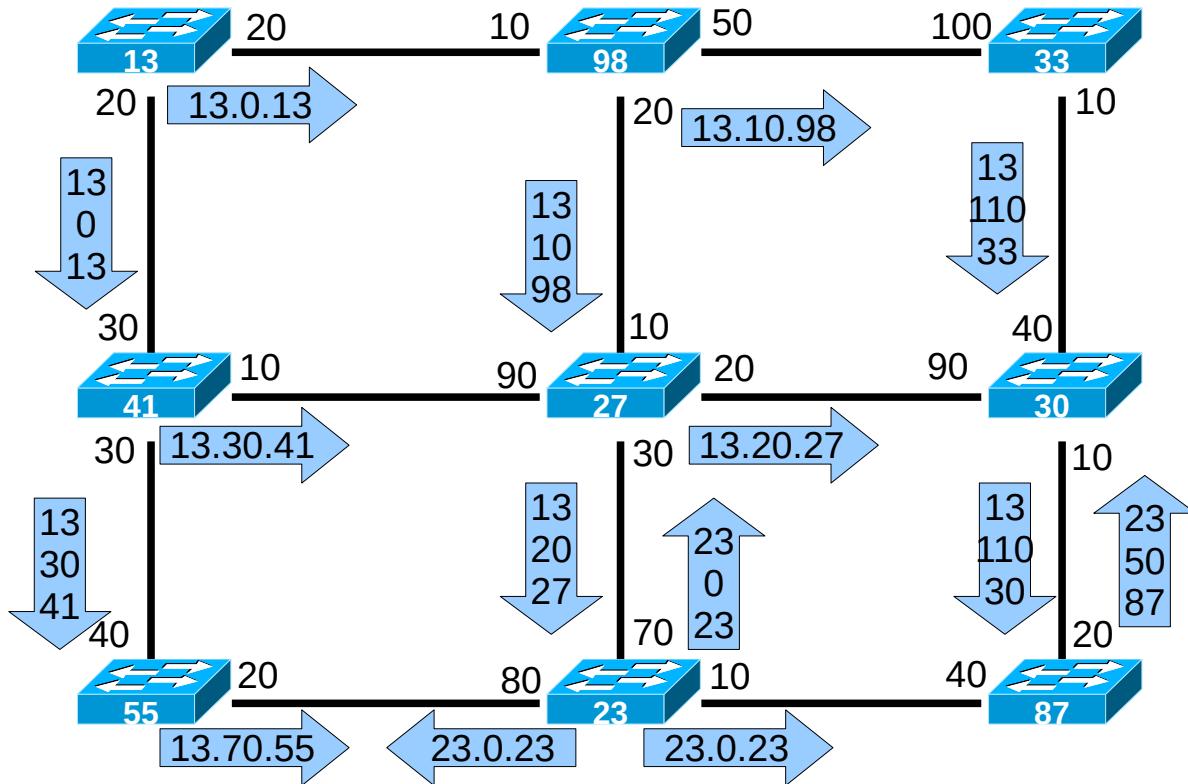
Raíz.Custo.ID

ST Construction - Messages Exchange



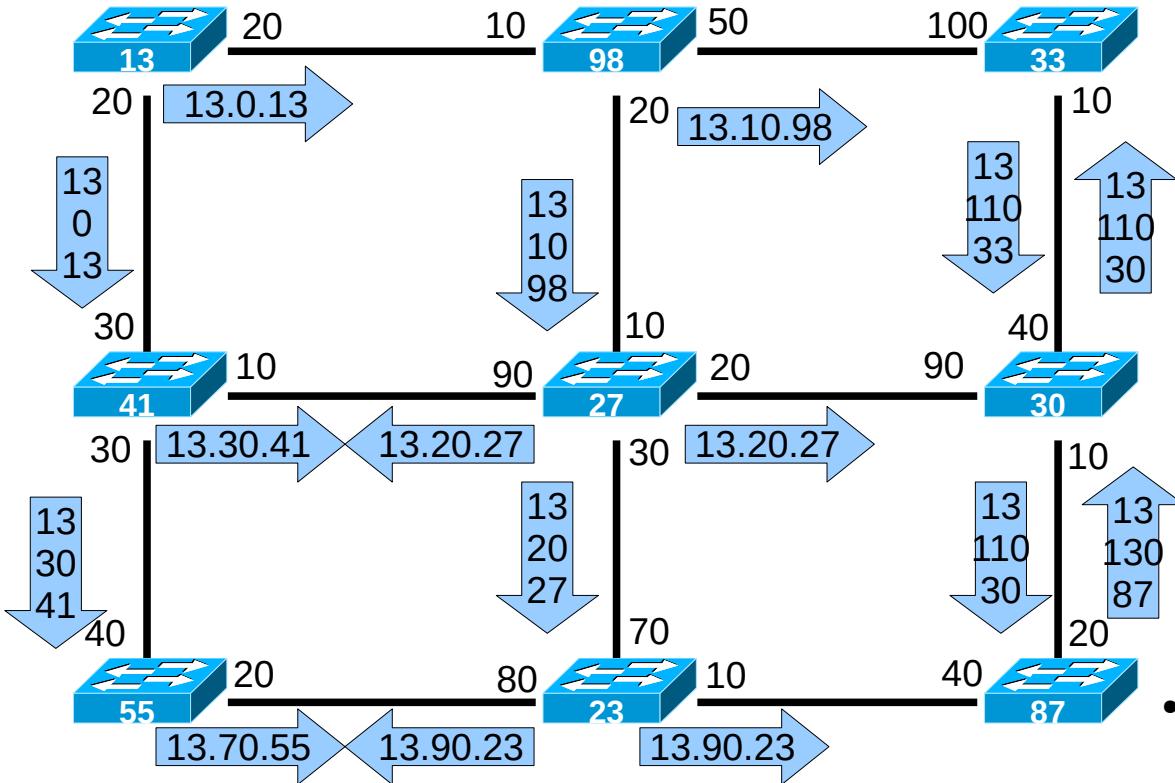
- Bridges only send Conf-BPDUs to the Ethernet segments where they are designated.
 - 13 remains root
 - 98 accepts 13 as root (cost 10)
 - 33 accepts 13 as root (cost 110 – via 98)
 - 41 accepts 13 as root (cost 30)
 - 27 accepts 13 as root (cost 20 – via 98)
 - 30 accepts 23 as root (cost 120 – via 27)
 - 55 accepts 13 as root (cost 70 – via 41)
 - 23 remains root
 - 87 accepts 23 as root (cost 40)

ST Construction - Messages Exchange



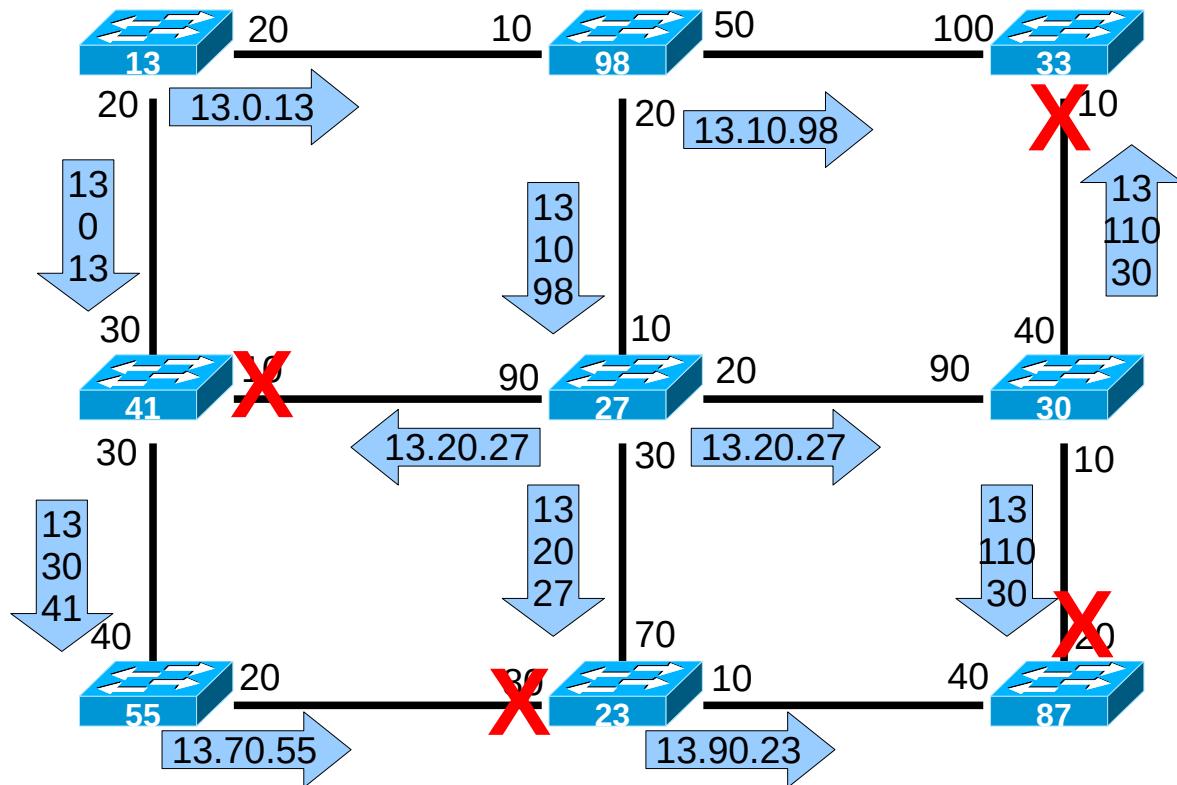
- Bridges only send Conf-BPDUs to the Ethernet segments where they are designated.
 - 13 remains root
 - 98 accepts 13 as root (cost 10)
 - 33 accepts 13 as root (cost 110 – via 98)
 - 41 accepts 13 as root (cost 30)
 - 27 accepts 13 as root (cost 20 – via 98)
 - 30 accepts 13 as root (cost 110 – via 27)
 - 55 accepts 13 as root (cost 70 – via 41)
 - 23 accepts 13 as root (cost 90 – via 27)
 - 87 accepts 13 as root (cost 130 – via 30)

ST Construction - Messages Exchange



- Bridges only send Conf-BPDUs to the Ethernet segments where they are designated.
 - 13 remains root
 - 98 accepts 13 as root (cost 10)
 - 33 accepts 13 as root (cost 110 – via 98)
 - 41 accepts 13 as root (cost 30)
 - 27 accepts 13 as root (cost 20 – via 98)
 - 30 accepts 13 as root (cost 110 – via 27)
 - 55 accepts 13 as root (cost 70 – via 41)
 - 23 accepts 13 as root (cost 90 – via 27)
 - 87 accepts 13 as root (cost 130 – via 23)
 - Cost 130 – via 23 is preferred because the bridge ID is lower ($23 < 30$)
- The designated bridge of a Ethernet segment is chosen according with the best messages sent.
 - Ethernet segment 41-27: designated bridge 27 (lowest cost)
 - Ethernet segment 55-23: designated bridge 55 (lowest cost)
 - Ethernet segment 30-33: designated bridge 30 (Lowest bridge ID)
 - Ethernet segment 30-87: designated bridge 30 (lowest cost)

ST Construction - Messages Exchange



Network Design Models

Redes de Comunicações II

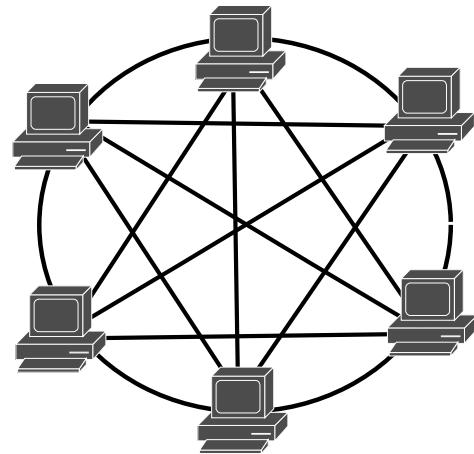
**Licenciatura em
Engenharia de Computadores e Informática
DETI-UA**



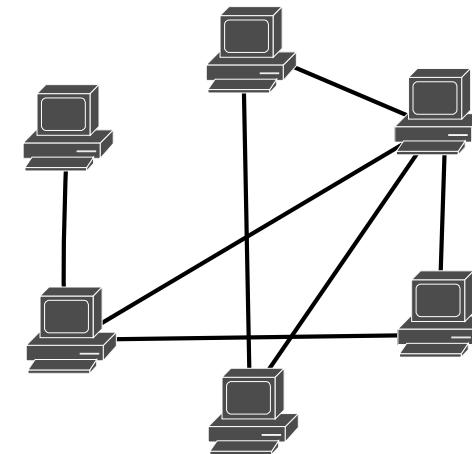
universidade de aveiro

deti.ua.pt

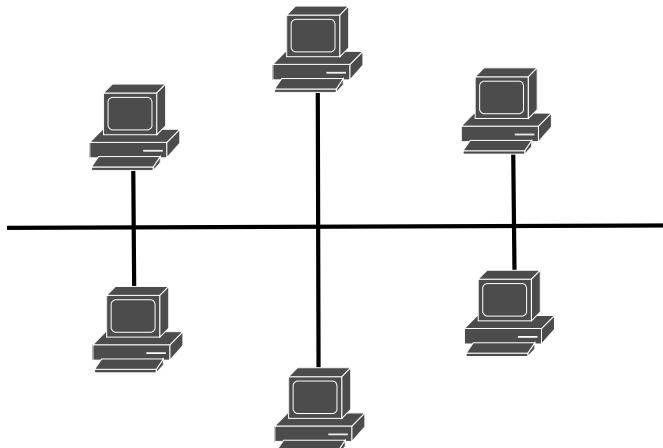
Types of Network Topology



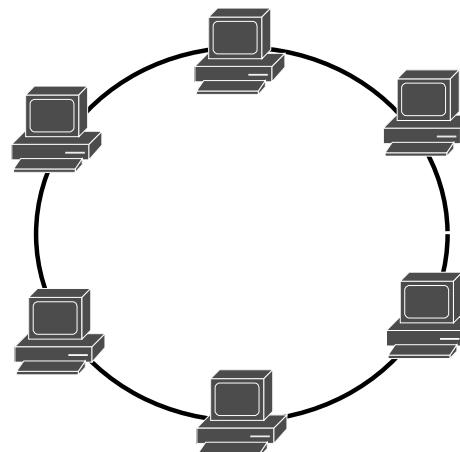
Fully Connected



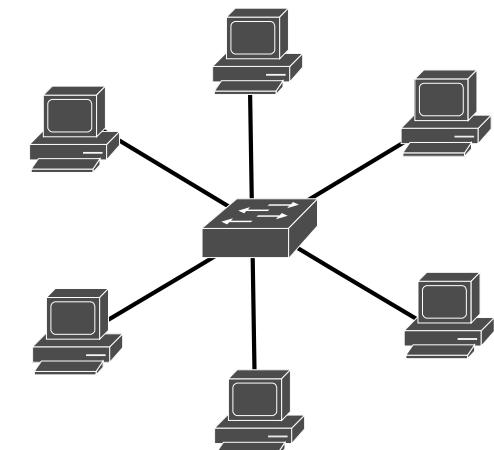
Mesh



Common Bus



Ring



Star

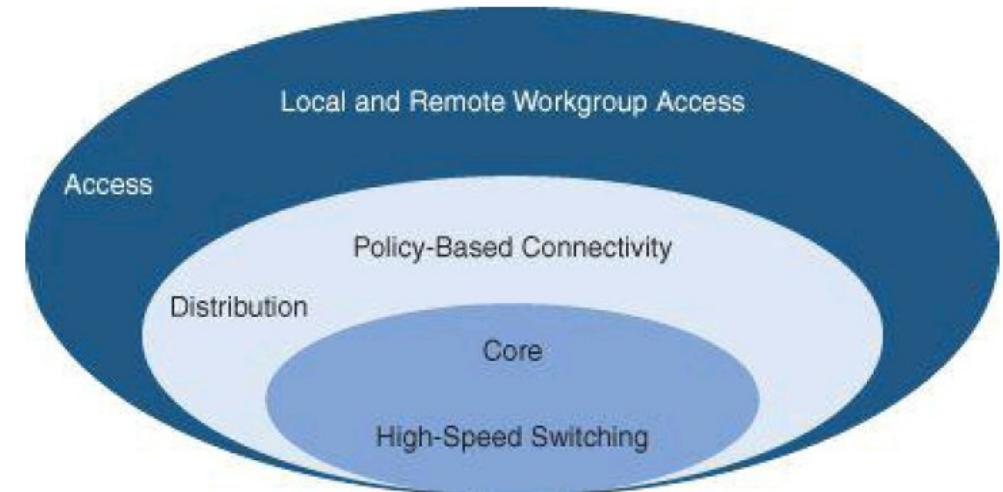


Objectives of Network Design

- Network should be **Modular**
 - ◆ Support growth and change.
 - ◆ Scaling the network is eased by adding new modules instead of complete redesigns.
- Network should be **Resilient**
 - ◆ Up-time close to 100 percent.
 - If network fails in some companies (e.g. financial), even for a second, may represent millions of lost revenue.
 - If network fails in a modern hospital, this may represent lost of lives.
 - ◆ Resilience has costs.
 - Resilience level should be a trade-off between available budget and acceptable risk.
- Network should have **Flexibility**
 - ◆ Businesses change and evolve.
 - ◆ Network should adapt quickly.



Hierarchical Network Model



- **Access layer**

- Provides user access to network.
- Generally incorporates switched LAN devices that provide connectivity to workstations, IP phones, servers, and wireless access points.
- For remote users or remote sites provide an entry to the network across WAN technology.

- **Distribution layer**

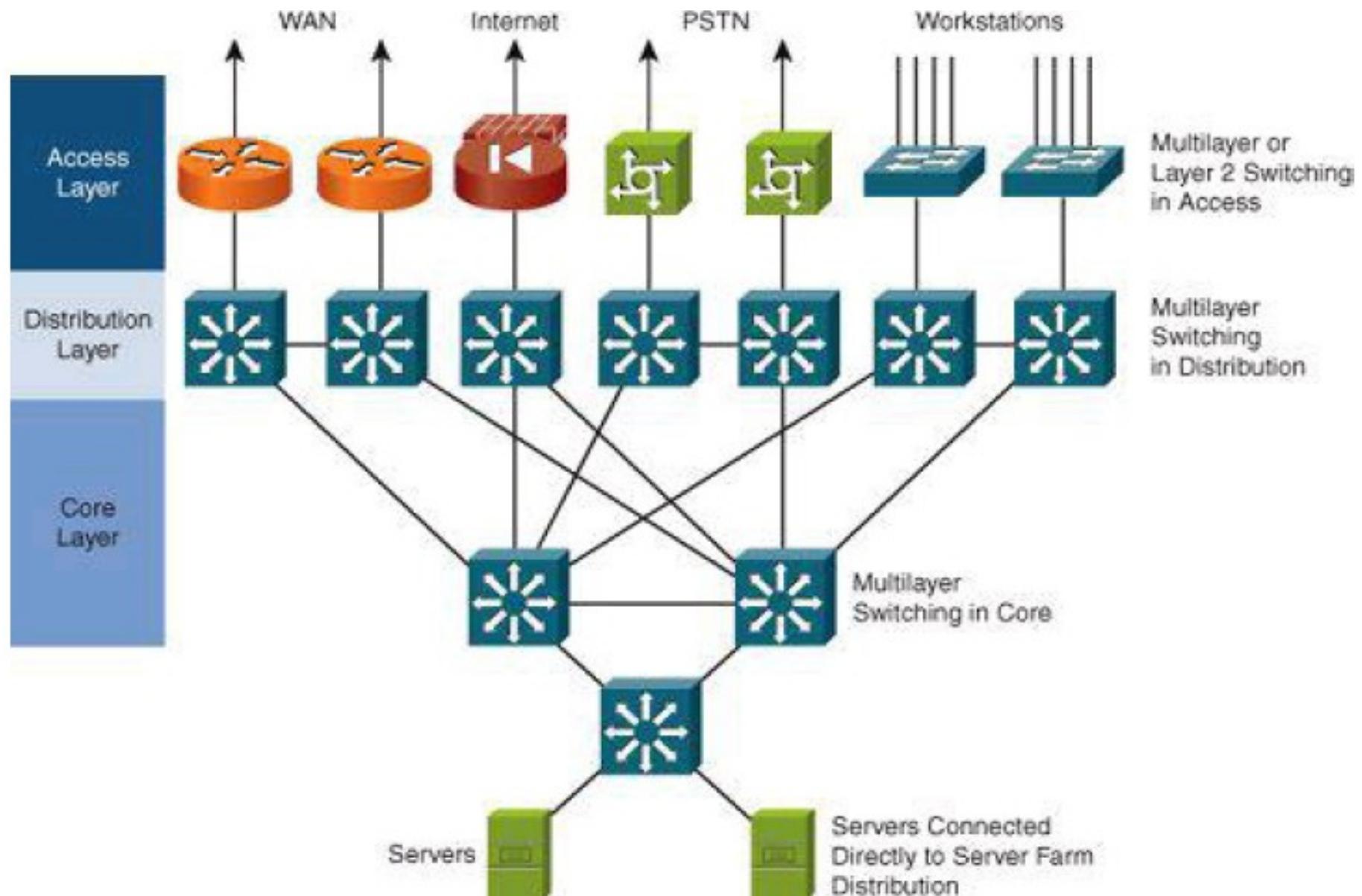
- Aggregates LAN devices.
- Segments work groups and isolate network problems.
- Aggregates WAN connections at the edge of the campus and provides policy-based connectivity.
- Implements QoS policies.

- **Core layer**

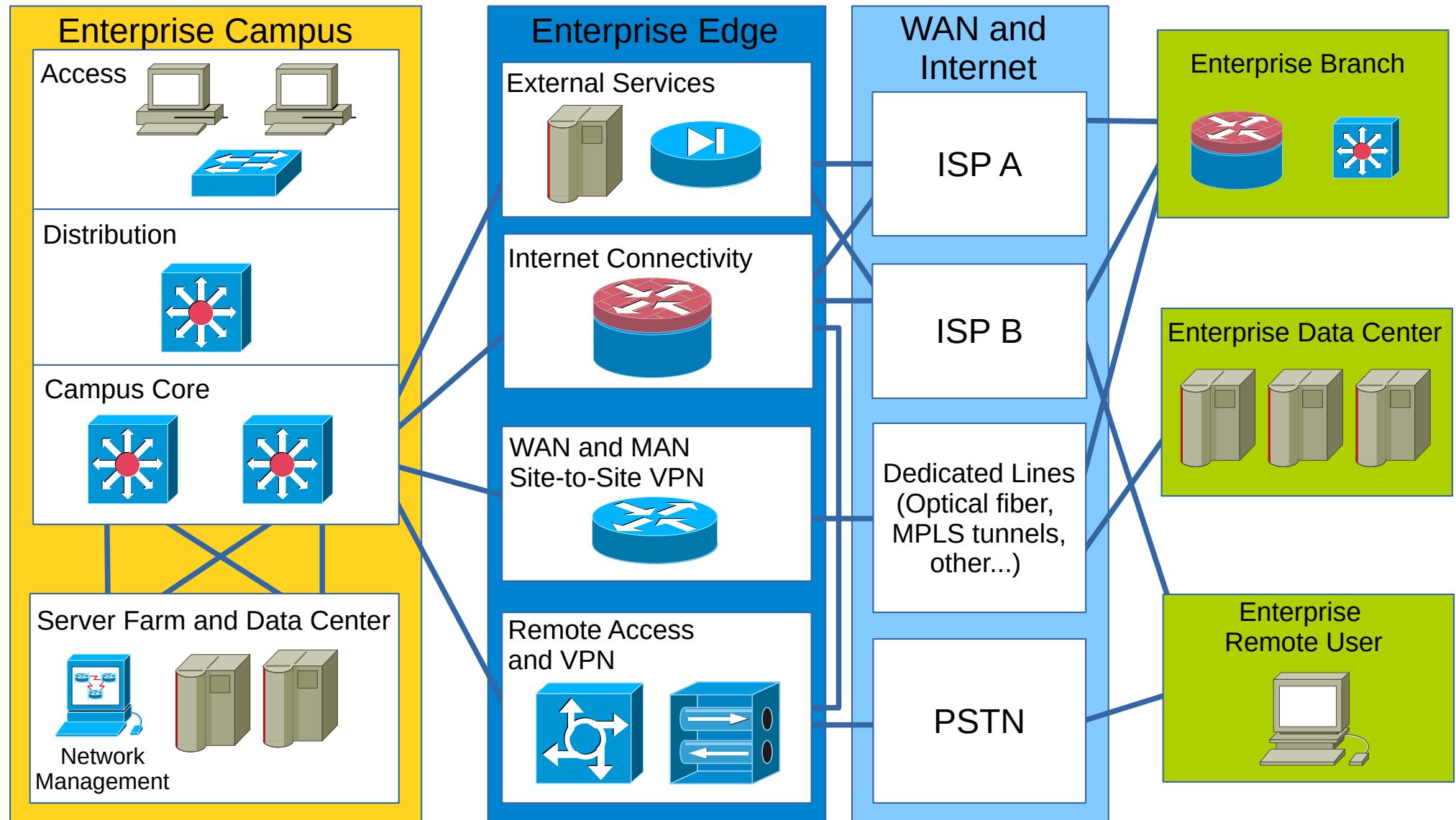
- A high-speed backbone.
- Core is critical for connectivity, must provide a high level of availability and adapt quickly to changes.
- Should provide scalability and fast convergence.
- Should provide an integration point for data center.



A Hierarchical Network



Modular Network Design



Network Modules (1)

- Campus
 - ◆ Operating center of an enterprise.
 - ◆ This module is where most users access the network.
 - ◆ Combines a core infrastructure of intelligent switching and routing with mobility, and advanced security.
- Data Center
 - ◆ Redundant data centers provide backup and application replication.
 - ◆ Network and devices offer server and application load balancing to maximize performance.
 - ◆ Allows the enterprise to scale without major changes to the infrastructure.
 - ◆ Can be located either at the campus as a server farm and/or at a remote facility.
- Branch
 - ◆ Allows enterprises to extend head-office applications and services to remote locations and users or to a small group of branches.
 - ◆ Provides secure access to voice, mission-critical data, and video applications.
 - ◆ Should provide a robust architecture with high levels of resilience for all the branch offices.



Network Modules (2)

- WAN and MAN

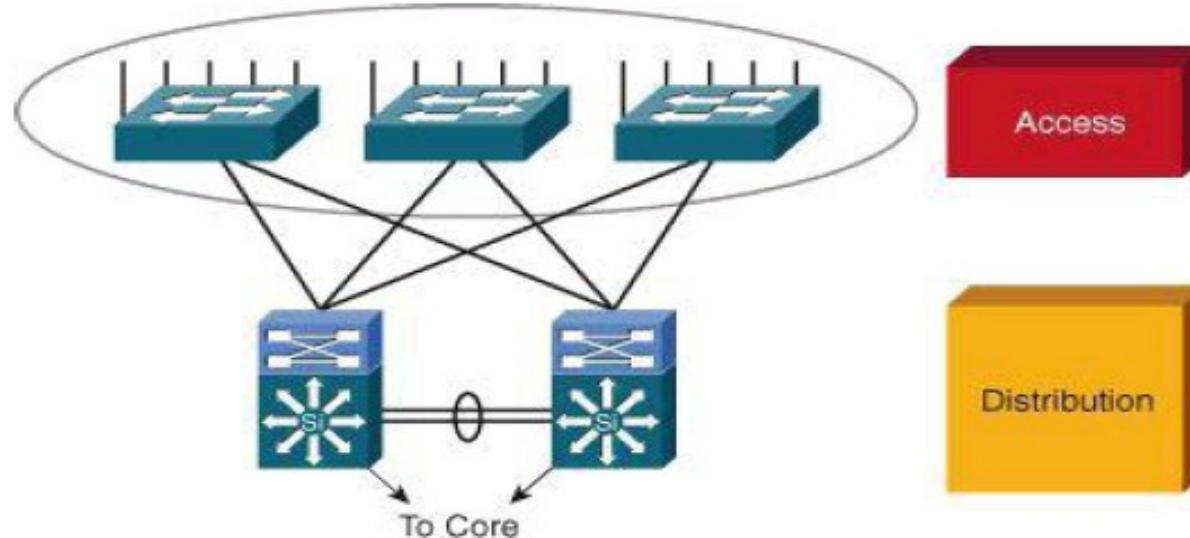
- Offers the convergence of voice, video, and data services.
- Enables the enterprise a cost-effectively presence in large geographic areas.
- QoS, granular service levels, and comprehensive encryption options help ensure the secure delivery to all sites.
- Security is provided with multiservice VPNs (IPsec and MPLS) over Layer 2 or Layer 3 communications.

- Remote User

- Allows enterprises to securely deliver voice and data services to a remote small office/home office (SOHO) over a standard broadband access service.
- Allows a secure log in to the network over a VPN and access to authorized applications and services.



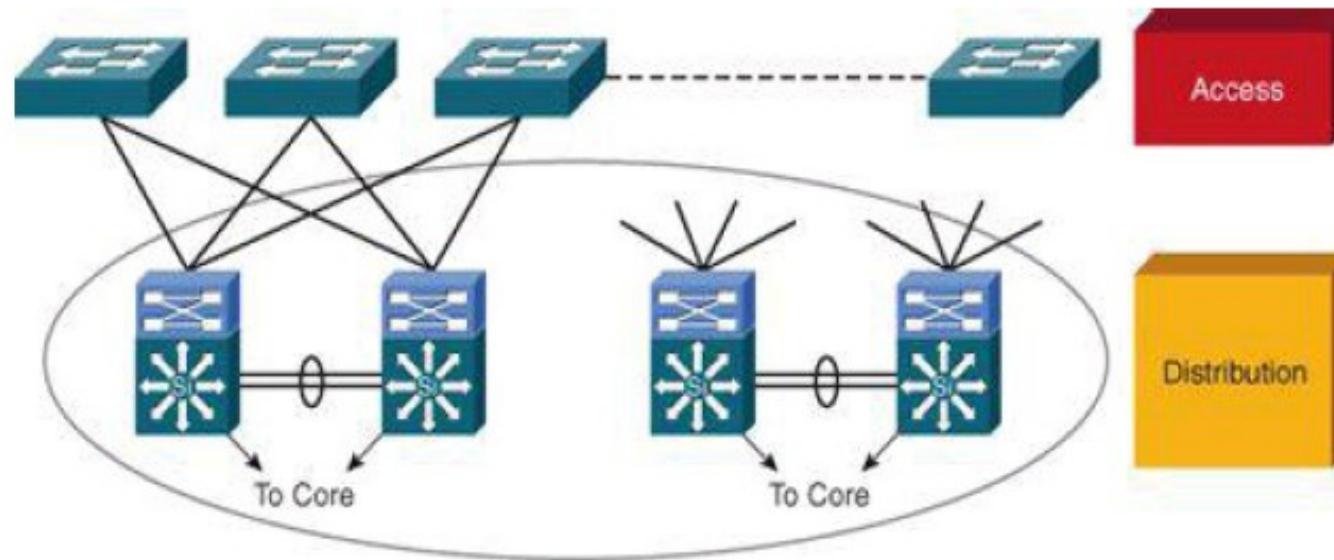
Designing the Access Layer



- High availability
 - ◆ Default gateway redundancy using multiple connections from access switches to redundant distribution layer switches.
 - ◆ Redundant power supplies.
- Other considerations
 - ◆ Convergence: the access layer should provide seamless convergence of voice into data network and providing roaming wireless LAN (WLAN).
 - ◆ Security: for additional security against unauthorized access to the network, the access layer should provide tools such as IEEE 802.1X, port security, DHCP snooping and dynamic ARP inspection (DAI).
 - ◆ Quality of service (QoS): The access layer should allow prioritization of critical network traffic using traffic classification and queuing as close to the ingress of the network as possible.
 - ◆ IP multicast: the access layer should support efficient network and bandwidth management using features such as Internet Group Management Protocol (IGMP) snooping.



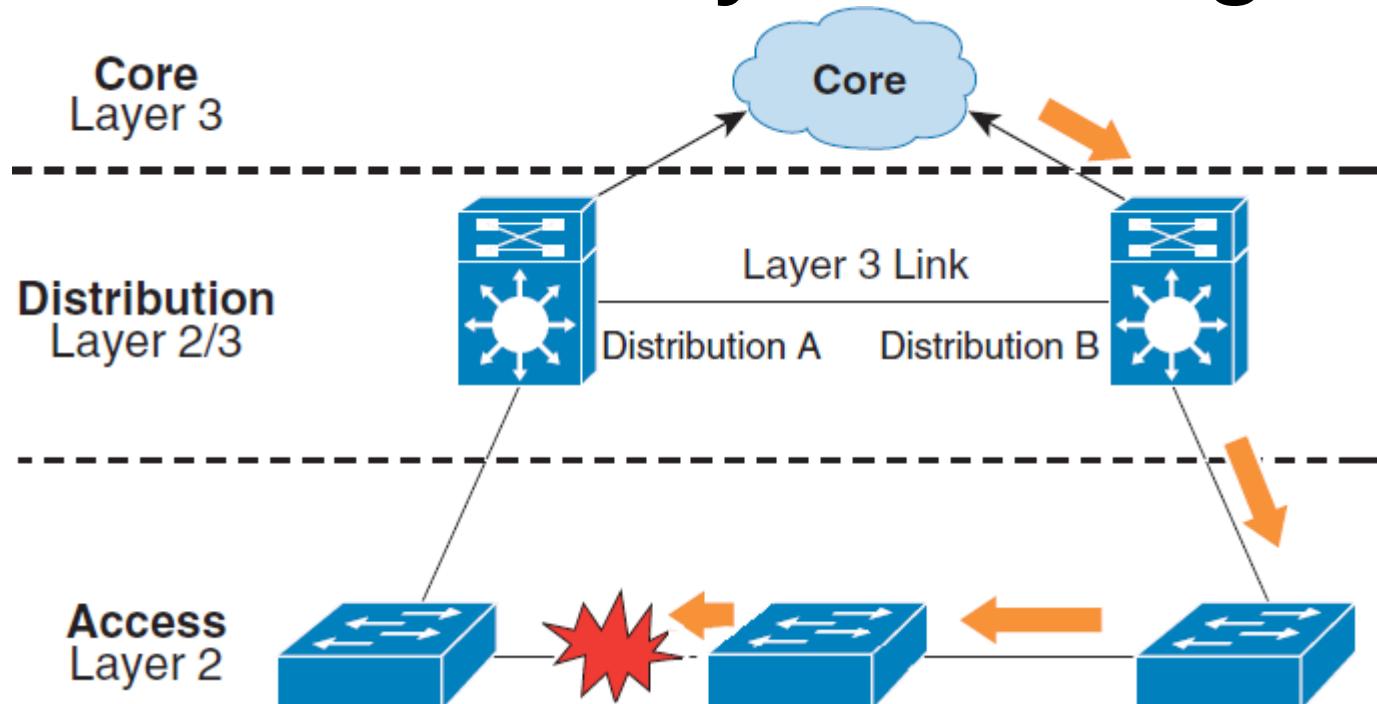
Designing the Distribution Layer



- Uses a combination of Layer 2 and multilayer switching to segment workgroups and isolate network problems, preventing them from impacting the core layer.
- Connects network services to the access layer and implements QoS, security, traffic loading balancing, and implements routing policies.
- Major design concerns: high availability, load balancing, QoS, and provisioning.
- In some networks, offers a default route to access layer routers and runs dynamic routing protocols when communicating with core routers.
- The distribution layer it is usually used to terminate VLANs from access layer switches.
- To further improve routing protocol performance, summarizes routes from the access layer.
- To implement policy-based connectivity, performs tasks such as controlled routing and filtering and QoS.



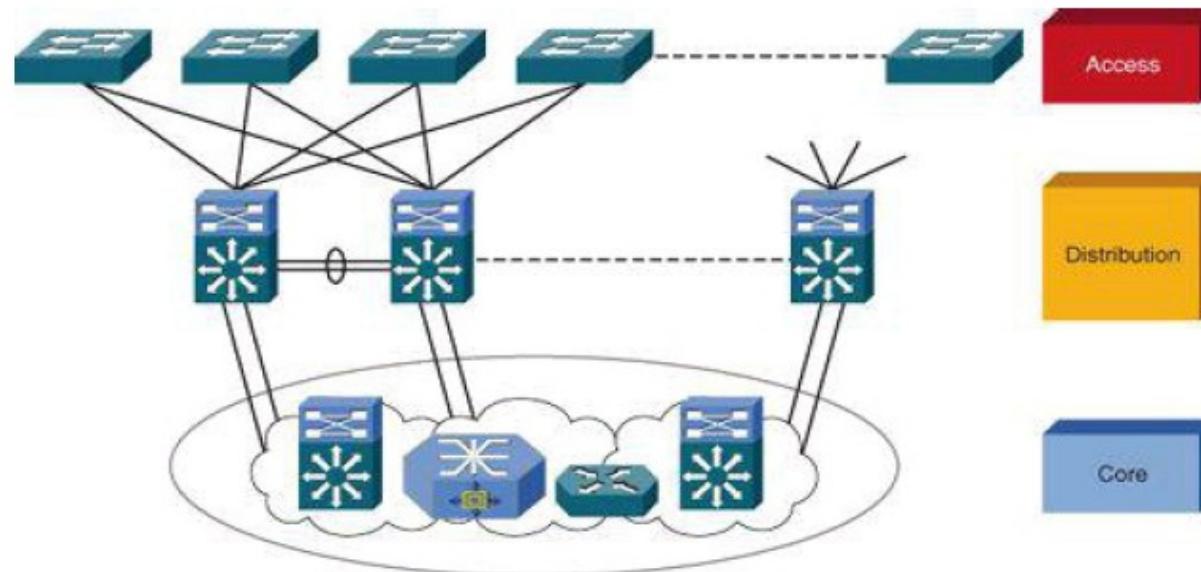
Avoid Daisy Chaining



- When using a L3 link between Distribution layer switches
 - ◆ In Access layer, any path from a switch should not require another switch from the Access layer.
 - ◆ In Distribution layer, any path between Distribution layer switches should not require a switch from the Access layer.
- When using a L2 link between Distribution layer switches
 - ◆ Daisy chain is acceptable, however
 - ◆ Could overload some Access layer switches.
 - ◆ Could increase STP convergence in case of failure.



Designing the Core Layer

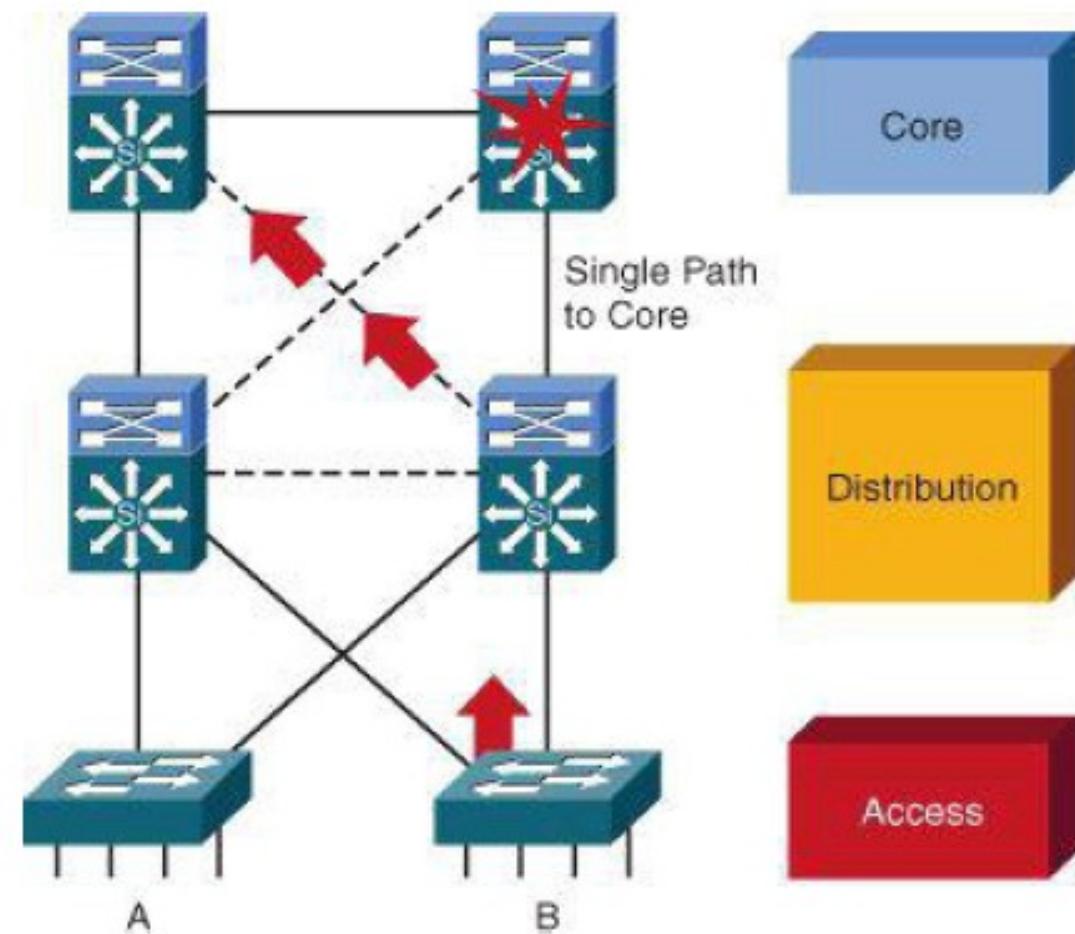


- Backbone for campus connectivity and is the aggregation point for the other layers.
- Should provide scalability, high availability, and fast convergence to the network.
 - ◆ The core layer should scale easily.
 - ◆ High-speed environment that should use hardware-acceleration, if possible.
 - ◆ The core should provide a high level of redundancy and adapt to changes quickly.
 - ◆ Core devices should be more reliable
 - ◆ Accommodate failures by rerouting traffic and respond quickly to changes in the network topology.
 - ◆ Implements scalable protocols and technologies.
 - ◆ Provides alternate paths and load balancing.
 - ◆ Packet manipulation should be avoided, such as checking access lists and filtering, which could slow down the switching of packets.
- Not all campus implementations require a campus core.
- The core and distribution layer functions can be combined at the distribution layer for a smaller campus.



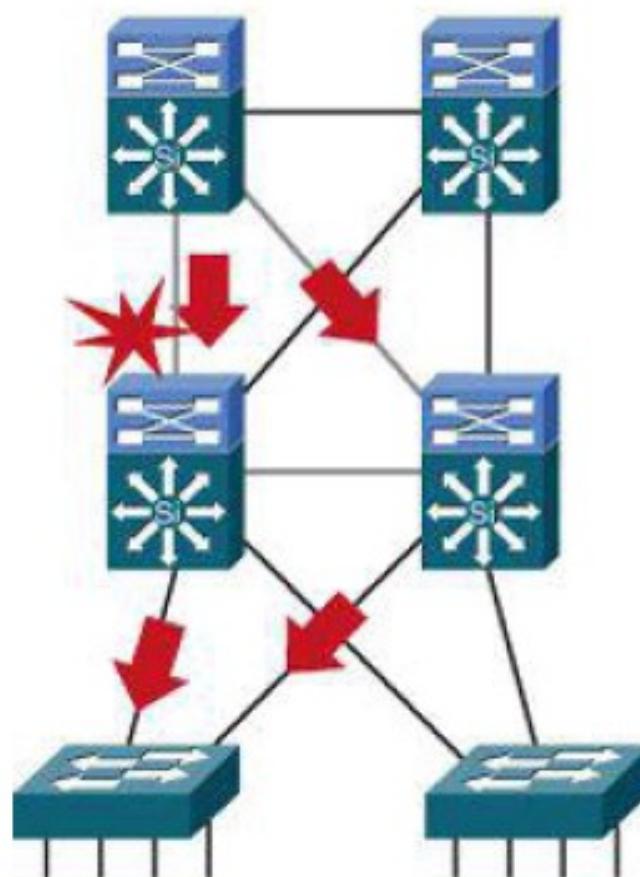
Provide Alternate Paths

- An additional link providing an alternate path to a second core switch from each distribution switch offers redundancy to support a single link or node failure.



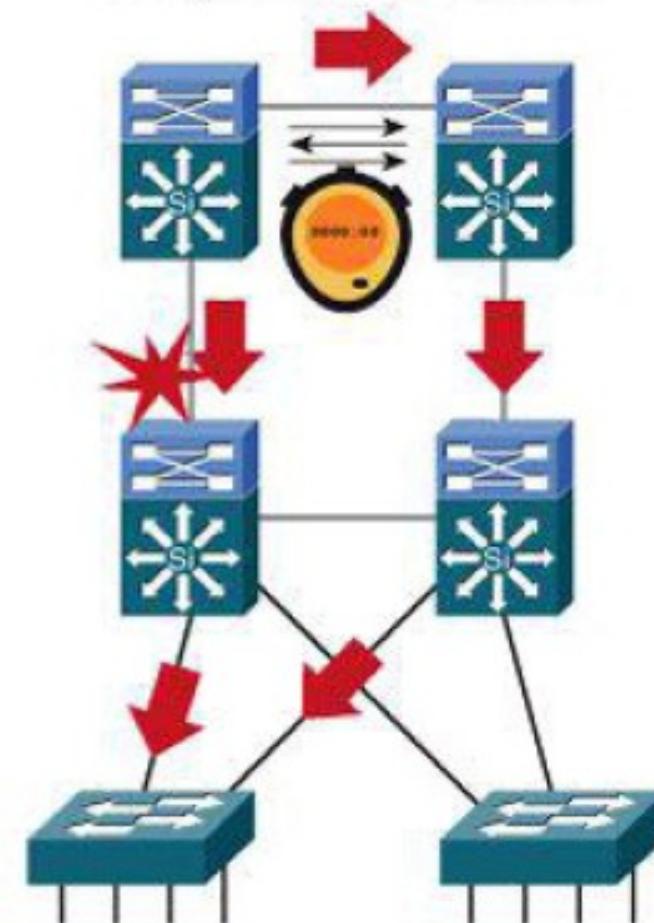
Core Redundant Triangles

Triangles: Link or box failure does *not* require routing protocol convergence.



Model A

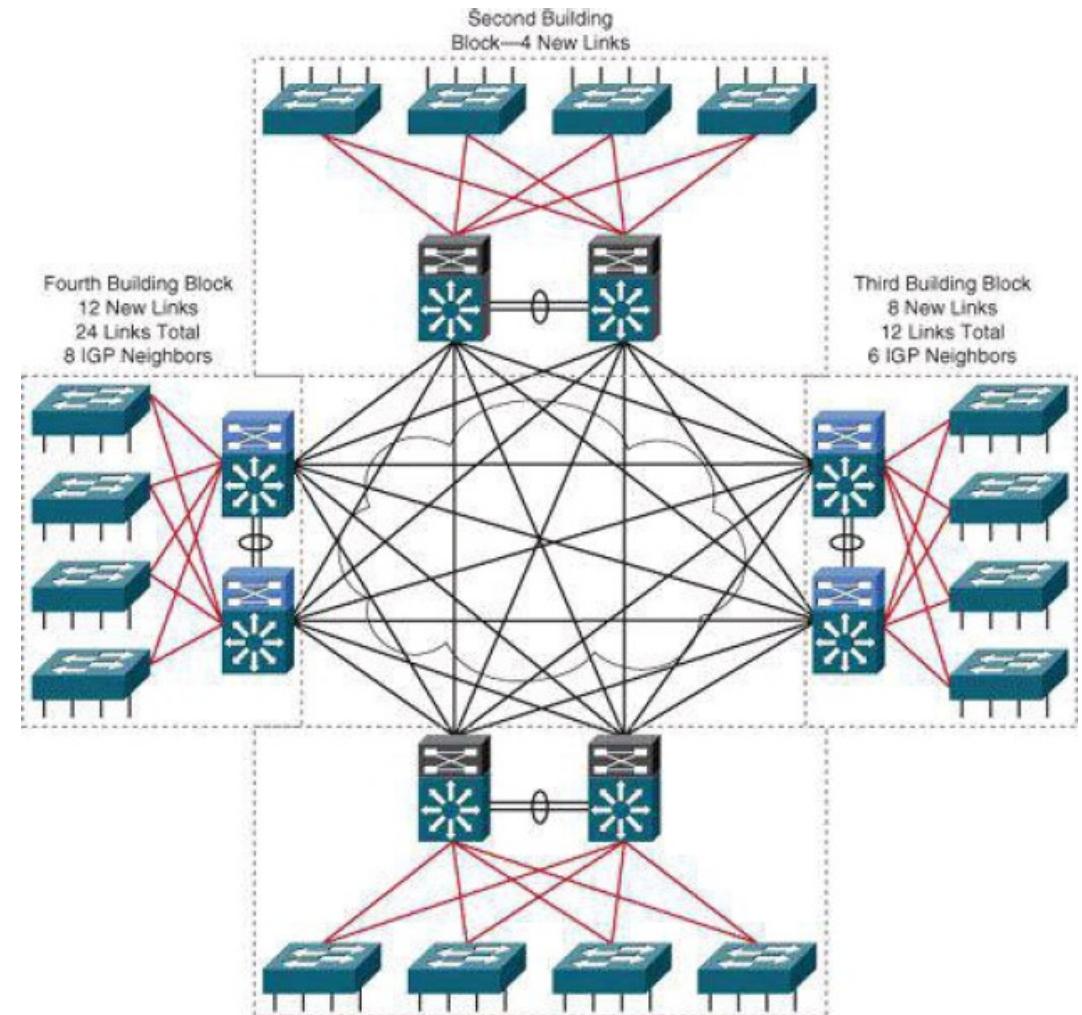
Squares: Link or box failure requires routing protocol convergence.



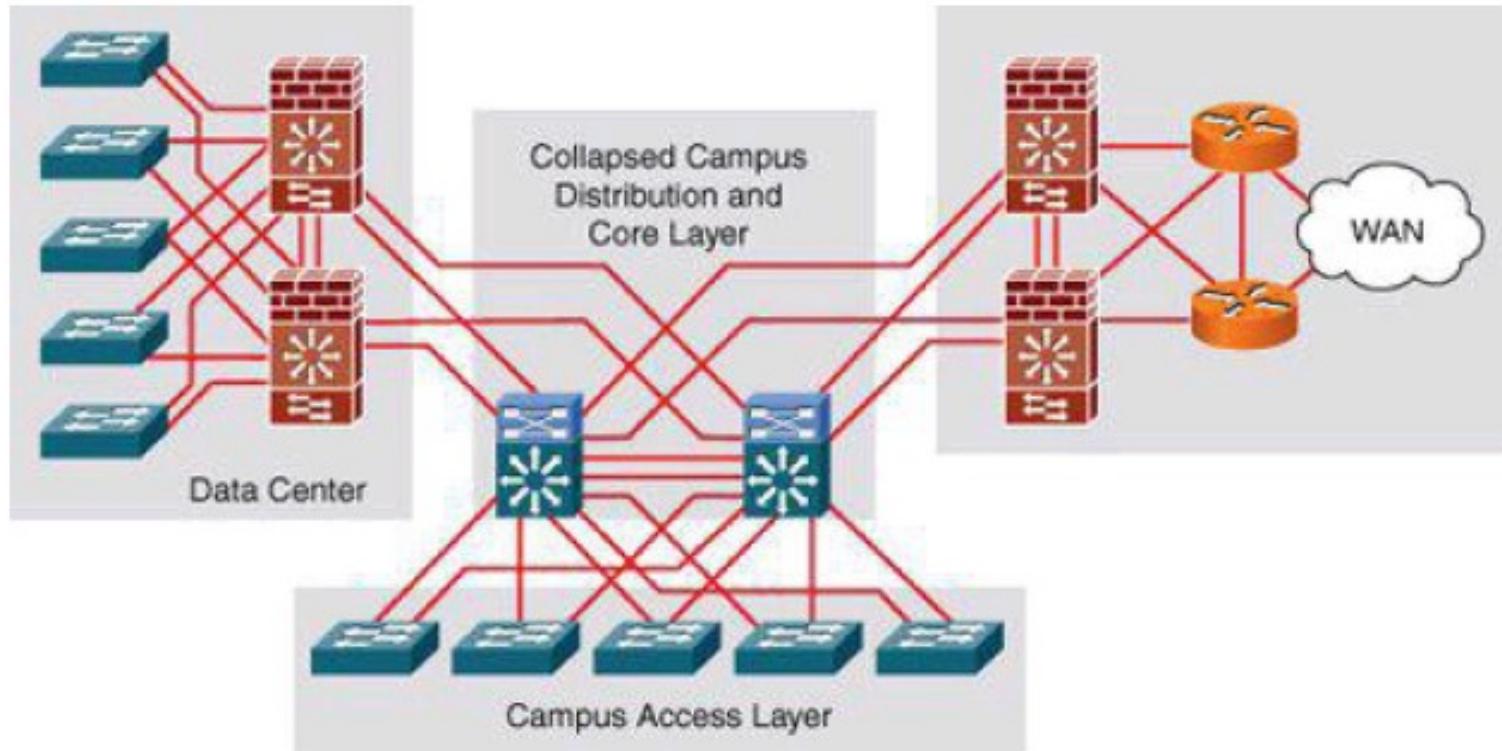
Model B

Without a Core Layer

- The distribution layer switches need to be fully meshed.
- Can be difficult to scale.
- Increases the cabling requirements.
- Routing complexity of a full-mesh design increases as new neighbors are added.
- Can be used in small campus with no perspective of growing.



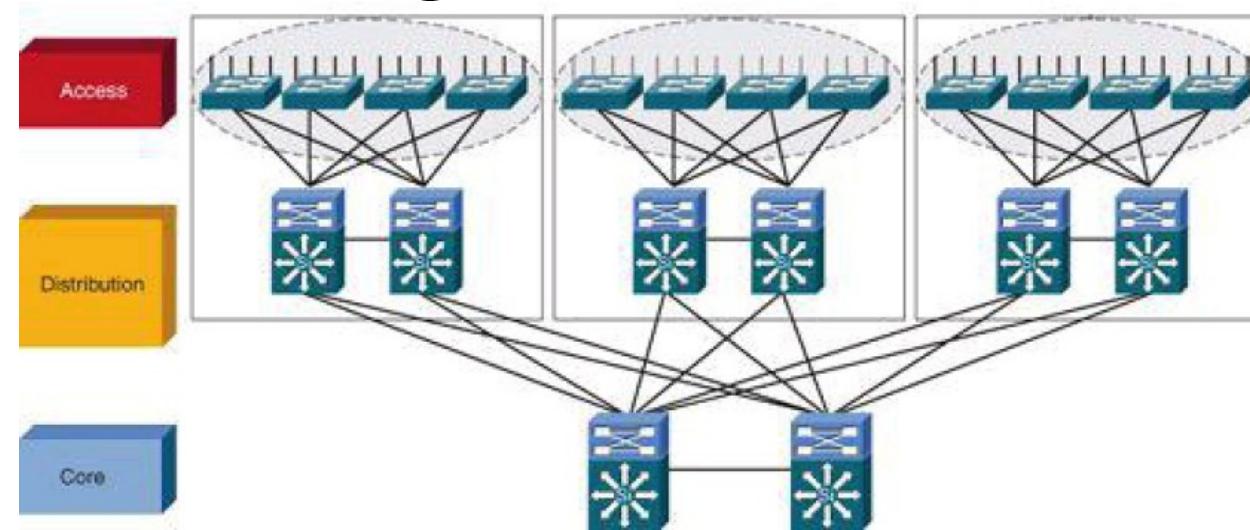
Collapsed Core Layer Architecture



- In smaller networks, the core and the distribution layer can be only one,
 - ◆ Eliminates the need for extra switching hardware and simplifies the network implementation.
- However, eliminates the advantages of the multilayer architecture, specifically fault isolation.



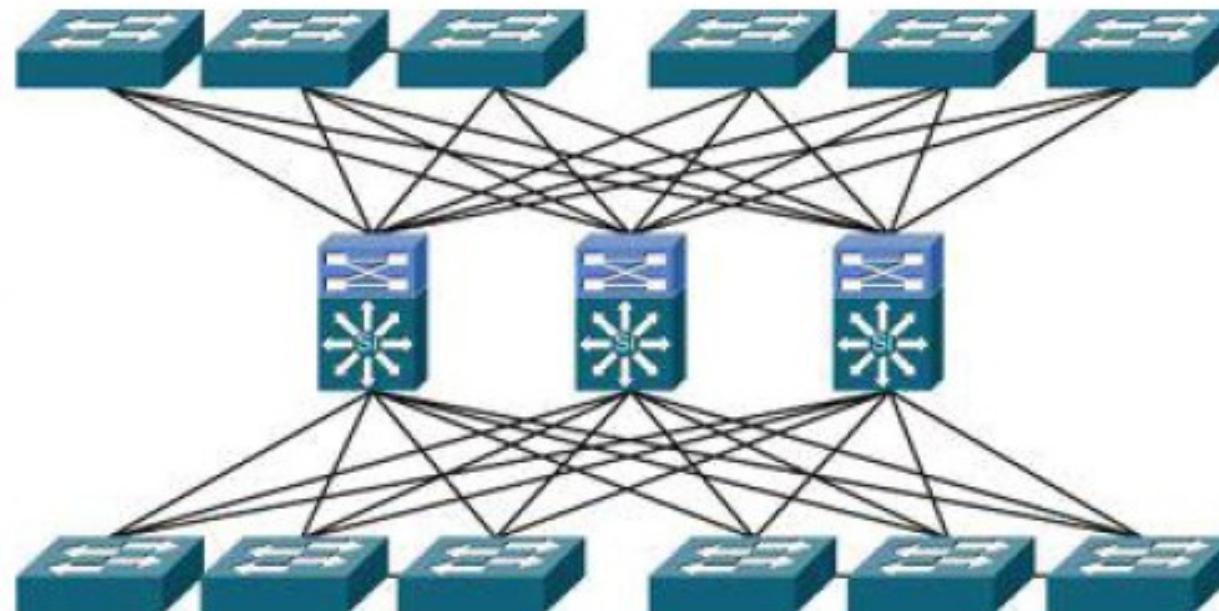
Avoid Single Points of Failure



- With an hierarchical design,
 - In Distribution and Core Layers the single points of failure are easy to avoid with redundant links.
 - Don't forget redundant power and cooling!
 - In Access Layer, all L2 switches are single points of failure (only) to the user connected to them,
 - Solution 1, redundant backup hardware activated by a (proprietary) supervision mechanism to "replace" faulty equipment.
 - Copies full configuration and state to backup hardware.
 - Solution 2, have multiple connections between each user terminal and different access switches
 - Requires multiple network cards in user terminals and more plugs/wiring.
 - Cheaper?



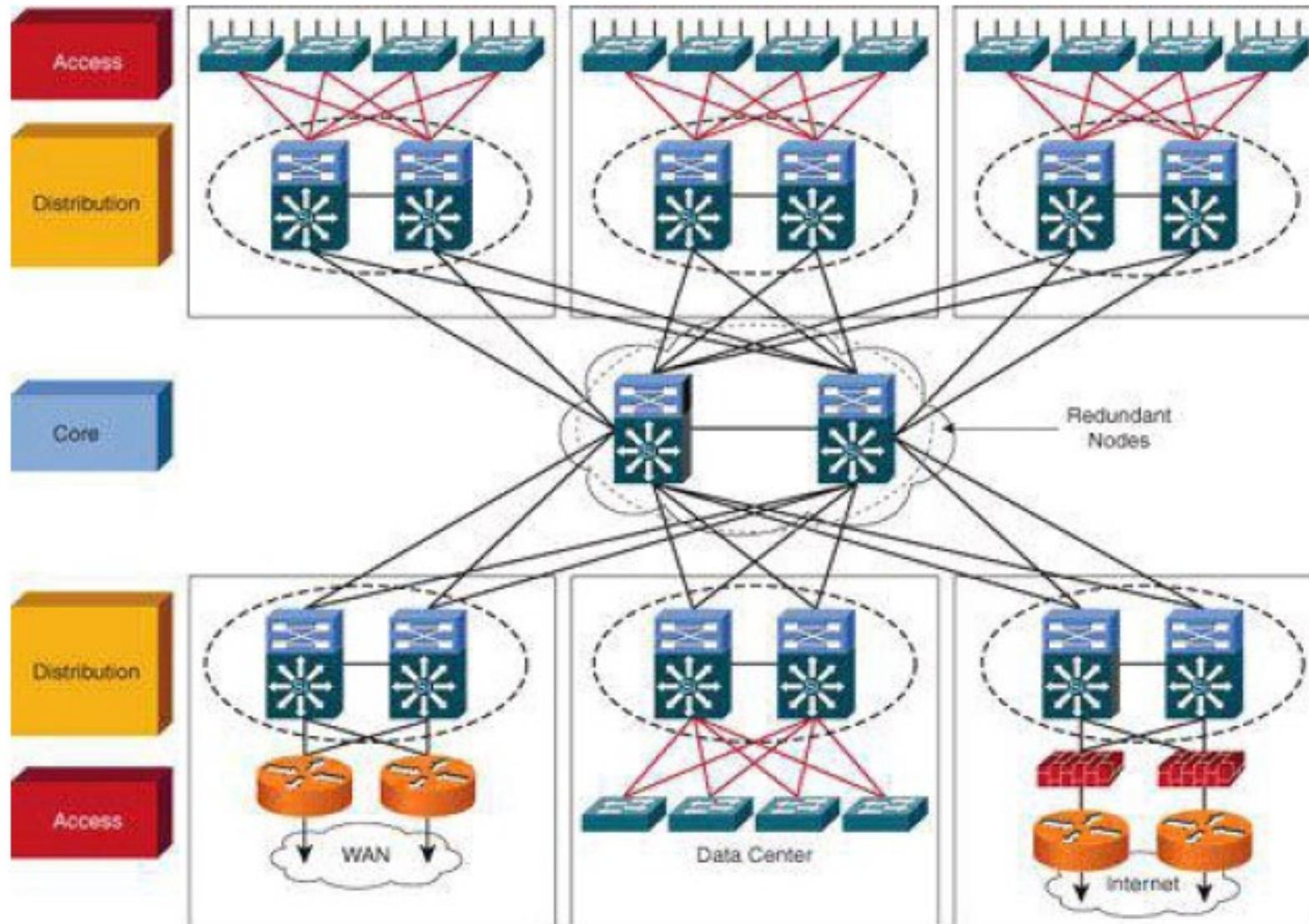
Avoid Too Much Redundancy



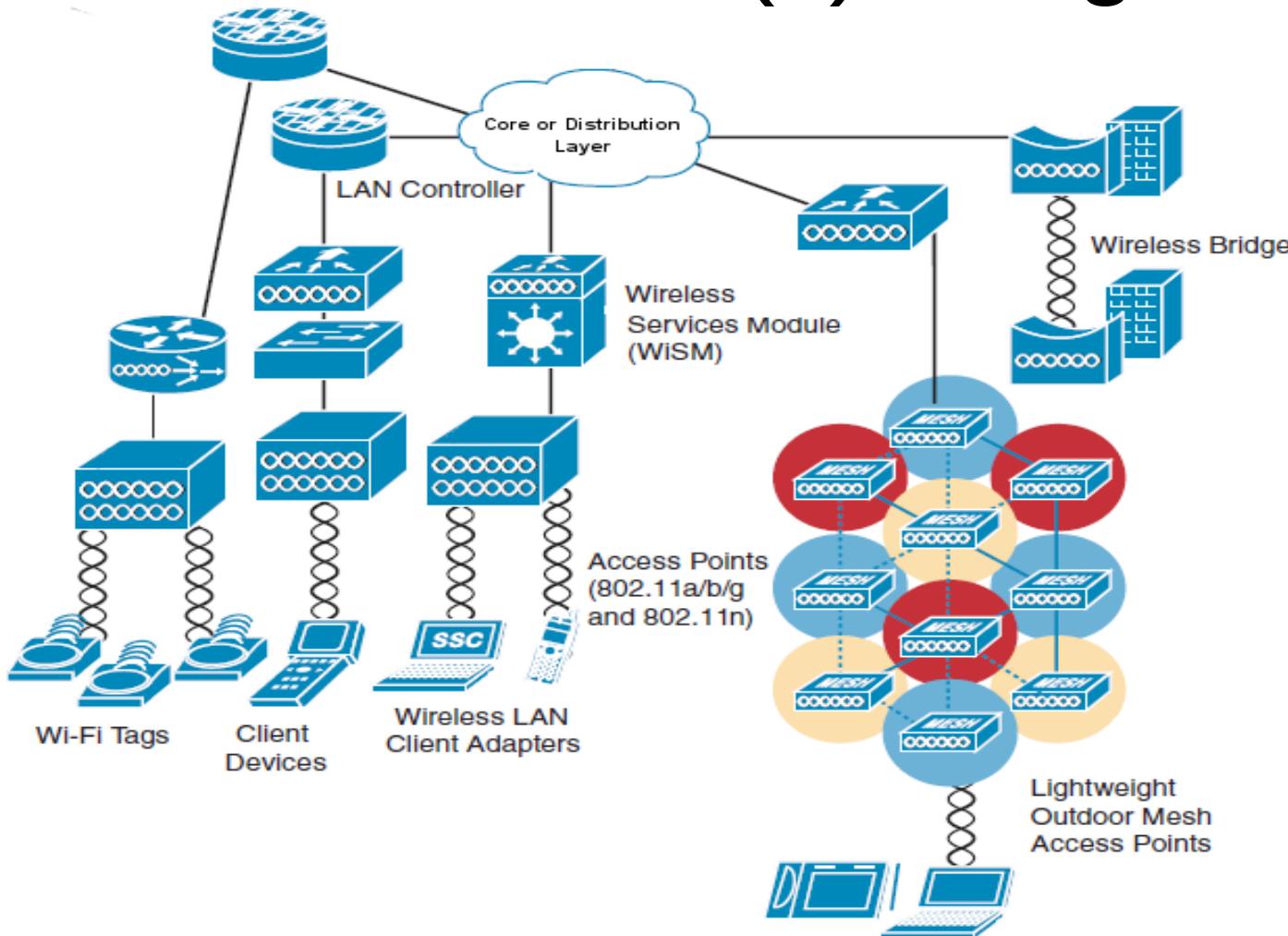
- Increases,
 - ◆ Routing complexity
 - ◆ Number of ports used
 - ◆ Wiring



Optimal Redundancy



Wireless Network(s) Integration



- Wireless networking technologies should have an integration point at core or distribution layers.
- In terms of network architecture a WLAN can be seen as any LAN.
 - ◆ Except that we have mobility and must have seamless roaming while moving.
- A large number of AP can be managed by a (Wireless) LAN Controller.



VLANs on Access Points

- AP have trunk ports to distribution/core switches.
- “Wired” VLANs must/can be extended to the wireless domain.
 - ◆ e.g., VLAN 30 “Green” and VLAN 10 “Red”.
- Each SSID can be mapped to a VLAN.
 - ◆ Different SSID/VLAN can have different security policies.
- Wireless VLANs should be configured as end-to-end.
 - ◆ Mobility and AP roaming should not break Layer 3 connectivity.
 - ◆ IP address should be the same → same VLAN with campus.
- A Native VLAN is required to provide management capability and client authentications.
 - ◆ Never extended to the wireless domain!!
 - ◆ e.g., VLAN 1.



IP Unicast Routing

Redes de Comunicações II

**Licenciatura em
Engenharia de Computadores e Informática
DETI-UA**



universidade de aveiro

deti.ua.pt

IP Routing Overview

- Routers forward packets toward destination networks.
- Routers must be aware of destination networks to be able to forward packets to them.
- A router knows about the networks directly attached to its interfaces
- For networks not directly connected to one of its interfaces, however, the router must rely on outside information.
- A router can be made aware of remote networks by:
 - ◆ **Static routing:** An administrator manually configure the information.
 - ◆ **Dynamic routing:** Learns from other routers.
 - ◆ **Policy based routing:** Manually routing rules that outweigh static/dynamic routing and may depend on parameters other than the destination address.



Default Routes

- In some circumstances, a router does not need to recognize the details of remote networks.
- The router can be configured to send all traffic (or all traffic for which there is not a more specific entry in the routing table) to a specific neighbor router.
- This is known as a default route.
- Default routes are either dynamically advertised using routing protocols or statically configured.
- IPv4 default route - 0.0.0.0/0
- IPv6 default route - ::/0

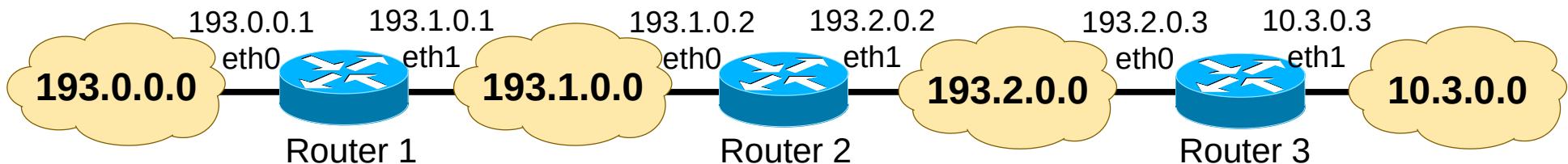


Static Routing

- Stating routing do not react to network topology changes.
 - ◆ If a link fails, the static route is no longer valid if it is configured to use that failed link, so a new static route must be configured.
 - ◆ Connectivity may be lost until intervention of an administrator.
- Static routing does not scale well when network grows.
 - ◆ Administrative burden to maintain routes may can become excessive.
- Static routes can be used in the following circumstances:
 - ◆ When the administrator needs total control over the routes used by the router.
 - ◆ When a backup to a dynamically recognized route is necessary.
 - ◆ When it is used to reach a network accessible by only one path (a stub network).
 - There is no backup link, so dynamic routing has no advantage.
 - ◆ When a router connects to its ISP and needs to have only a default route pointing toward the ISP router, rather than learning many routes from the ISP.
 - Again, a single path of access without backup.
 - ◆ When a router is underpowered and does not have the CPU or memory resources necessary to handle a dynamic routing protocol.
 - ◆ When it is undesirable to have dynamic routing updates forwarded across low bandwidth links.



Static Routing Examples



- Example 1

- Router2 do not know networks 193.0.0.0/24 and 10.3.0.0/24
- Necessary static routes:
 - 193.0.0.0/24 accessible through 193.1.0.1 (eth1, Router1)
 - 10.3.0.0/24 accessible through 193.2.0.3 (eth0, Router3)

- Example 2

- Router1 do not know networks 193.2.0.0/24 and 10.3.0.0/24
- Necessary static routes:
 - 193.2.0.0/24 accessible through 193.1.0.2 (eth0, Router2)
 - 10.3.0.0/24 accessible through 193.1.0.2 (eth0, Router2)
- OR
- Using default route: 0.0.0.0/0 accessible through 193.1.0.2 (eth0, Router2)



Dynamic Routing

- Dynamic routing allows the network to adjust to changes in the topology automatically, without administrator involvement.
- Routers exchange information about the reachable networks and the state of each network/link.
 - ◆ Routers exchange information only with other routers running the same routing protocol.
 - ◆ When the network topology changes, the new information is dynamically propagated throughout the network, and each router updates its routing table to reflect the changes.



(Complex) Routing Tables

- An IP address may have multiple matches on a Routing Table:
 - ◆ Example: 192.168.1.12
 - ◆ Will match:
 - 192.168.1.0/25 via ...
 - 192.168.1.0/24 via ...
 - 192.168.0.0/23 via ...
 - 192.168.0.0/16 via ...
 - ...
 - ◆ Router will choose entry with the largest network prefix (most specific network).
 - i.e., 192.168.1.0/25 via ...
- Load balancing
 - ◆ Routing tables may have more than one path for each network
 - ◆ Traffic will be divided by all entries.
 - ◆ By packet, flow (TCP session, UDP IPs/port), etc...
 - E.g, packet 1 path 1, packet 2 path 2, packet 3 path 1, ...
 - Flow 1 path 1, flow 2 path 2, flow 3 path 3, flow 4 path 1, flow 5 path 2, ...



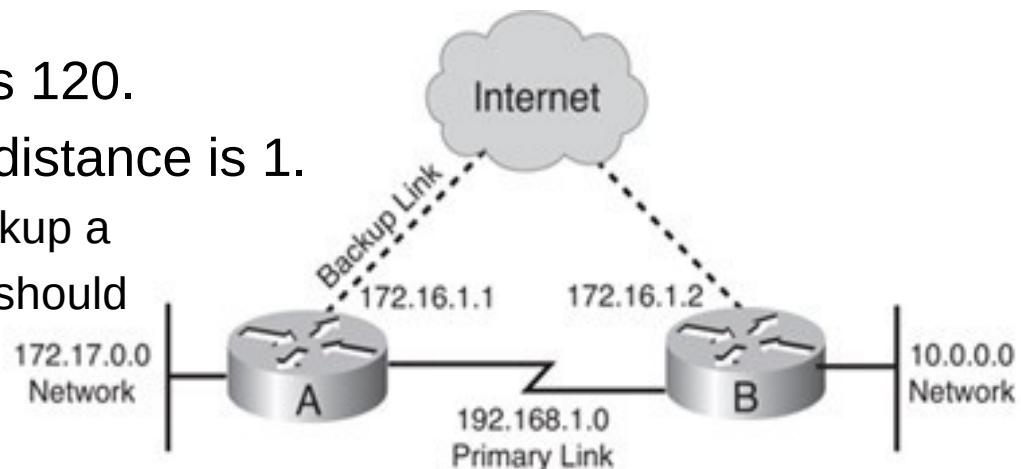
Administrative Distance

- Most routing protocols have metric structures and algorithms that are incompatible with other protocols.
- It is critical that a network using multiple routing protocols be able to seamlessly exchange route information and be able to select the best path across multiple protocols.
- Routers use a value called administrative distance to select the best path when they learn from different routing protocols the same destination (same network prefix and mask length).
- The Protocol/Method with the lowest Administrative Distance is preferred
 - ◆ The Administrative Distance value is configurable.
- Example:
 - ◆ Static [1/1] 192.168.1.0/24 via ... ← Chosen!
 - ◆ RIP [120/1] 192.168.1.0/24 via ...
 - ◆ OSPF [110/1] 192.168.1.0/24 via ...



Floating Static Routes

- Based on the default administrative distances, routers use static routes over any dynamically learned route.
 - ◆ However, this default behavior might not be the desired behavior.
 - ◆ For example, when you configure a static route as a backup to a dynamically learned route, you do not want the static route to be used as long as the dynamic route is available.
- A static route that appears in the routing table only when the primary route goes away is called a floating static route.
 - ◆ The administrative distance of the static route is configured to be higher than the administrative distance of the primary route and it “floats” above the primary route, until the primary route is no longer available.
- RIP default administrative distance is 120.
- Static Routes default administrative distance is 1.
 - ◆ To create a floating static route (to backup a RIP route) the administrative distance should be greater than 120.
 - ◆ In example: 200.

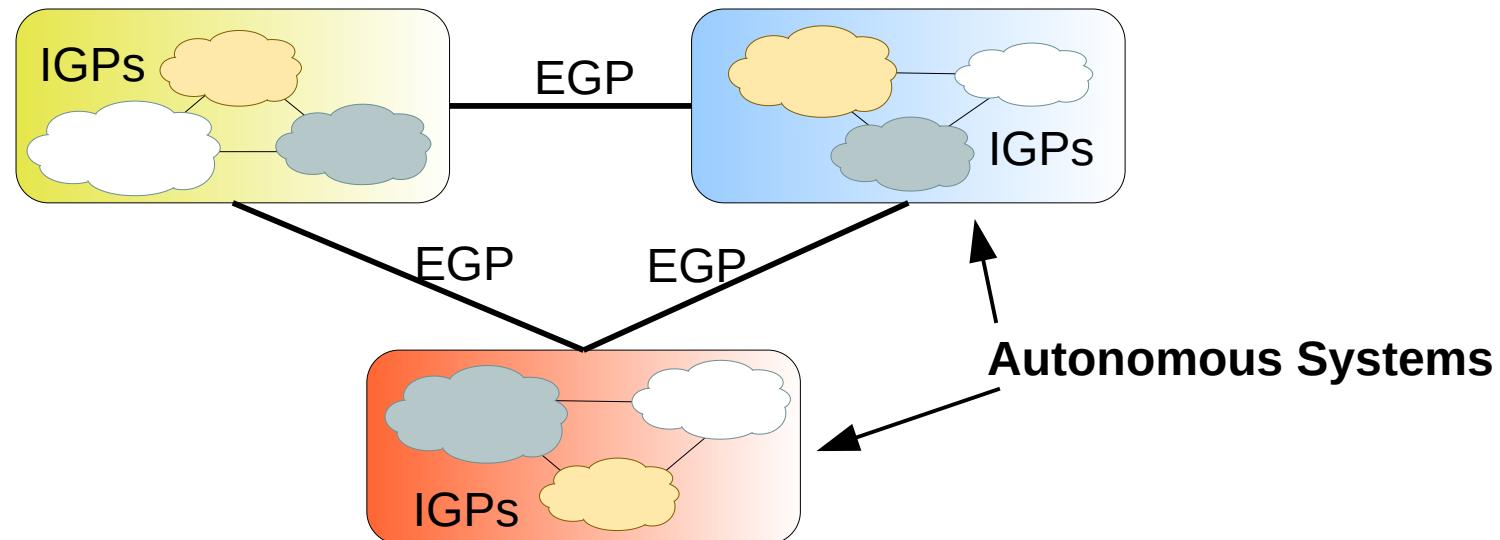


```
ip route 10.0.0.0 255.0.0.0 172.16.1.2 200
router rip
  network 192.168.1.0
  network 172.17.0.0
```

```
ip route 172.17.0.0 255.255.0.0 172.16.1.1 200
router rip
  network 192.168.1.0
  network 10.0.0.0
```

Autonomous Systems

- AS (Autonomous System) – set of routers/networks with a common routing policy and under the same administration.
- Routing inside an AS is performed by IGPs (Interior Gateway Protocols) such as RIPv1, RIPv2, OSPF, IS-IS and EIGRP.
 - ◆ Called Internal Routing
- Routing between AS is performed by EGPs (Exterior Gateway Protocols) such as BGP.
- IGPs and EGPs have different objectives:
 - ◆ IGPs: optimize routing performance
 - ◆ EGPs: optimize routing performance obeying political, economic and security policies.



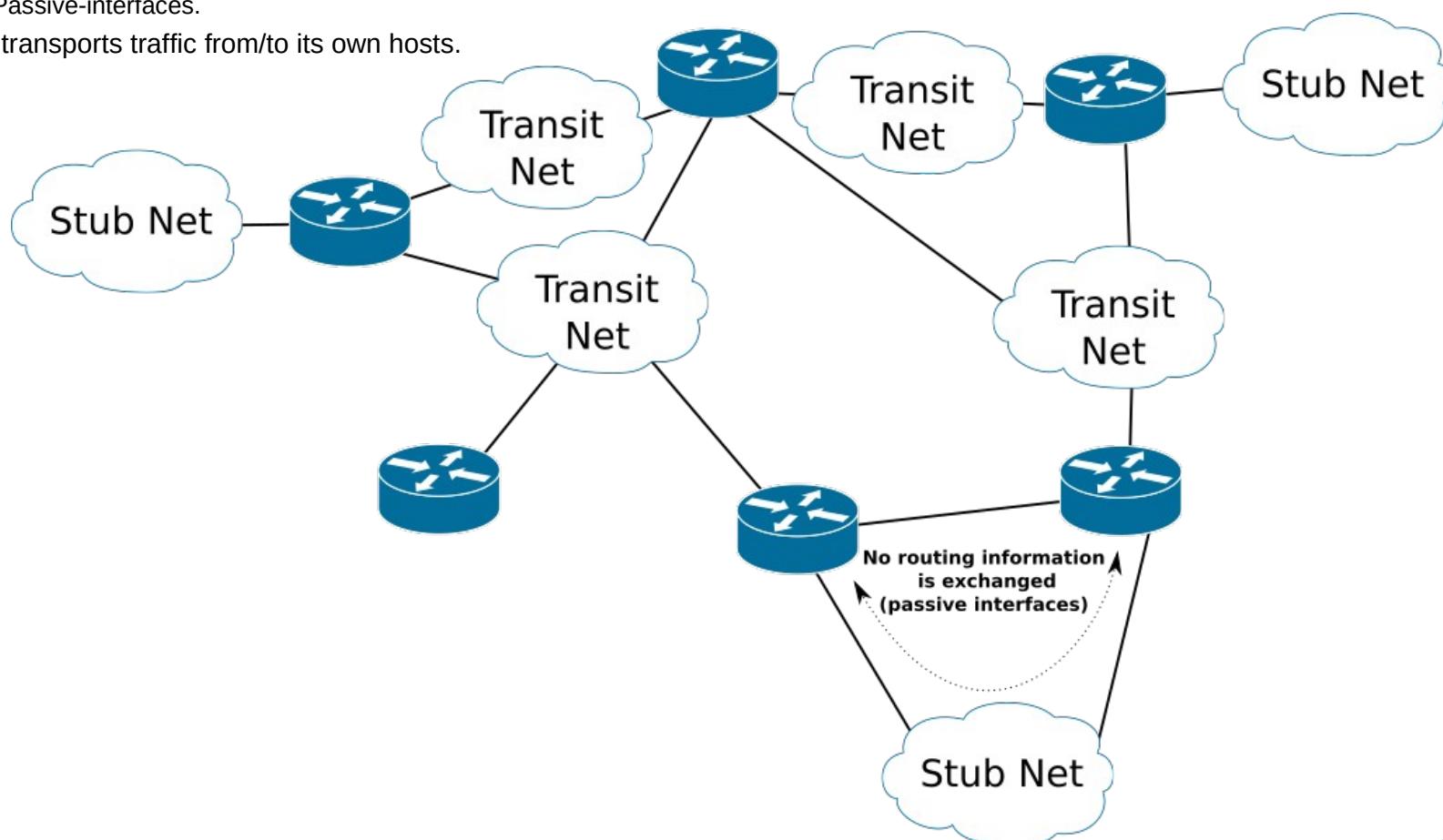
Type of Networks

Transit/Transport

- Used to interconnect networks.
 - Routers exchange routing information using it.
- Transports traffic from/to other network hosts and from/to its own hosts.

Stub

- Single router network.
- or multiple routers network, if routers do not exchange routing information.
 - Passive-interfaces.
- Only transports traffic from/to its own hosts.



Distance Vector versus Link State Protocols

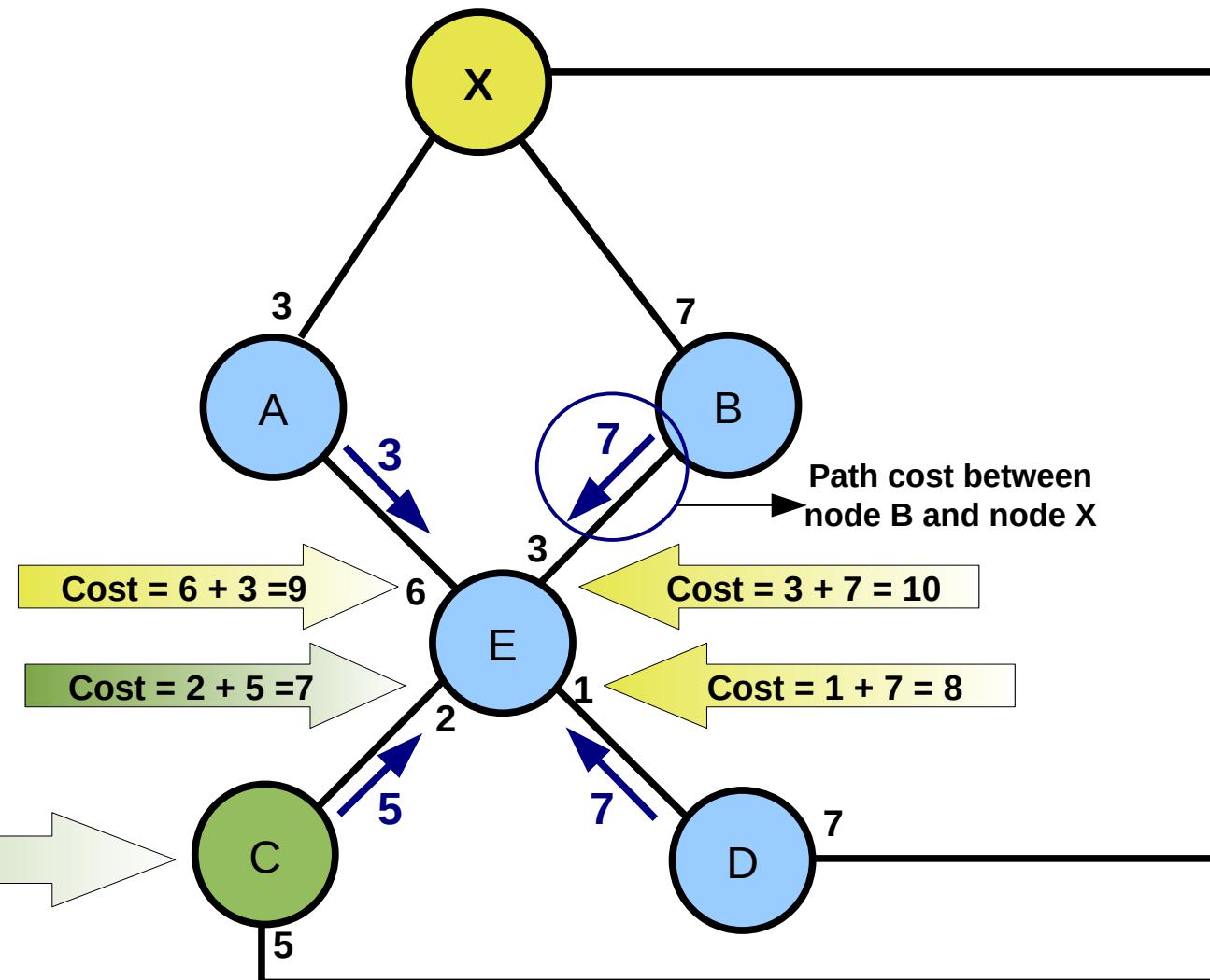
- Distance vector
 - ◆ Each routers learns networks and best path based on the information sent periodically by its neighbors.
 - ◆ Network and cost (distance) to that network.
 - ◆ Each router determines the shortest paths to all know networks based on a distributed and asynchronous version of the Bellman-Ford algorithm.
 - ◆ Examples: RIPv1, RIPv2, IGRP, EIGRP.
- Link state
 - ◆ Routers learn the complete network topology and use a centralized algorithm to determine the shortest paths to all known networks.
 - ◆ The information necessary to construct and maintain in each router a data base with the network topology is obtain by a flooding process.
 - ◆ Network information is only exchanged on bootstrap and after any topology change.
 - ◆ Examples: OSPF, IS-IS.



Distributed and Asynchronous Bellman-Ford Algorithm

- Each node periodically transmits to its neighboring nodes (its estimation of) the cost to reach a destination node.
- Each node recalculates its own estimation of the cost to reach a destination node
 - ◆ Adds the received estimated cost to the destination to the cost of the connection/port where it received the neighbor information.
 - ◆ Chooses the lowest cost.

Neighbor chosen by node E to route traffic to node X



RIP (Routing Information Protocol)

- Is a *distance vector* protocol
 - ◆ Each router maintains a list of known networks and, for each network, an estimation of the cost to reach it – this is called a distance vector.
 - ◆ Each router periodically send to its neighboring routers its own distance vector (partially or complete) – announcement/update.
 - ◆ Each router uses the distance vector sent by its neighbors to update its own distance vector.
- The path cost to a destination is given by the number of routers/hops in the path.
 - ◆ Maximum cost is 15.
 - ◆ A cost of 16 is considered infinite (or unattainable destination).
- Each router determines the entries in its own routing table, based on the constructed distance vector.
 - ◆ For each destination (network) learned, it adds an entry to that network that uses the path (or paths) with the lowest cost, using as next-hop the neighboring router(s) that announced that network with that lowest cost path.



RIP Version 1

- RIP Version 1 (RIPv1) is a classfull protocol.
 - ◆ Does not announces (sub-)networks masks, only network prefixes.
 - ◆ Network masks are assumed based on the incoming interface mask.
 - ◆ If all networks have the same mask it works perfectly, however, when networks with different masks exist it is problematic.
- RIPv1 uses the broadcast address 255.255.255.255 to send announcements/updates.
 - ◆ All network devices must process the packets.
- Does not support authentication.
 - ◆ Messages may be forged by an attacker.



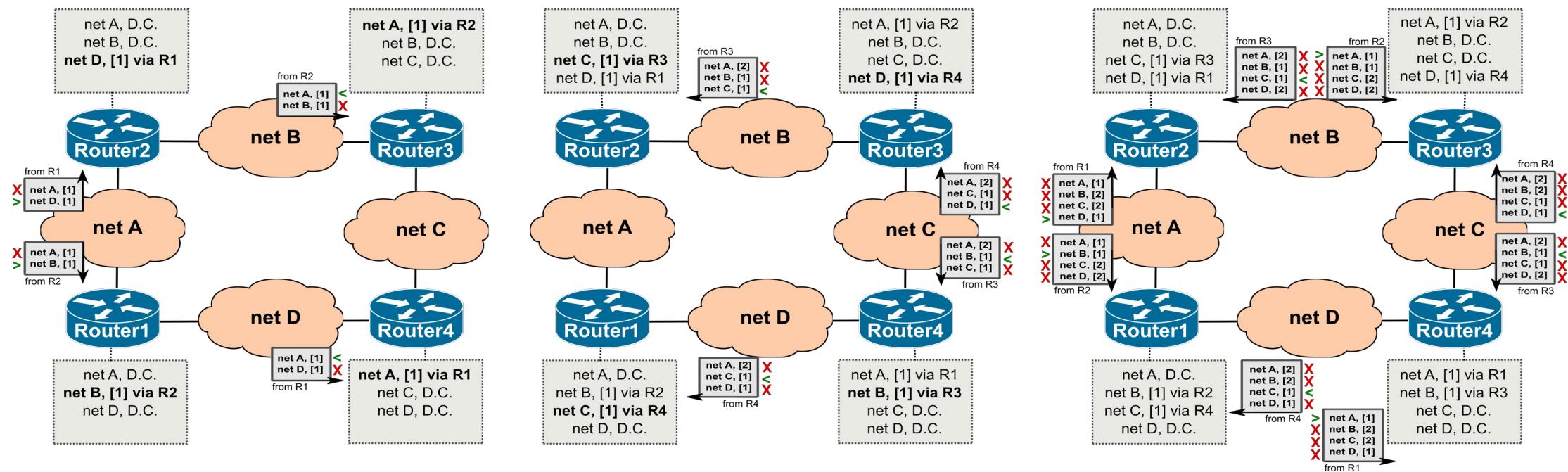
RIP Version 2

- RIP Version 2 (RIPv2) is a *classless* protocol.
 - ◆ RIPv2 announcements include network prefix and mask.
 - ◆ Supports variable length masks.
- RIPv2 used the multicast address 224.0.0.9 to send announcements/updates only to routers running RIPv2.
- RIPv2 supports authentication using message-digest and clear text password.
 - ◆ Clear text password authentication should not be used!



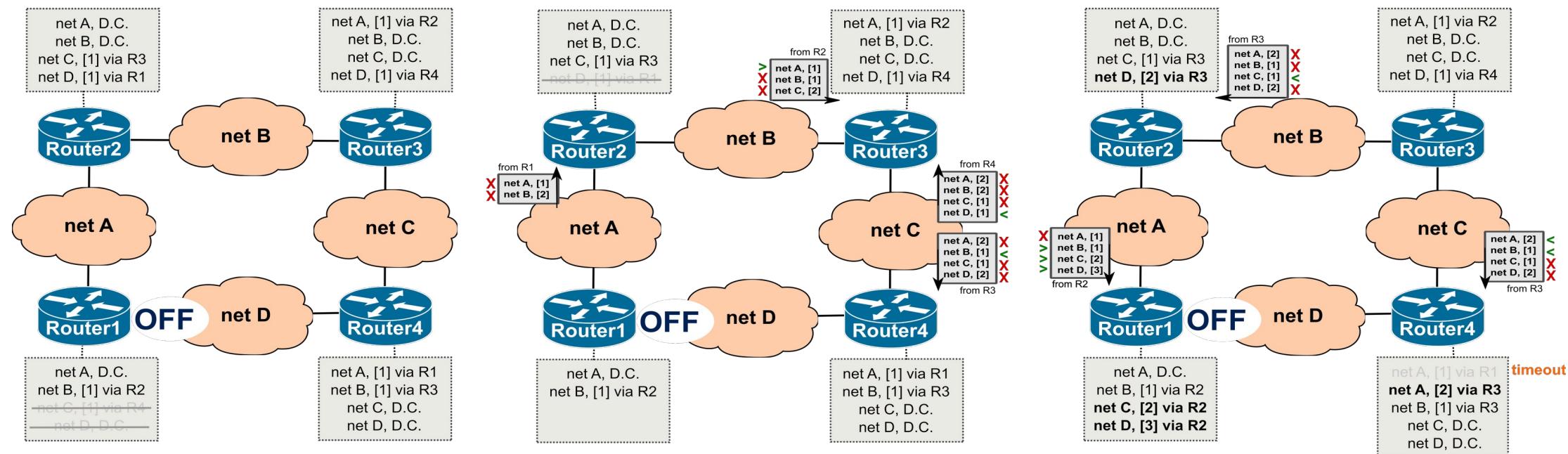
RIP Algorithm (1)

- Assuming that Router1 and Router2 send announcements first.
 - With split-horizon disabled.



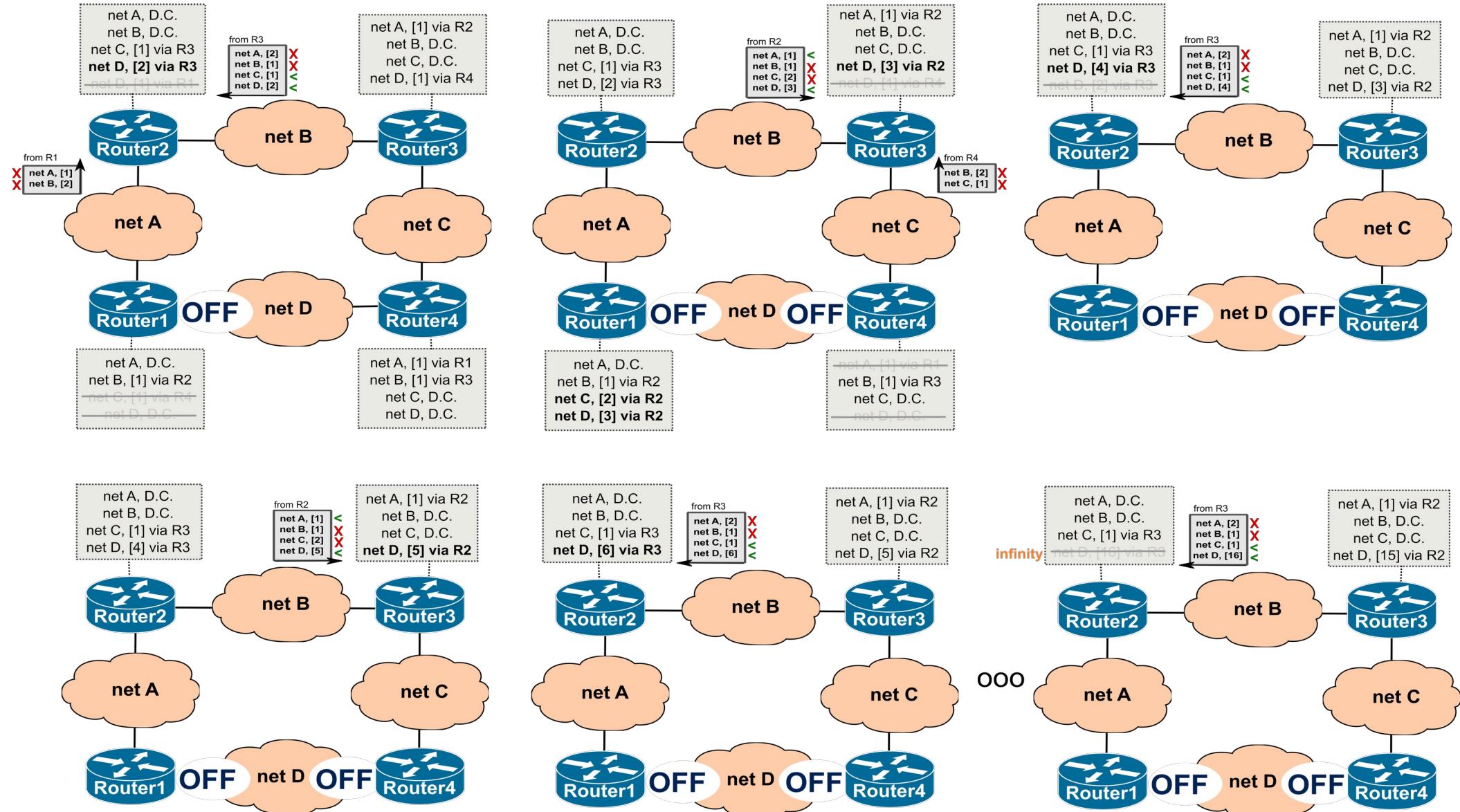
RIP Algorithm (2)

- Assuming Router1 connection to network D goes down.
 - No triggered updates.



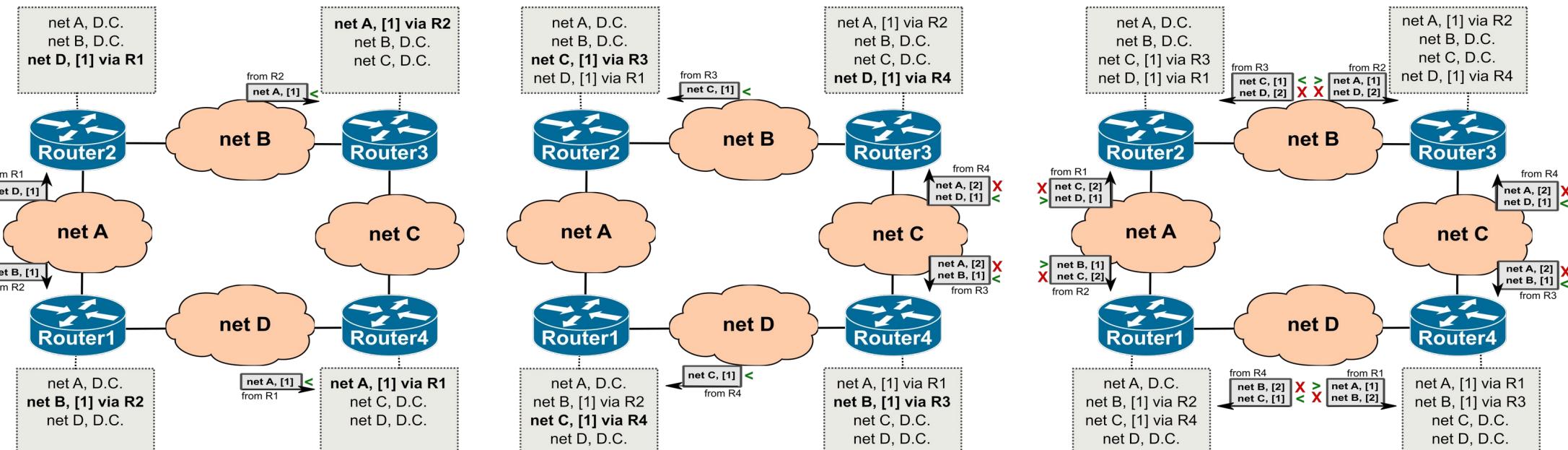
Count to Infinity Problem

- When multiple failures occur before algorithm convergence!



Split-Horizon (1)

- Solution for the count to infinity problem.
- Each Router, in each interface, announces only the networks in which that interface is not used to provide the best path to that destination.
- Split horizon lowers the convergence time of the routing tables when there is a topology change.
 - ◆ RIPv1 e RIPv2 supports it.
- Assuming Router1 and Router2 start sending announcements first:

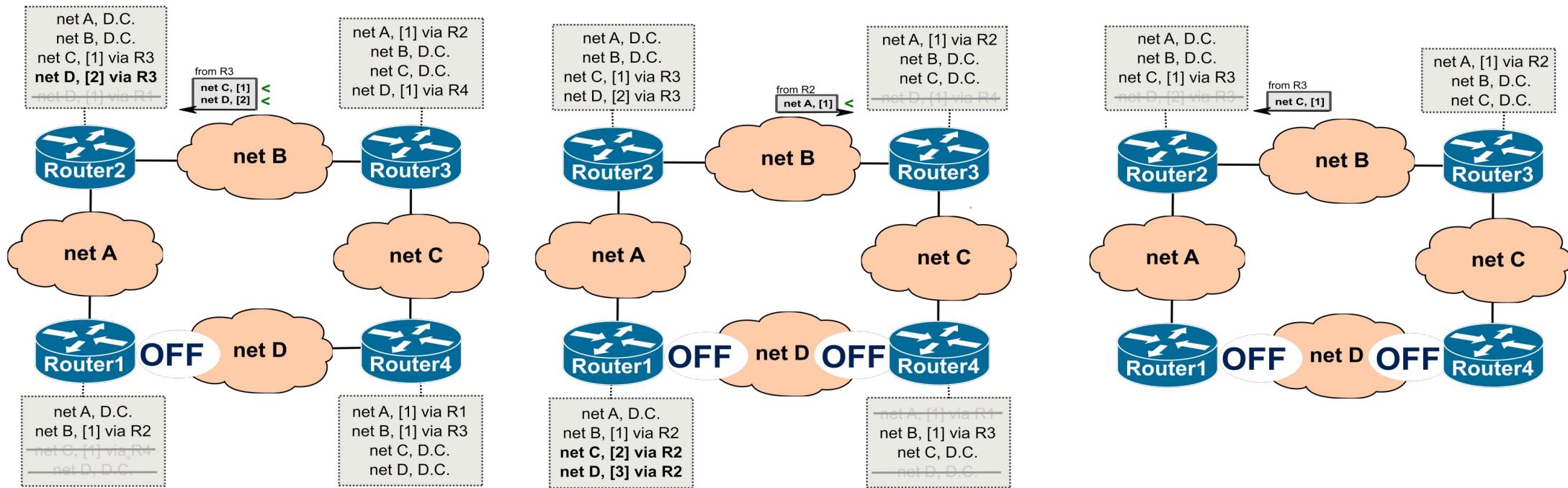


- In Split horizon with Poisoned Reverse, routers announce all networks but set metric to infinity (16) for networks learned by the interface by which they are sending the announcement.
 - ◆ Larger update messages.



Split-Horizon (2)

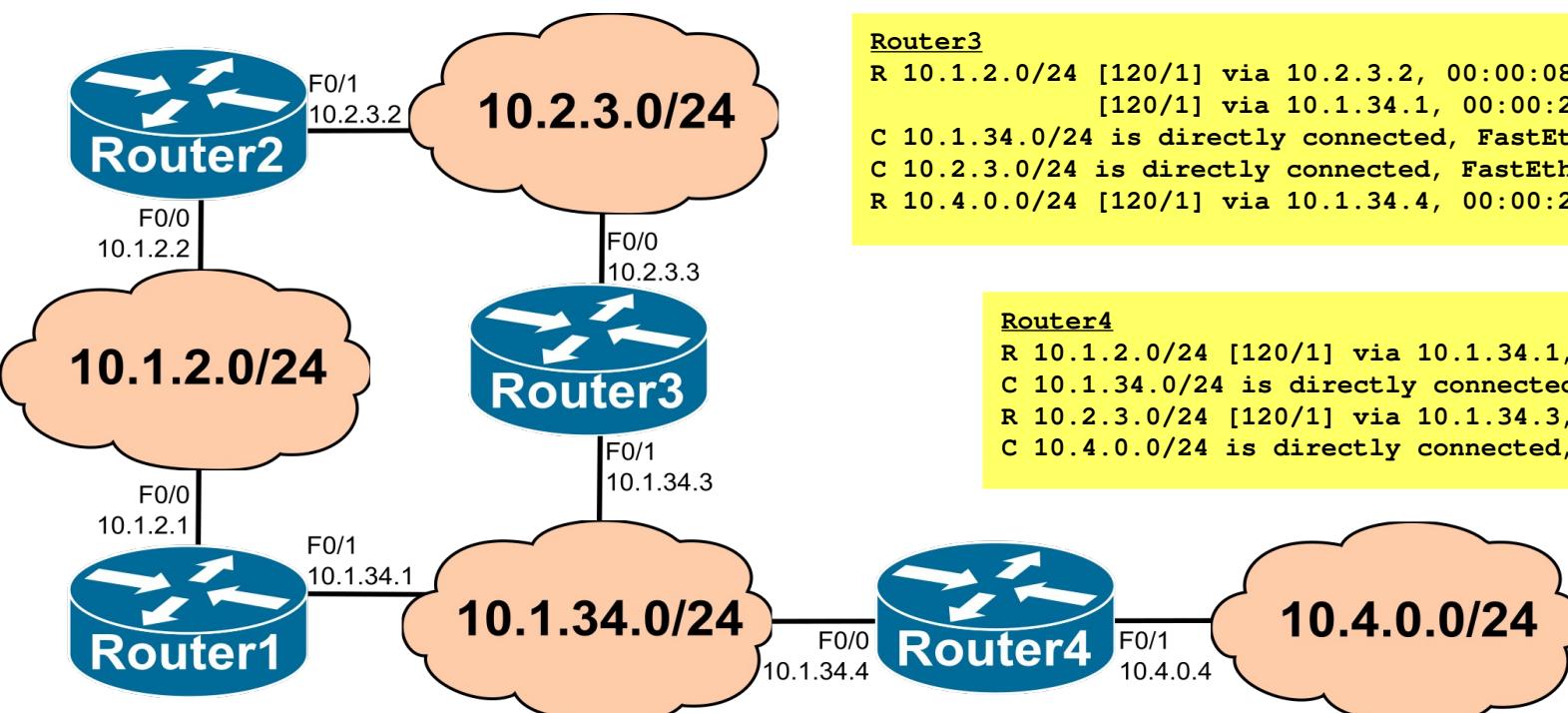
- Solution for the count to infinity problem.
- Prevents any routing loops that involve two routers.
 - ◆ It is possible to end up with patterns in which three or more routers are engaged in mutual deception.
- Assuming Router1 and Router4 loose connection to network D almost simultaneously:



Routing Tables with RIP

Router2

```
C 10.1.2.0/24 is directly connected, FastEthernet0/0
R 10.1.34.0/24 [120/1] via 10.2.3.3, 00:00:21, FastEthernet0/1
          [120/1] via 10.1.2.1, 00:00:11, FastEthernet0/0
C 10.2.3.0/24 is directly connected, FastEthernet0/1
R 10.4.0.0/24 [120/2] via 10.2.3.3, 00:00:21, FastEthernet0/1
          [120/2] via 10.1.2.1, 00:00:11, FastEthernet0/0
```



Router3

```
R 10.1.2.0/24 [120/1] via 10.2.3.2, 00:00:08, FastEthernet0/0
          [120/1] via 10.1.34.1, 00:00:28, FastEthernet0/1
C 10.1.34.0/24 is directly connected, FastEthernet0/1
C 10.2.3.0/24 is directly connected, FastEthernet0/0
R 10.4.0.0/24 [120/1] via 10.1.34.4, 00:00:24, FastEthernet0/1
```

Router4

```
R 10.1.2.0/24 [120/1] via 10.1.34.1, 00:00:18, FastEthernet0/0
C 10.1.34.0/24 is directly connected, FastEthernet0/0
R 10.2.3.0/24 [120/1] via 10.1.34.3, 00:00:29, FastEthernet0/0
C 10.4.0.0/24 is directly connected, FastEthernet0/1
```

Router1

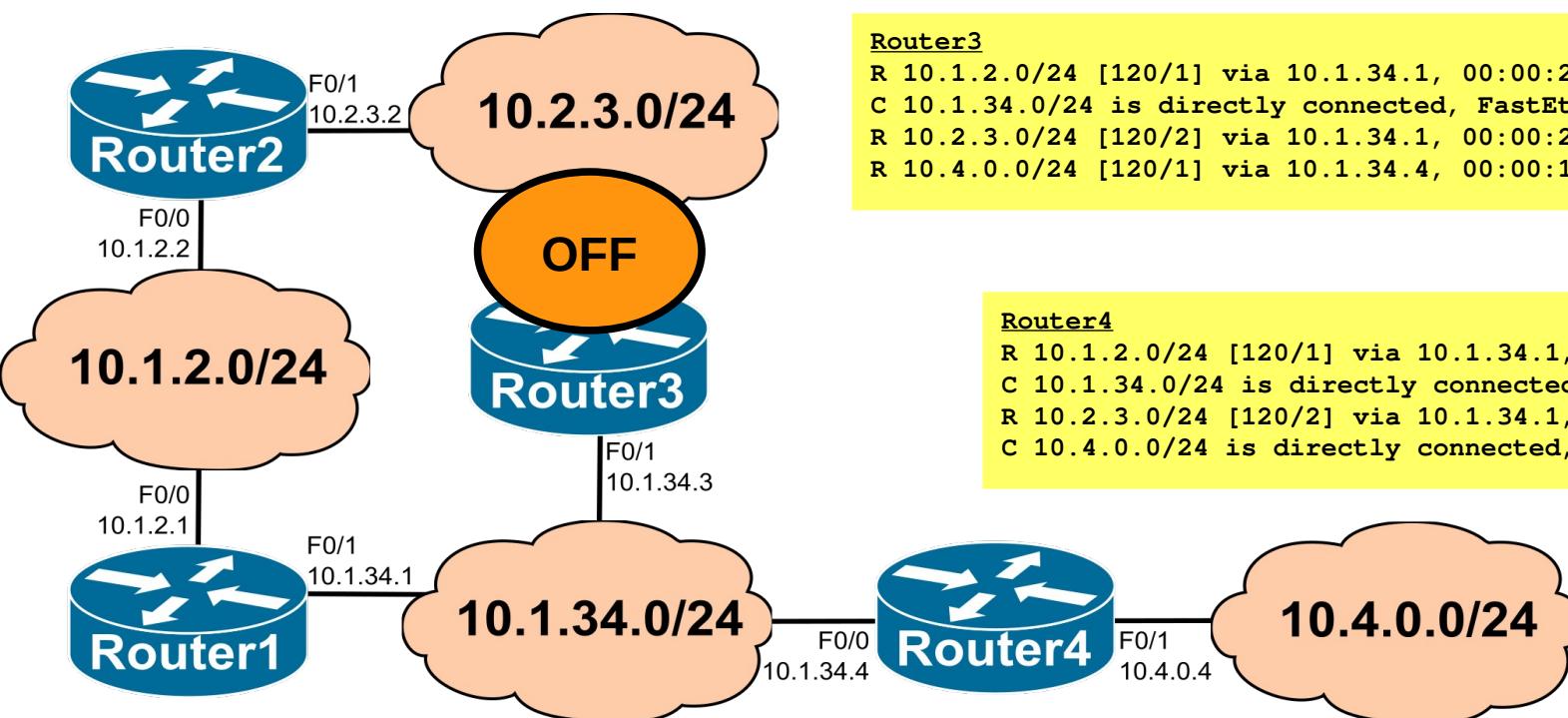
```
C 10.1.2.0/24 is directly connected, FastEthernet0/0
C 10.1.34.0/24 is directly connected, FastEthernet0/1
R 10.2.3.0/24 [120/1] via 10.1.34.3, 00:00:11, FastEthernet0/1
          [120/1] via 10.1.2.2, 00:00:01, FastEthernet0/0
R 10.4.0.0/24 [120/1] via 10.1.34.4, 00:00:24, FastEthernet0/1
```



Routing Tables with RIP

Router2 (AFTER 3 minutes TIMEOUT)

```
C 10.1.2.0/24 is directly connected, FastEthernet0/0
R 10.1.34.0/24 [120/1] via 10.1.2.1, 00:00:25, FastEthernet0/0
C 10.2.3.0/24 is directly connected, FastEthernet0/1
R 10.4.0.0/24 [120/2] via 10.1.2.1, 00:00:25, FastEthernet0/0
```



Router3

```
R 10.1.2.0/24 [120/1] via 10.1.34.1, 00:00:22, FastEthernet0/1
C 10.1.34.0/24 is directly connected, FastEthernet0/1
R 10.2.3.0/24 [120/2] via 10.1.34.1, 00:00:22, FastEthernet0/1
R 10.4.0.0/24 [120/1] via 10.1.34.4, 00:00:19, FastEthernet0/11
```

Router4

```
R 10.1.2.0/24 [120/1] via 10.1.34.1, 00:00:18, FastEthernet0/0
C 10.1.34.0/24 is directly connected, FastEthernet0/0
R 10.2.3.0/24 [120/2] via 10.1.34.1, 00:00:29, FastEthernet0/0
C 10.4.0.0/24 is directly connected, FastEthernet0/1
```

Router1

```
C 10.1.2.0/24 is directly connected, FastEthernet0/0
C 10.1.34.0/24 is directly connected, FastEthernet0/1
R 10.2.3.0/24 [120/1] via 10.1.2.2, 00:00:01, FastEthernet0/0
R 10.4.0.0/24 [120/1] via 10.1.34.4, 00:00:24, FastEthernet0/1
```



RIP Message Types

- RIP Response
 - ◆ Distance vector announcement/update message.
 - ◆ Contains the distance vector.
 - ◆ It is sent:
 - ◆ 1 – Periodically (~30 seconds by default, there is a random component).
 - ◆ 2 – Optionally, when some information changes (triggered updates).
 - ◆ 3 – In response to a RIP Request.
 - ◆ In cases 1 and 2:
 - In RIPv1, is sent to the broadcast address.
 - In RIPv2, is sent to the multicast address 224.0.0.9 (Routers com RIP).
 - ◆ In case 3, it is sent only (unicast) to the router that sent the RIP Request.

- RIP Request (Optional)
 - ◆ Sent by a router that was recently started (bootstrap) or, when the validity of some of the distance vector information has expired (default timeout = 180 seconds)
 - ◆ It may request specific information (a specific network) or, the complete neighbor distance vector.



RIPv1 vs. RIPv2 Responses (1)

- New RIPv2 message fields in Response packets:
 - ◆ Subnet mask
 - Supports variable length masks.
 - Makes RIPv2 *classless* protocol.
 - ◆ Route tag
 - Attribute assigned to a specific network that must be reserved a re-announced.
 - Provides a method to separate internal (to the RIP domain) and external networks.
 - ◆ Next hop
 - Address to which the packets must be routed.
 - 0.0.0.0 indicates that the packets must be routed to the router that sent the RIP message.

RIPv1

Bit	0	7	8	15	16	23	24	31
	Command = 1 or 2	Version = 1			Must be zero			
Route Entry			Address family identifier (2 = IP)			Must be zero		
						IP Address (Network Address)		
						Must be zero		
						Must be zero		
						Metric (Hops)		
	Multiple Route Entries, up to a maximum of 25							

RIPv2

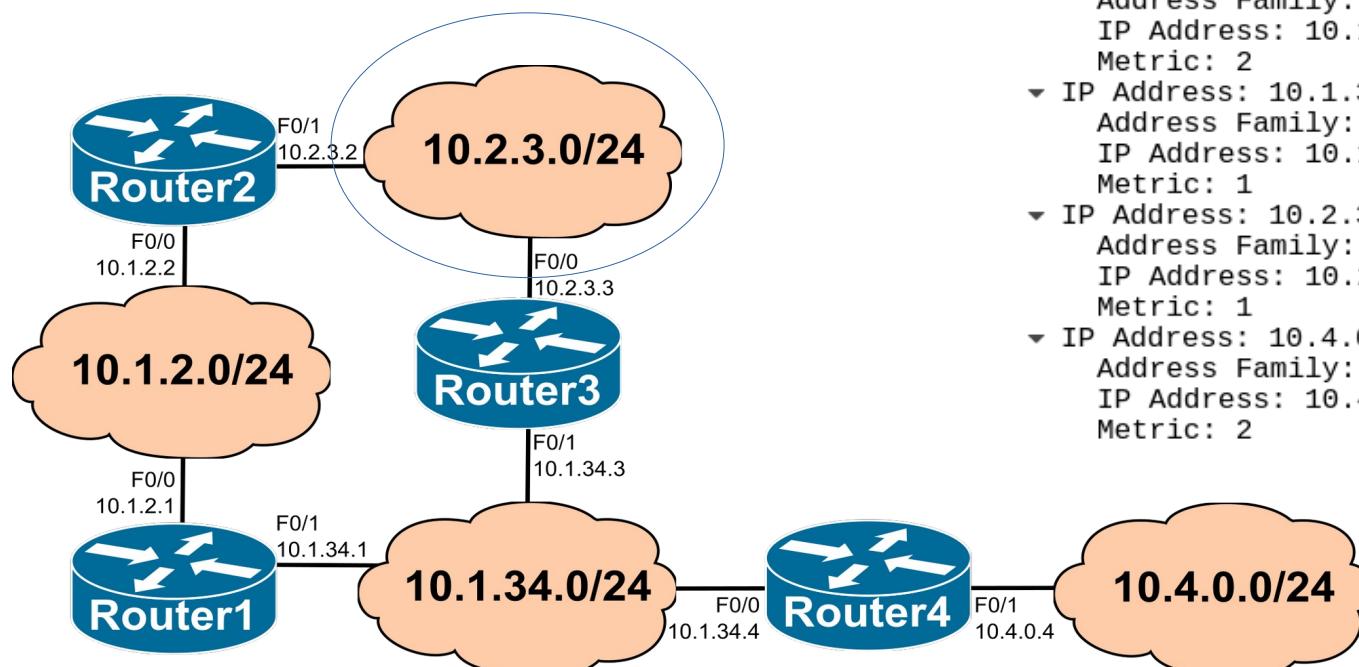
Bit	0	7	8	15	16	23	24	31
	Command = 1 or 2	Version = 2			Must be zero			
Route Entry			Address family identifier (2 = IP)			Route Tag		
						IP Address (Network Address)		
						Subnet Mask		
						Next Hop		
						Metric (Hops)		
	Multiple Route Entries, up to a maximum of 25							



RIPv1 Messages (Example)

Sent by Router3 with Split-Horizon

- ▶ Internet Protocol Version 4, Src: 10.2.3.3, Dst: 255.255.255.255
- ▶ User Datagram Protocol, Src Port: 520, Dst Port: 520
- ▶ Routing Information Protocol
 - Command: Response (2)
 - Version: RIPv1 (1)
 - IP Address: 10.1.34.0, Metric: 1
 - Address Family: IP (2)
 - IP Address: 10.1.34.0
 - Metric: 1
 - IP Address: 10.4.0.0, Metric: 2
 - Address Family: IP (2)
 - IP Address: 10.4.0.0
 - Metric: 2



Sent by Router3 without Split-Horizon

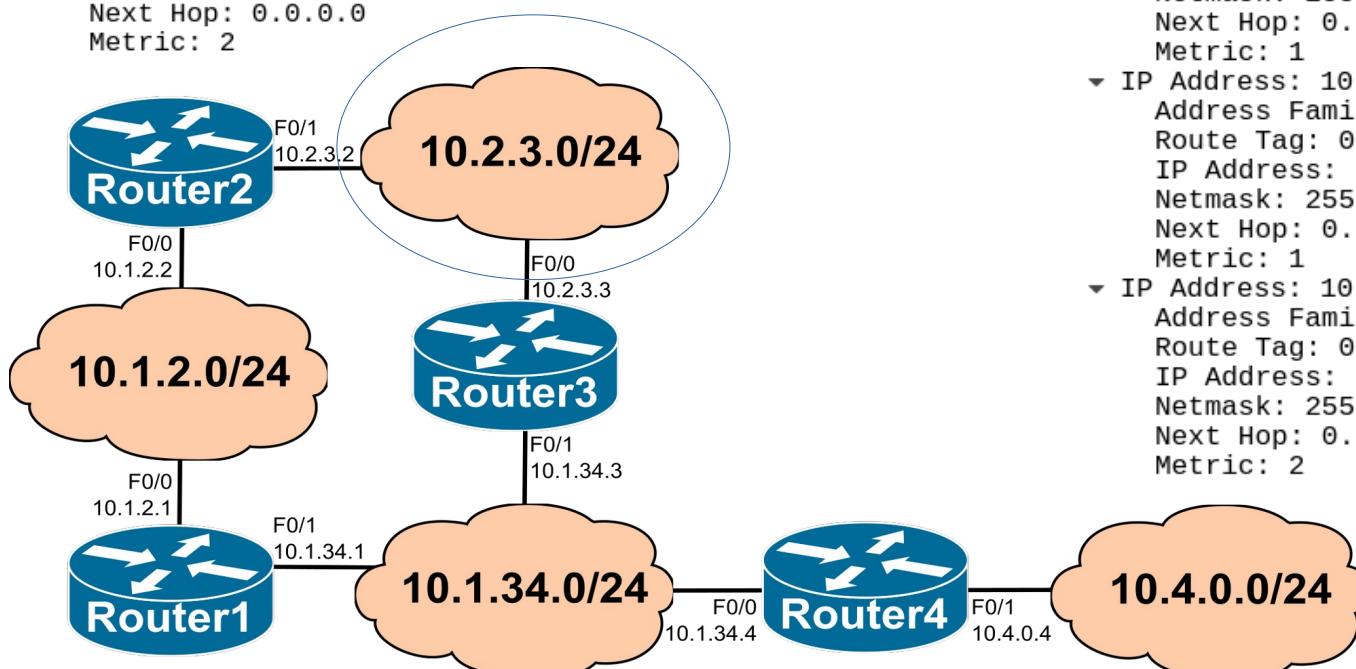
- ▶ Internet Protocol Version 4, Src: 10.2.3.3, Dst: 255.255.255.255
- ▶ User Datagram Protocol, Src Port: 520, Dst Port: 520
- ▶ Routing Information Protocol
 - Command: Response (2)
 - Version: RIPv1 (1)
 - IP Address: 10.1.2.0, Metric: 2
 - Address Family: IP (2)
 - IP Address: 10.1.2.0
 - Metric: 2
 - IP Address: 10.1.34.0, Metric: 1
 - Address Family: IP (2)
 - IP Address: 10.1.34.0
 - Metric: 1
 - IP Address: 10.2.3.0, Metric: 1
 - Address Family: IP (2)
 - IP Address: 10.2.3.0
 - Metric: 1
 - IP Address: 10.4.0.0, Metric: 2
 - Address Family: IP (2)
 - IP Address: 10.4.0.0
 - Metric: 2



RIPv2 Messages (Example)

Sent by Router3 with Split-Horizon

```
► Internet Protocol Version 4, Src: 10.2.3.3, Dst: 224.0.0.9
► User Datagram Protocol, Src Port: 520, Dst Port: 520
▼ Routing Information Protocol
    Command: Response (2)
    Version: RIPv2 (2)
    ▼ IP Address: 10.1.34.0, Metric: 1
        Address Family: IP (2)
        Route Tag: 0
        IP Address: 10.1.34.0
        Netmask: 255.255.255.0
        Next Hop: 0.0.0.0
        Metric: 1
    ▼ IP Address: 10.4.0.0, Metric: 2
        Address Family: IP (2)
        Route Tag: 0
        IP Address: 10.4.0.0
        Netmask: 255.255.255.0
        Next Hop: 0.0.0.0
        Metric: 2
```



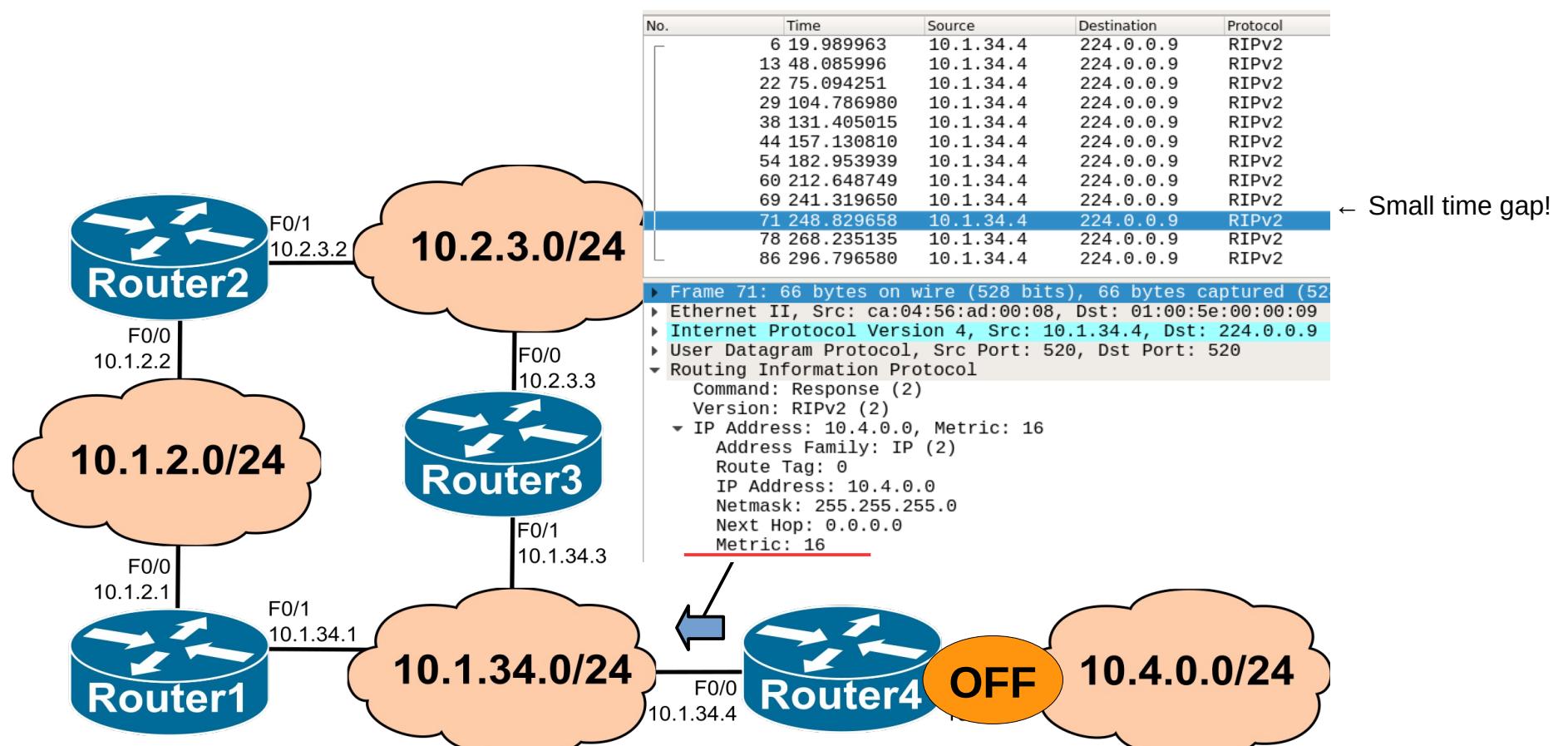
Sent by Router3 without Split-Horizon

```
► Internet Protocol Version 4, Src: 10.2.3.3, Dst: 224.0.0.9
► User Datagram Protocol, Src Port: 520, Dst Port: 520
▼ Routing Information Protocol
    Command: Response (2)
    Version: RIPv2 (2)
    ▼ IP Address: 10.1.2.0, Metric: 2
        Address Family: IP (2)
        Route Tag: 0
        IP Address: 10.1.2.0
        Netmask: 255.255.255.0
        Next Hop: 10.2.3.2
        Metric: 2
    ▼ IP Address: 10.1.34.0, Metric: 1
        Address Family: IP (2)
        Route Tag: 0
        IP Address: 10.1.34.0
        Netmask: 255.255.255.0
        Next Hop: 0.0.0.0
        Metric: 1
    ▼ IP Address: 10.2.3.0, Metric: 1
        Address Family: IP (2)
        Route Tag: 0
        IP Address: 10.2.3.0
        Netmask: 255.255.255.0
        Next Hop: 0.0.0.0
        Metric: 1
    ▼ IP Address: 10.4.0.0, Metric: 2
        Address Family: IP (2)
        Route Tag: 0
        IP Address: 10.4.0.0
        Netmask: 255.255.255.0
        Next Hop: 0.0.0.0
        Metric: 2
```



Triggered Updates

- Prevents any routing loops that involve more than two routers.
- Whenever a router changes the metric for a route, it is required to send update messages almost immediately, even if it is not yet time for one of the regular update message.
- Neighboring routers update routing tables faster and overall convergence is faster.
 - Including entries that were removed by timeout!



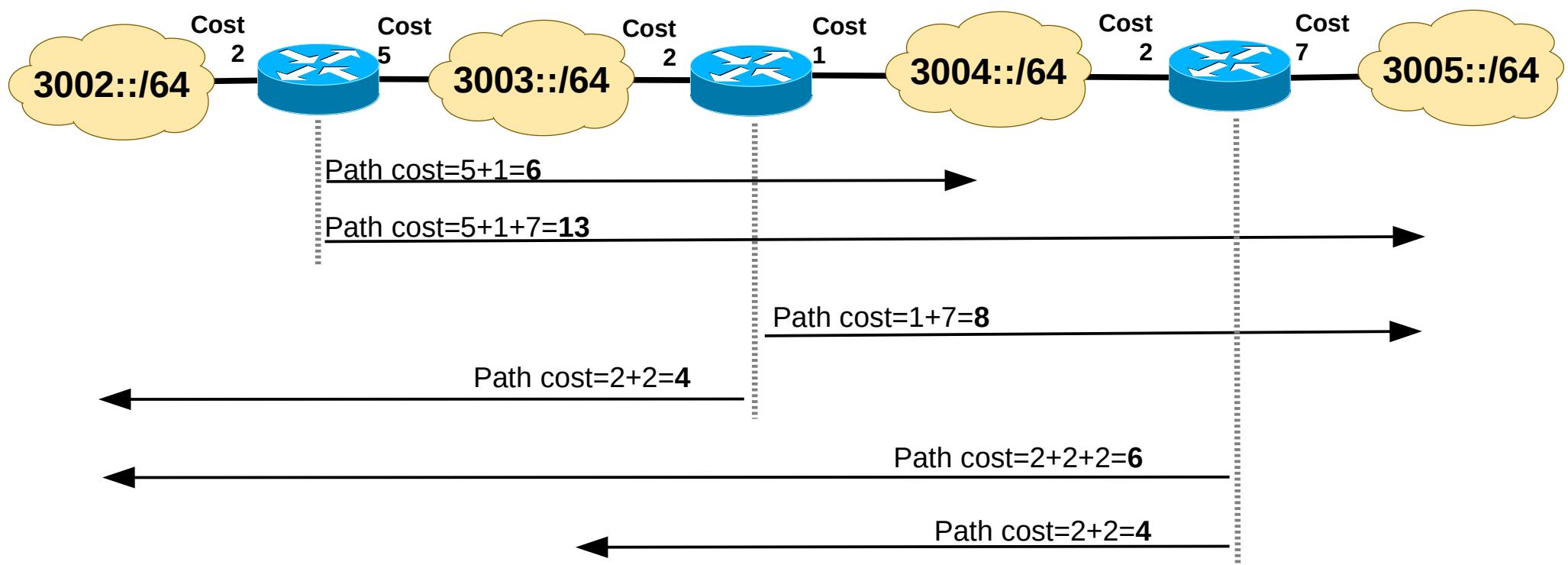
RIPng for IPv6 Routing

- Similar to IPv4 RIPv2:
 - ◆ Distance-vector concept, radius of 15 hops, infinity metric is 16, split-horizon, triggered update.
- Differences between RIPv2 and RIPng
 - ◆ Uses IPv6 for transport.
 - ◆ Uses link-local addresses (not the global ones).
 - ◆ IPv6 prefix, next-hop IPv6 link-local address.
 - ◆ Uses multicast group address FF02::9 (all-RIP-routers) as the destination address for RIP updates.
 - ◆ Routers always add the cost of the interface to the metric received.
 - ◆ Metric is sum of “output interfaces” costs to destination and not number of hops.
 - ◆ If all costs are 1, metric is number of “output interfaces” to destination.
 - ◆ Allows for node/interface costs other than 1.
 - ◆ Cisco calls it “cost offset” per interface (out or in direction).
 - ◆ Cost to network is given by the sum of all output interfaces costs along the path.
 - ◆ With the infinity metric value at 16, this require careful configurations.
 - ◆ Routers always announce directed connected networks.
 - ◆ in IOS Cisco
 - ◆ Activation per interface, named process, more than one active process.



RIPng Path Costs

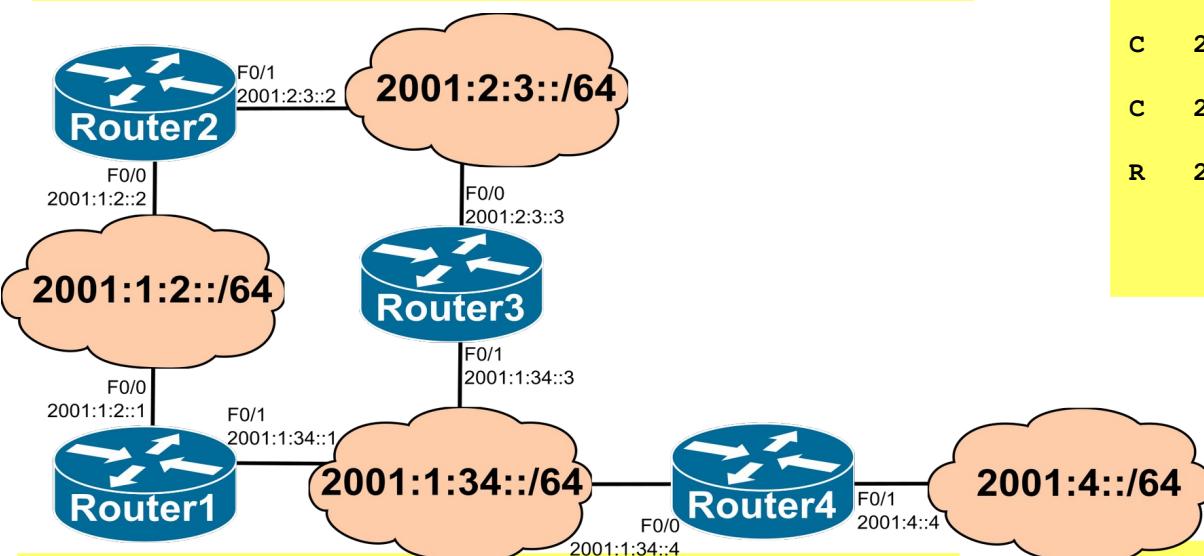
- Each router link/interface has an associated RIPng cost.
- The total cost between a router and a network is given by the sum of all RIPng costs of the (routers) output interfaces along the path.
 - ◆ Routers to access directly connect networks never use RIPng paths.



IPv6 Routing Tables with RIPng

Router2

```
C  2001:1:2::/64 [0/0]
    via FastEthernet0/0, directly connected
R  2001:1:34::/64 [120/2]
    via FE80::C801:54FF:FE41:8, FastEthernet0/0
    via FE80::C803:56FF:FE0A:8, FastEthernet0/1
C  2001:2:3::/64 [0/0]
    via FastEthernet0/1, directly connected
R  2001:4::/64 [120/3]
    via FE80::C801:54FF:FE41:8, FastEthernet0/0
    via FE80::C803:56FF:FE0A:8, FastEthernet0/1
```



Router1

```
C  2001:1:2::/64 [0/0]
    via FastEthernet0/0, directly connected
C  2001:1:34::/64 [0/0]
    via FastEthernet0/1, directly connected
R  2001:2:3::/64 [120/2]
    via FE80::C802:54FF:FEF5:8, FastEthernet0/0
    via FE80::C803:56FF:FE0A:6, FastEthernet0/1
R  2001:4::/64 [120/2]
    via FE80::C804:56FF:FEAD:8, FastEthernet0/1
```

Assuming all interfaces with cost 1.

Router3

```
R  2001:1:2::/64 [120/2]
    via FE80::C802:54FF:FEF5:6, FastEthernet0/0
    via FE80::C801:54FF:FE41:6, FastEthernet0/1
C  2001:1:34::/64 [0/0]
    via FastEthernet0/1, directly connected
C  2001:2:3::/64 [0/0]
    via FastEthernet0/0, directly connected
R  2001:4::/64 [120/2]
    via FE80::C804:56FF:FEAD:8, FastEthernet0/1
```

Router4

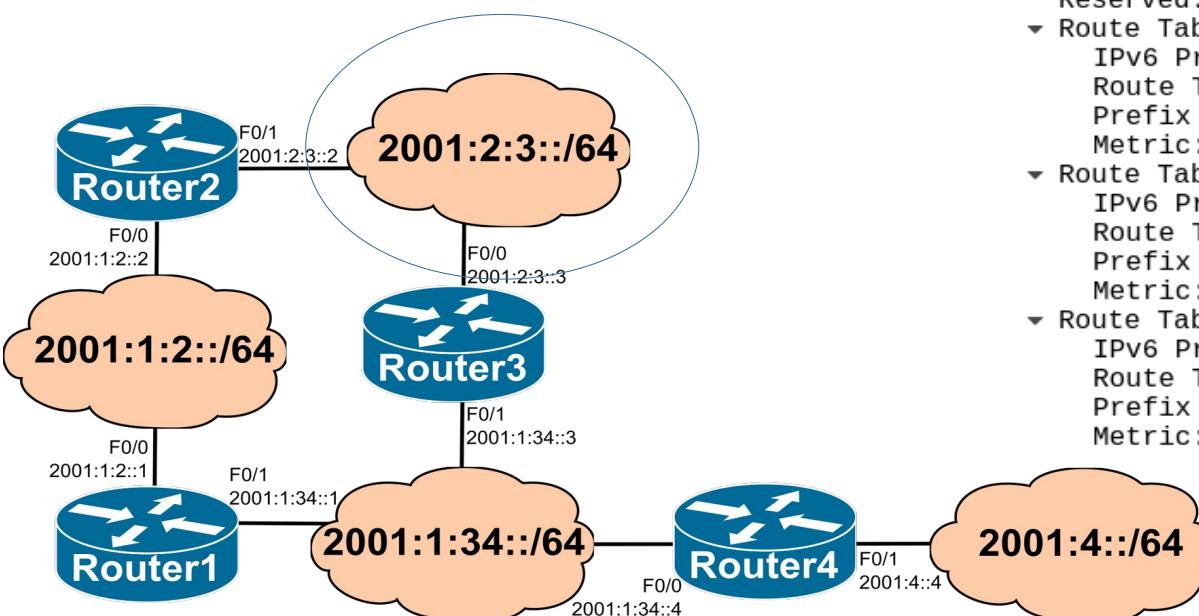
```
R  2001:1:2::/64 [120/2]
    via FE80::C801:54FF:FE41:6, FastEthernet0/0
C  2001:1:34::/64 [0/0]
    via FastEthernet0/0, directly connected
R  2001:2:3::/64 [120/2]
    via FE80::C803:56FF:FE0A:6, FastEthernet0/0
C  2001:4::/64 [0/0]
    via FastEthernet0/1, directly connected
```



RIPng Messages (Example)

Sent by Router2 with Split-Horizon

```
► Internet Protocol Version 6, Src: fe80::c802:54ff:fef5:6, Dst: ff02::9
► User Datagram Protocol, Src Port: 521, Dst Port: 521
▼ RIPng
  Command: Response (2)
  Version: 1
  Reserved: 0000
▼ Route Table Entry: IPv6 Prefix: 2001:1:2::/64 Metric: 1
  IPv6 Prefix: 2001:1:2::
  Route Tag: 0x0000
  Prefix Length: 64
  Metric: 1
▼ Route Table Entry: IPv6 Prefix: 2001:2:3::/64 Metric: 1
  IPv6 Prefix: 2001:2:3::
  Route Tag: 0x0000
  Prefix Length: 64
  Metric: 1
```



Sent by Router3 with Split-Horizon

```
► Internet Protocol Version 6, Src: fe80::c803:56ff:fe0a:8, Dst: ff02::9
► User Datagram Protocol, Src Port: 521, Dst Port: 521
▼ RIPng
  Command: Response (2)
  Version: 1
  Reserved: 0000
▼ Route Table Entry: IPv6 Prefix: 2001:2:3::/64 Metric: 1
  IPv6 Prefix: 2001:2:3::
  Route Tag: 0x0000
  Prefix Length: 64
  Metric: 1
▼ Route Table Entry: IPv6 Prefix: 2001:1:34::/64 Metric: 1
  IPv6 Prefix: 2001:1:34::
  Route Tag: 0x0000
  Prefix Length: 64
  Metric: 1
▼ Route Table Entry: IPv6 Prefix: 2001:4::/64 Metric: 2
  IPv6 Prefix: 2001:4::
  Route Tag: 0x0000
  Prefix Length: 64
  Metric: 2
```



Open Shortest Path First (OSPF) Protocol

- OSPF is an open-standard protocol based primarily on RFC 2328.
- OSPF is a link-state routing protocol
 - ◆ Respond quickly to network changes,
 - ◆ Send triggered updates when a network change occurs,
 - ◆ Send periodic updates, known as link-state refresh, at long time intervals, such as every 30 minutes.
- Routers running OSPF collect routing information from all other routers in the network (or from within a defined area of the network)
- And then each router independently calculates its best paths to all destinations in the network, using Dijkstra's (SPF) algorithm.



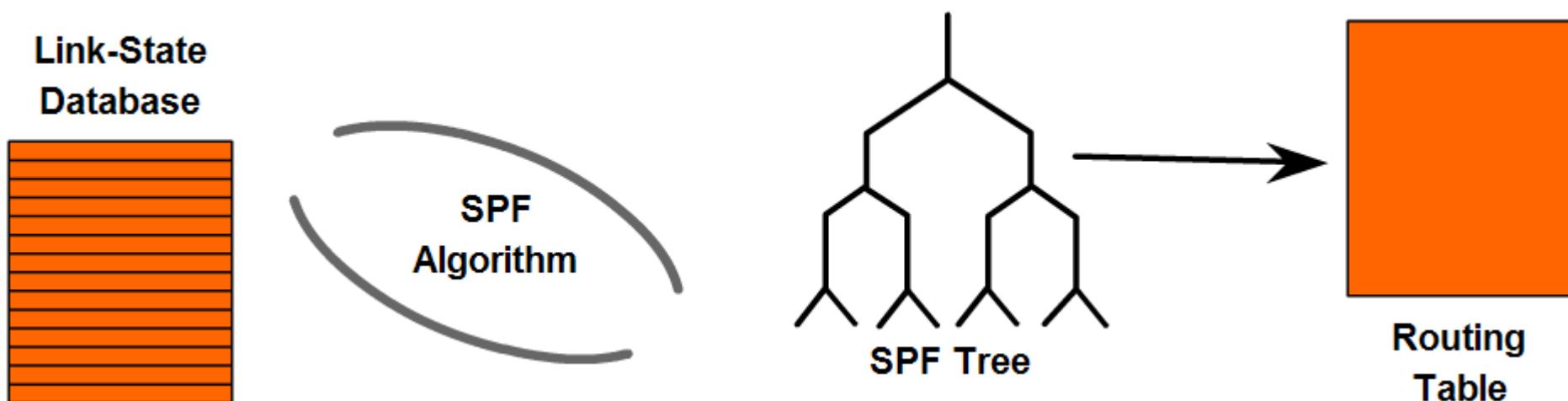
OSPF Necessary Routing Information

- For all the routers in the network to make consistent routing decisions, each link-state router must keep a record of the following information:
 - Its immediate neighbor routers
 - If the router loses contact with a neighbor router, within a few seconds it invalidates all paths through that router and recalculates its paths through the network.
 - For OSPF, adjacency information about neighbors is stored in the OSPF neighbor table, also known as an adjacency database.
 - All the other routers in the network, or in its area of the network, and their attached networks
 - The router recognizes other routers and networks through LSAs, which are flooded through the network.
 - LSAs are stored in a topology table or database (which is also called an LSDB).
 - The best paths to each destination
 - Each router independently calculates the best paths to each destination in the network using Dijkstra's (SPF) algorithm.
 - All paths are kept in the LSDB.
 - The best paths are then offered to the routing table (also called the forwarding database).
 - Packets arriving at the router are forwarded based on the information held in the routing table.



Link-State Protocol Operation

- Link-state routing protocols generate routing updates only when a change occurs in the network topology.
- When a link changes state, the device that detected the change creates a Link-State Advertisement (LSA) concerning that link.
 - ◆ LSA propagates to neighbor devices using a special multicast address.
- Each router stores the LSA, forwards the LSA to neighboring devices and updates its Link-State DataBase (LSDB).
- Link-state routers find the best paths to a destination by applying Dijkstra's algorithm, also known as SPF, against the LSDB to build the SPF tree.
- Each router selects the best paths from their SPF tree and places them in their routing table.



Link-State Advertisement (LSA)

- LSAs report the state of routers and the links between routers.
- Link-state information must be synchronized between routers.
- LSAs have the following characteristics:
 - ◆ LSAs are reliable. There is a method for acknowledging their delivery.
 - ◆ LSAs are flooded throughout the area (or throughout the domain if there is only one area).
 - ◆ LSAs have a sequence number and a set lifetime, so each router recognizes that it has the most current version of the LSA.
 - ◆ LSAs are periodically refreshed to confirm topology information before they age out of the LSDB.



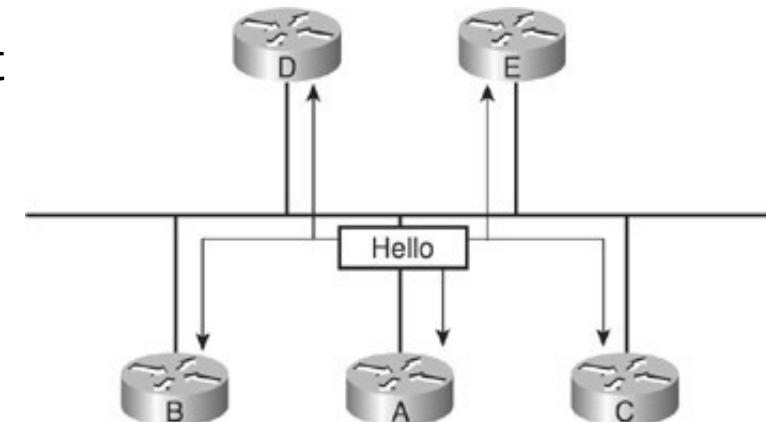
OSPF Router ID (RID)

- The Router ID identifies the router and is:
 - ◆ The highest IPv4 address of all router interfaces at the moment of the OSPF process activation.
 - ◆ A value administratively defined.
- If a physical interface address is being used as the router ID, and that physical interface fails, and the router (or OSPF process) is restarted, the router ID will change.
 - ◆ This change in router ID makes it more difficult for network administrators to troubleshoot and manage OSPF.
- Administratively defining the RID or using loopback interfaces for the router ID forces the router ID to stay the same, regardless of the state of the physical interfaces.



OSPF Adjacencies

- A router running a link-state routing protocol must first establish neighbor adjacencies, by exchanging hello packets with the neighboring routers
- The router sends and receives Hello packets to and from its neighboring routers.
 - The destination address is typically a multicast address.
 - It is possible to define unicast OSPF relations.
- The routers exchange hello packets subject to protocol-specific parameters, such as checking whether the neighbor is in the same area, using the same hello interval, and so on.
 - Routers declare the neighbor up when the exchange is complete.
- Two OSPF routers on a point-to-point serial link, usually encapsulated in High-Level Data Link Control (HDLC) or Point-to-Point Protocol (PPP), form a full adjacency with each other.
- However, OSPF routers on broadcast networks, such as LAN links, elect one router as the designated router (DR) and another as the backup designated router (BDR).
 - All other routers on the LAN form full adjacencies with these two routers and pass LSAs only to them.



DR and BDR Election

- The first OSPF router to boot becomes the Designated Router (DR).
- The second router to boot becomes the Backup Designated Router (BDR).
- If multiple routers boot simultaneously,
 - ◆ The DR it will be the router with the highest priority. The BDR the second.
 - ◆ The OSPF priority is a administratively defined parameter.
 - ◆ In case of tie, it will be chosen the router with the highest Router ID (RID).
- When the DR fails, the BDR assumes the role of DR.
 - ◆ The BDR does not perform any DR functions when the DR is operating.
 - ◆ The choice of the new BDR is done according to some criteria of the initial election.
- After the election, the DR and BDR maintain that role, independently of which routers join the OSPF process.
- The ID of an OSPF Network is the IP address of the network's Designated Router (DR) interface.



OSPF LS Database

- The OSPF database (LSDB) is organized in two tables.
 - Router Link States – Routers related information table.
 - The routers are identified by theirs RID.
 - Net Link States – Networks/Links related information table.
 - Networks are identified by their ID.

OSPF Router with ID (20.20.20.1) (Process ID 1)					
Router Link States (Area 0)					
Link ID	ADV Router	Age	Seq#	Checksum	Link count
20.20.20.1	20.20.20.1	40	0x8000000A	0x00E7FB	2
30.30.30.2	30.30.30.2	69	0x80000006	0x002906	2
30.30.30.3	30.30.30.3	41	0x80000007	0x00283D	2
Net Link States (Area 0)					
Link ID	ADV Router	Age	Seq#	Checksum	
10.10.10.3	30.30.30.3	41	0x80000001	0x00051C	
20.20.20.2	30.30.30.2	70	0x80000001	0x00A164	
30.30.30.3	30.30.30.3	154	0x80000001	0x00A91C	



OSPF LS Database Tables (1)

- Router Link States

- For each router, it contains the information about the networks directly connected to that router.

```
LS age: 321
Options: (No TOS-capability, DC)
LS Type: Router Links
Link State ID: 20.20.20.1 ← Router ID
Advertising Router: 20.20.20.1
LS Seq Number: 8000000A
Checksum: 0xE7FB
Length: 48
Number of Links: 2 ← Number of Links

Link connected to: a Transit Network ← Network Type
(Link ID) Designated Router address: 20.20.20.2 ← Network ID
(Link Data) Router Interface address: 20.20.20.1 ← Interface IP Address
Number of TOS metrics: 0
TOS 0 Metrics: 1 ← Interface Cost

Link connected to: a Transit Network
(Link ID) Designated Router address: 10.10.10.3
(Link Data) Router Interface address: 10.10.10.1
Number of TOS metrics: 0
TOS 0 Metrics: 1
```



OSPF LS Database Tables (2)

- Network Link States

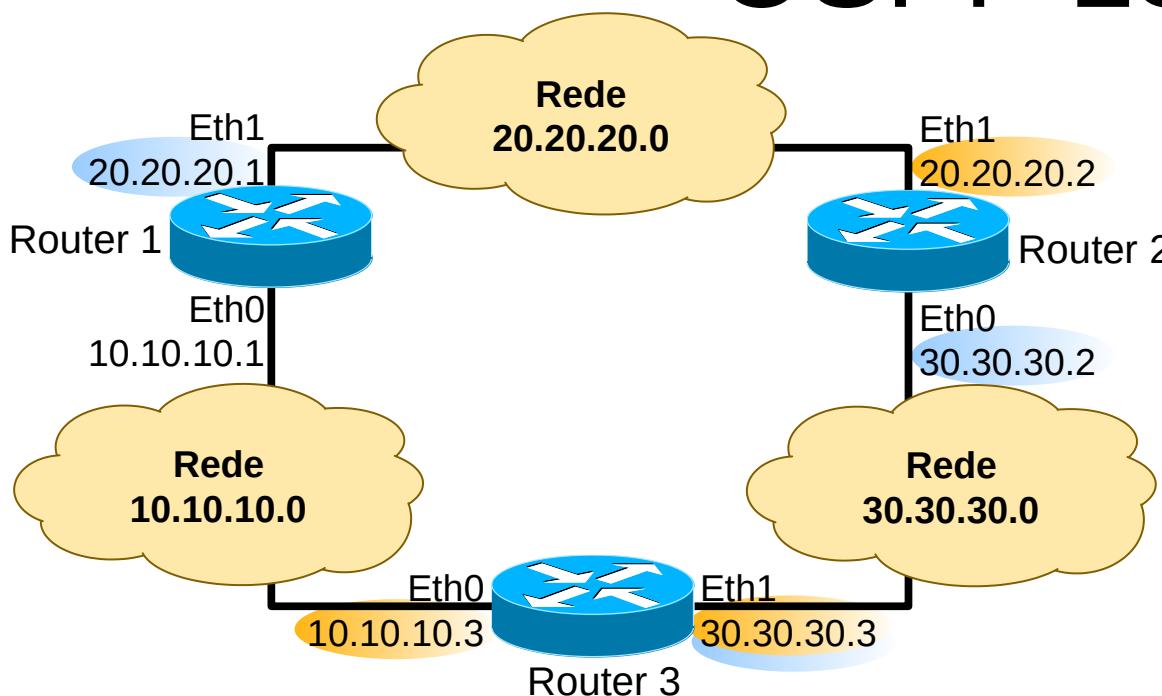
- For each network, it contains the information about the routers directly attached to that network.

```
Routing Bit Set on this LSA
LS age: 483
Options: (No TOS-capability, DC)
LS Type: Network Links
Link State ID: 10.10.10.3 (address of Designated Router)
Advertising Router: 30.30.30.3
LS Seq Number: 80000001
Checksum: 0x51C
Length: 32
Network Mask: /24
Attached Router: 30.30.30.3
Attached Router: 20.20.20.1 }
```

Network ID
Attached routers (RID)



OSPF LSDatabase Example



Routing Bit Set on this LSA

LS age: 208

Options: (No TOS-capability, DC)

LS Type: Network Links

Link State ID: 20.20.20.2 (address of Designated Router)
Advertising Router: 30.30.30.2

LS Seq Number: 80000001

Checksum: 0xA164

Length: 32

Network Mask: /24

Attached Router: 30.30.30.2

Attached Router: 20.20.20.1

Network 20.20.20.0's Network Link State

LS age: 321

Options: (No TOS-capability, DC)

LS Type: Router Links

Link State ID: 20.20.20.1

Advertising Router: 20.20.20.1

LS Seq Number: 8000000A

Checksum: 0xE7FB

Length: 48

Number of Links: 2

Link connected to: a Transit Network

(Link ID) Designated Router address: 20.20.20.2

(Link Data) Router Interface address: 20.20.20.1

Number of TOS metrics: 0

TOS 0 Metrics: 1

Link connected to: a Transit Network

(Link ID) Designated Router address: 10.10.10.3

(Link Data) Router Interface address: 10.10.10.1

Number of TOS metrics: 0

TOS 0 Metrics: 1

Router 1's Router Link State



OSPF Packets

- Hello - Discovers neighbors and builds adjacencies between them.
- Database Description (DBD) - Checks for database synchronization between routers.
- Link-State Request (LSR) - Requests specific link-state records from another router.
- Link-State Update (LSU) - Sends specifically requested link-state records.
- LSAck - Acknowledges the other packet types.



OSPF Packet Format

- Version Number

- Set to 2 for OSPF Version 2, the IPv4 version of OSPF.
- Set to 3 for OSPF Version 3, the IPv6 version of OSPF.

- Type

- Differentiates the five OSPF packet types.

- Packet Length

- The length of the OSPF packet in bytes.

- Router ID

- Defines which router is the packet's source.

- Area ID

- Defines the area in which the packet originated.

- Checksum

- Used for packet header error detection to ensure that the OSPF packet was not corrupted during transmission.

- Authentication Type

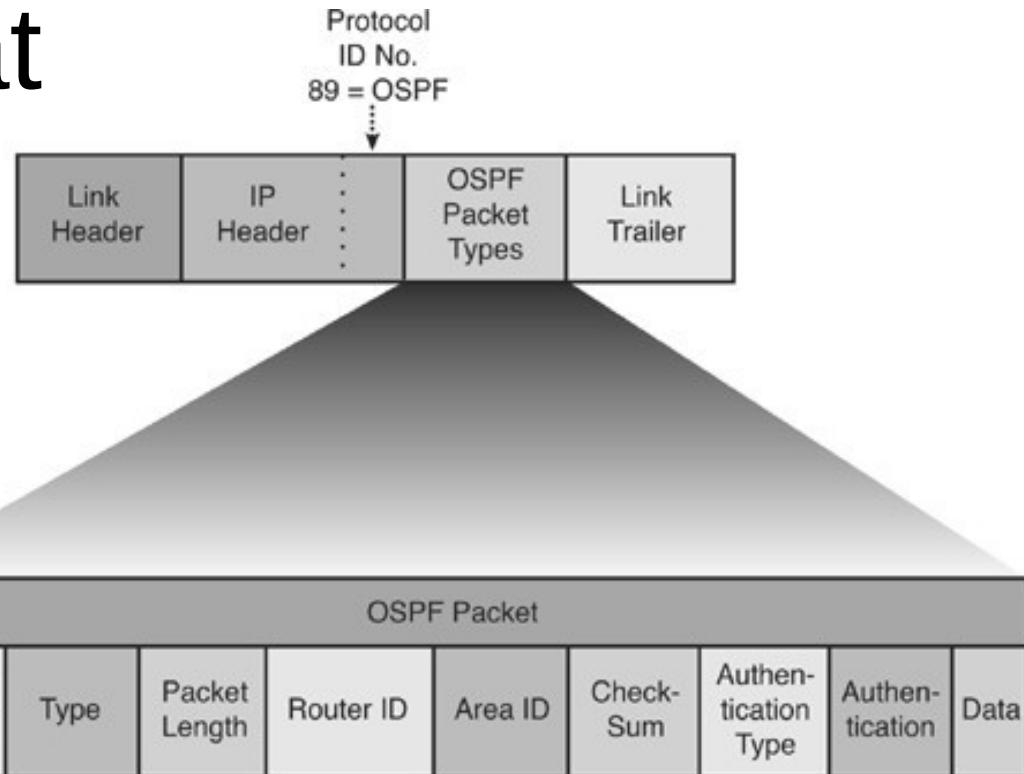
- An option in OSPF that describes either no authentication, clear-text passwords, or encrypted message digest 5 (MD5) for router authentication.

- Authentication

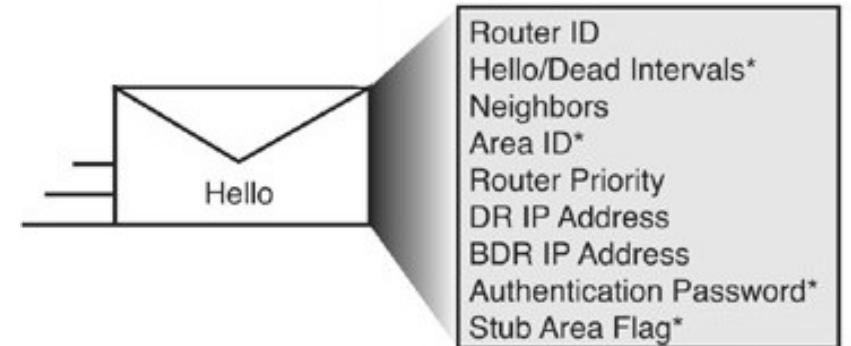
- Used with authentication type.

- Data, contains different information, depending on the OSPF packet type:

- For the Hello packet - Contains a list of known neighbors.
- For the DBD packet - Contains a summary of the LSDB, which includes all known router IDs and their last sequence number, among several other fields.
- For the LSR packet - Contains the type of LSU needed and the router ID of the router that has the needed LSU.
- For the LSU packet - Contains the full LSA entries. Multiple LSA entries can fit in one OSPF update packet.
- For the LSAck packet - This data field is empty.



OSPF Hello Packets



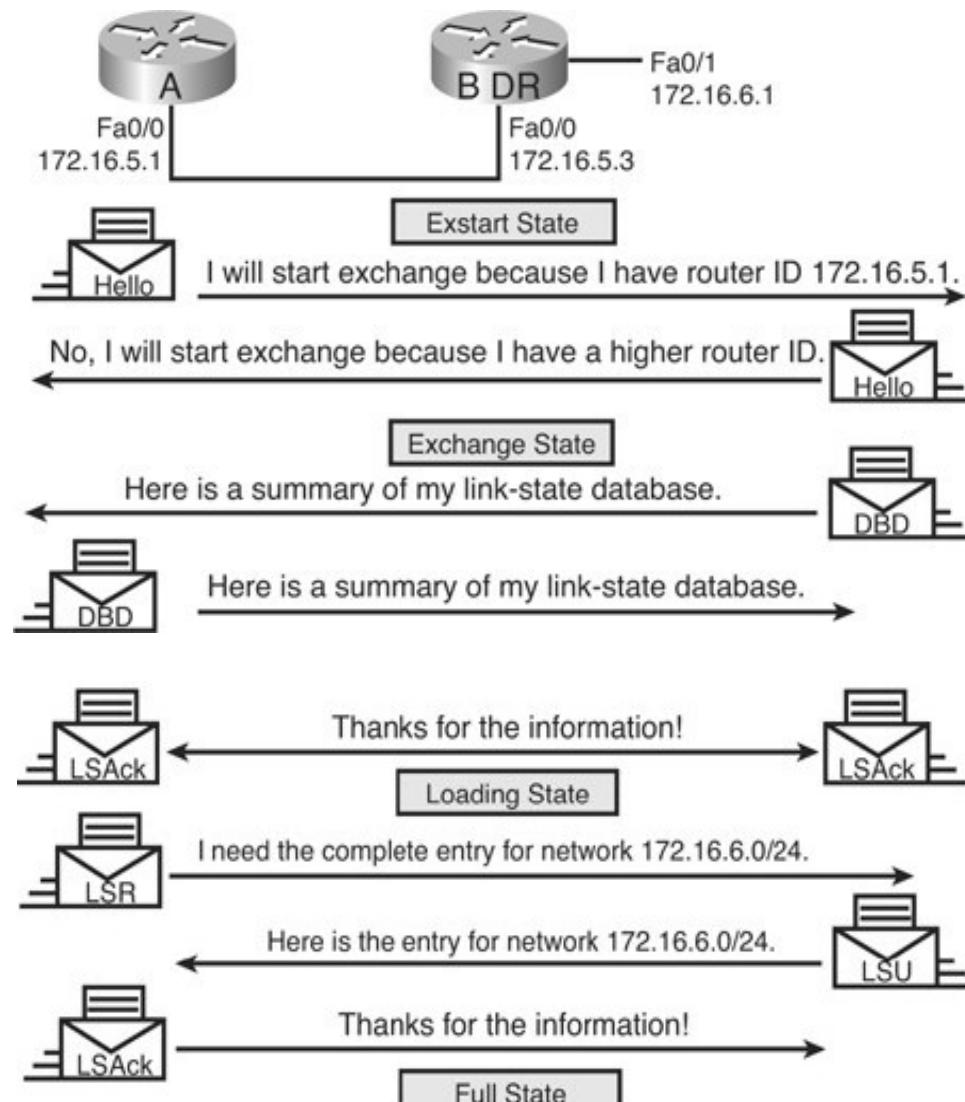
- An hello packet contains the following information:

- Router ID
 - A 32-bit number that uniquely identifies the router.
- Hello and dead intervals
 - The hello interval specifies how often, in seconds, a router sends hello packets (10 seconds is the default on multiaccess networks).
 - The dead interval is the amount of time in seconds that a router waits to hear from a neighbor before declaring the neighbor router out of service (the dead interval is four times the hello interval by default).
 - These timers must be the same on neighboring routers; otherwise an adjacency will not be established.
- Neighbors
 - The Neighbors field lists the adjacent routers with which this router has established bidirectional communication.
 - Bidirectional communication is indicated when the router sees itself listed in the Neighbors field of the hello packet from the neighbor.
- Area ID
 - To communicate, two routers must share a common segment, and their interfaces must belong to the same OSPF area on that segment.
 - These routers will all have the same link-state information for that area.
- Router priority
 - An 8-bit number that indicates a router's priority. Priority is used when electing a DR and BDR.
- DR and BDR IP addresses
 - If known, the IP addresses of the DR and BDR for the specific multiaccess network.
- Authentication password
 - If router authentication is enabled, two routers must exchange the same password.
 - Authentication is not required, but if it is enabled, all peer routers must have the same password.
- Stub area flag
 - A stub area is a special area.
 - The stub area technique reduces routing updates by replacing them with a default route.
 - Two neighboring routers must agree on the stub area flag in the hello packets.

- Hello Interval, Dead Interval, Area ID, Authentication Password and Stub Area Flag fields must match on neighboring routers for them to establish an adjacency.



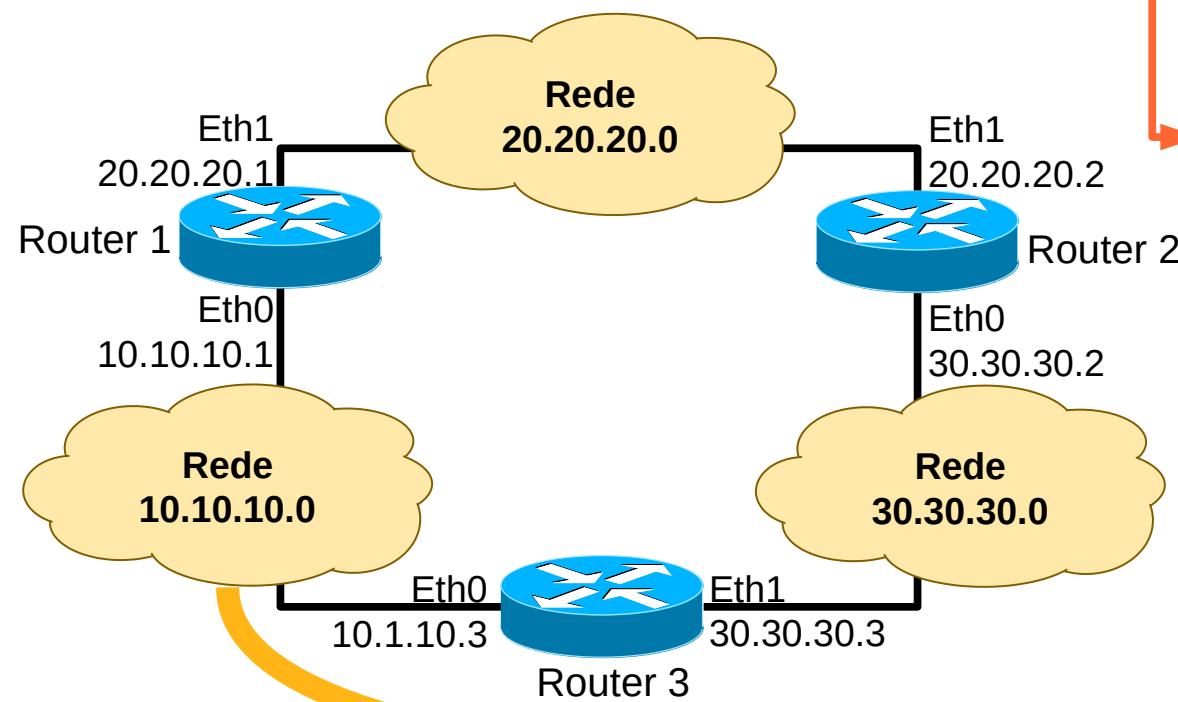
Discovering the Network Routes



- A master and slave relationship is created between each router and its adjacent DR and BDR.
 - ◆ Only the DR exchanges and synchronizes link-state information with the routers to which it has established adjacencies.
- The master and slave routers exchange one or more DBD packets.
 - ◆ A DBD includes information about the LSA entry header that appears in the router's LSDB.
 - ◆ The entries can be about a link or about a network.
 - ◆ Each LSA entry header includes information about the link-state type, the address of the advertising router, the link's cost, and the sequence number.
 - ◆ The router uses the sequence number to determine the "newness" of the received link-state information.
- It acknowledges the receipt of the DBD using the LSAck packet.
 - ◆ It compares the information it received with the information it has in its own LSDB.
- If the DBD has a more current link-state entry, the router sends an LSR to the other router.
- The other router responds with the complete information about the requested entry in an LSU packet.
- Again, when the router receives an LSU, it sends an LSAck.
- The router adds the new link-state entries to its LSDB.



OSPF Example



Time	Source	Destination	Protocol Info
0.000000	10.10.10.1	224.0.0.5	OSPF Hello Packet
10.002318	10.10.10.1	224.0.0.5	OSPF Hello Packet
20.003116	10.10.10.1	224.0.0.5	OSPF Hello Packet

80.000000	10.10.10.3	224.0.0.5	OSPF Hello Packet
83.683033	10.10.10.3	224.0.0.5	OSPF LS Update
83.715683	10.10.10.3	224.0.0.5	OSPF Hello Packet
83.717864	10.10.10.1	10.10.10.3	OSPF Hello Packet
83.726166	10.10.10.3	10.10.10.1	OSPF DB Descr.
83.726258	10.10.10.3	10.10.10.1	OSPF Hello Packet
83.728433	10.10.10.1	10.10.10.3	OSPF DB Descr.
83.732590	10.10.10.3	10.10.10.1	OSPF DB Descr.
83.734733	10.10.10.1	10.10.10.3	OSPF DB Descr.
83.738942	10.10.10.3	10.10.10.1	OSPF LS Request
83.741083	10.10.10.1	10.10.10.3	OSPF LS Update
84.240362	10.10.10.3	224.0.0.5	OSPF LS Update
86.245792	10.10.10.3	224.0.0.5	OSPF LS Acknowledge
86.380876	10.10.10.1	224.0.0.5	OSPF Hello Packet
86.741036	10.10.10.1	224.0.0.5	OSPF LS Acknowledge
93.721376	10.10.10.3	224.0.0.5	OSPF Hello Packet
96.380005	10.10.10.1	224.0.0.5	OSPF Hello Packet

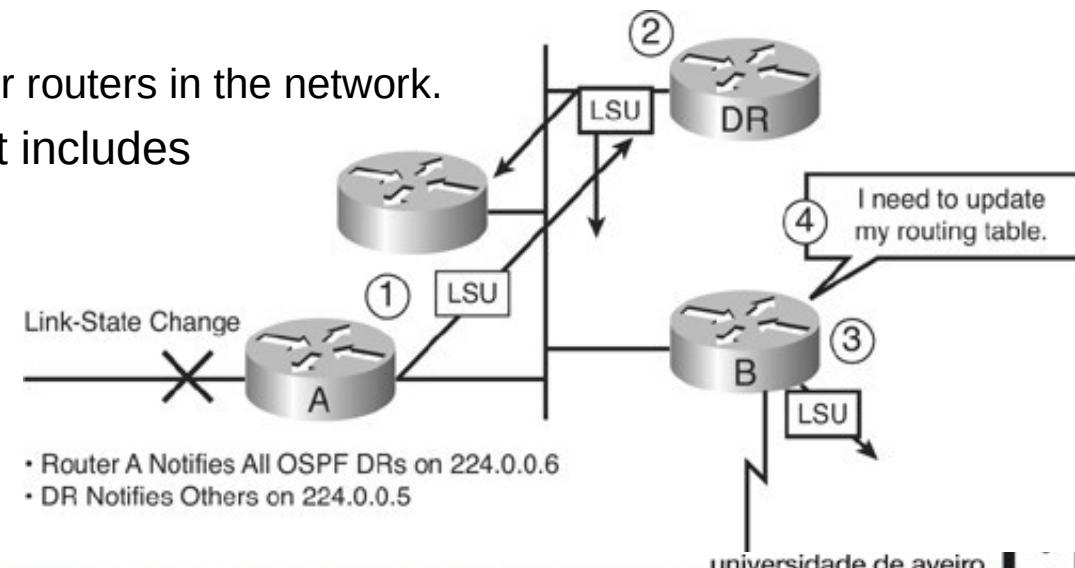
213.780338	10.10.10.3	224.0.0.5	OSPF Hello Packet
216.542473	10.10.10.1	224.0.0.5	OSPF Hello Packet
216.568852	10.10.10.1	224.0.0.5	OSPF LS Update
217.048427	10.10.10.1	224.0.0.5	OSPF LS Update
217.084909	10.10.10.1	224.0.0.5	OSPF LS Update
219.067748	10.10.10.3	224.0.0.5	OSPF LS Acknowledge
219.650308	10.10.10.1	224.0.0.5	OSPF LS Update
222.150349	10.10.10.3	224.0.0.5	OSPF LS Acknowledge
223.779492	10.10.10.3	224.0.0.5	OSPF Hello Packet
224.284149	10.10.10.3	224.0.0.5	OSPF LS Update
224.789598	10.10.10.1	224.0.0.5	OSPF LS Update
224.789775	10.10.10.3	224.0.0.5	OSPF LS Update
226.545718	10.10.10.1	224.0.0.5	OSPF Hello Packet
226.785254	10.10.10.1	224.0.0.5	OSPF LS Acknowledge
227.294756	10.10.10.3	224.0.0.5	OSPF LS Acknowledge
233.779863	10.10.10.3	224.0.0.5	OSPF Hello Packet
236.544658	10.10.10.1	224.0.0.5	OSPF Hello Packet



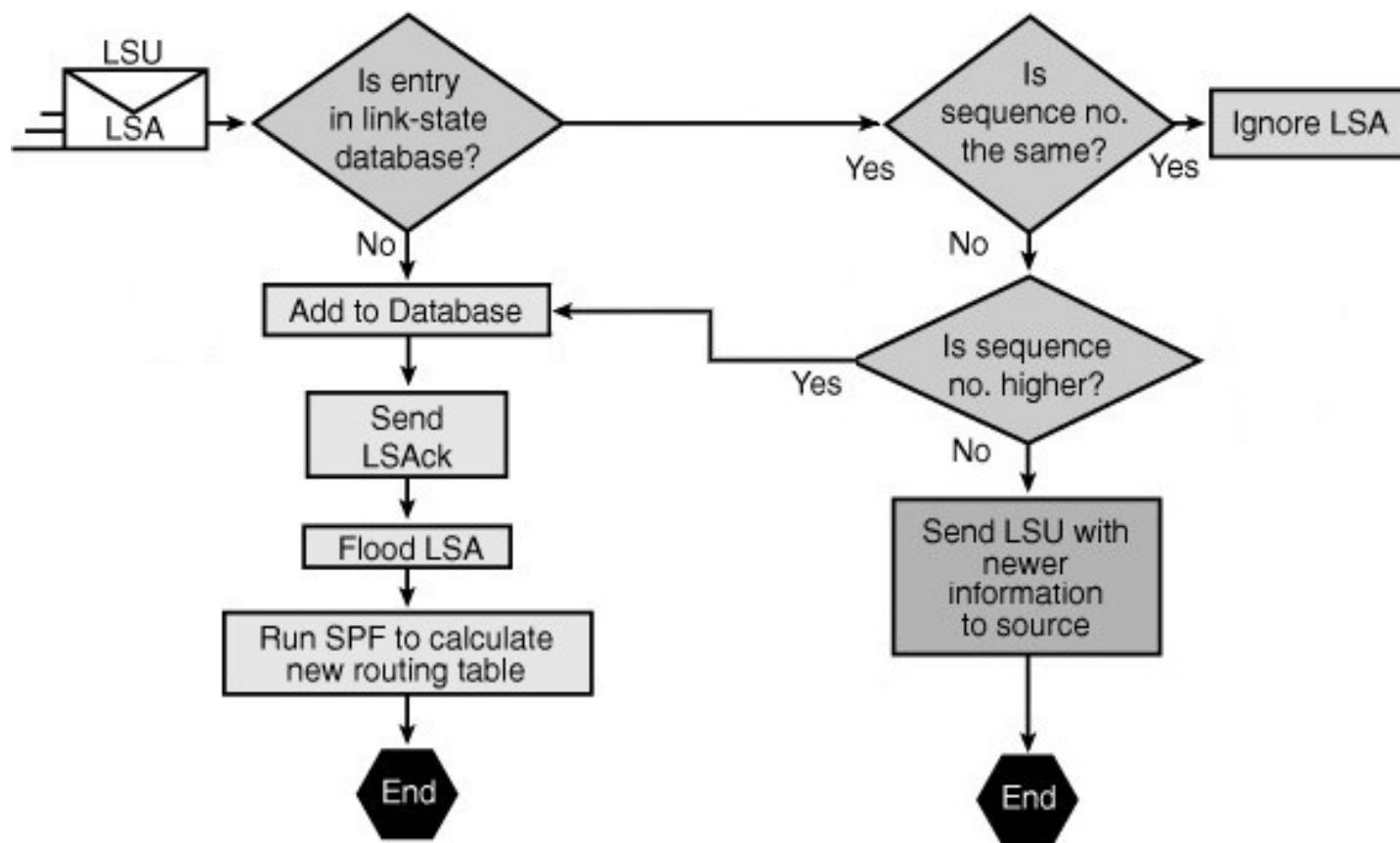
Maintaining Routing Information

- Flooding process:

- ◆ A router notices a change in a link state and multicasts an LSU packet, which includes the updated LSA entry with the sequence number incremented, to 224.0.0.6.
 - This address goes to all OSPF DRs and BDRs.
 - On point-to-point links, the LSU is multicast to 224.0.0.5.)
 - An LSU packet might contain several distinct LSAs.
- ◆ The DR receives the LSU, processes it, acknowledges the receipt of the change and floods the LSU to other routers on the network using the OSPF multicast address 224.0.0.5.
 - After receiving the LSU, each router responds to the DR with an LSAck.
 - To make the flooding procedure reliable, each LSA must be acknowledged separately.
- ◆ If a router is connected to other networks, it floods the LSU to those other networks by forwarding the LSU to the DR of the other network (or to the adjacent router if in a point-to-point network).
 - That DR, in turn, multicasts the LSU to the other routers in the network.
- ◆ The router updates its LSDB using the LSU that includes the changed LSA.
- ◆ It then recomputes the SPF algorithm against the updated database after a short delay and updates the routing table as necessary.



LSA Operation



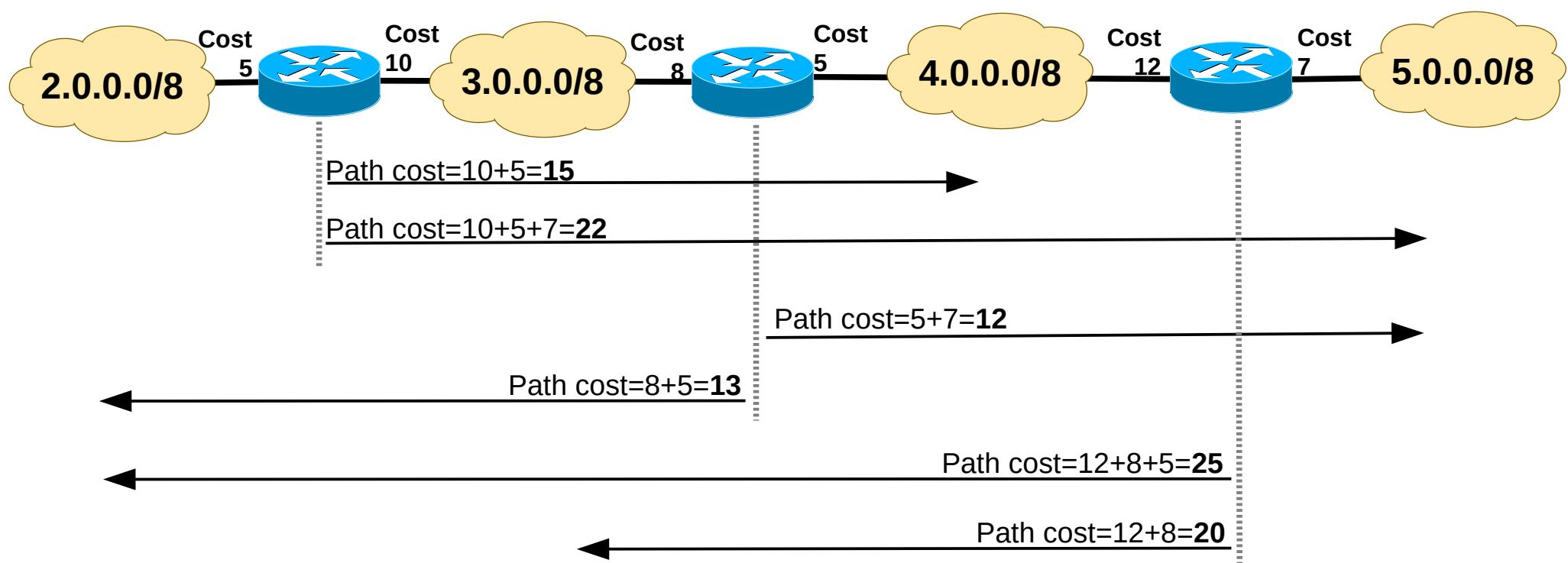
- When each router receives the LSU:

- If the LSA entry does not already exist, the router adds the entry to its LSDB, sends back a link-state acknowledgment (LSAck), floods the information to other routers, runs SPF, and updates its routing table.
- If the entry already exists and the received LSA has the same sequence number, the router ignores the LSA entry.
- If the entry already exists but the LSA includes newer information (it has a higher sequence number), the router adds the entry to its LSDB, sends back an LSAck, floods the information to other routers, runs SPF, and updates its routing table.
- If the entry already exists but the LSA includes older information, it sends an LSU to the sender with its newer information.

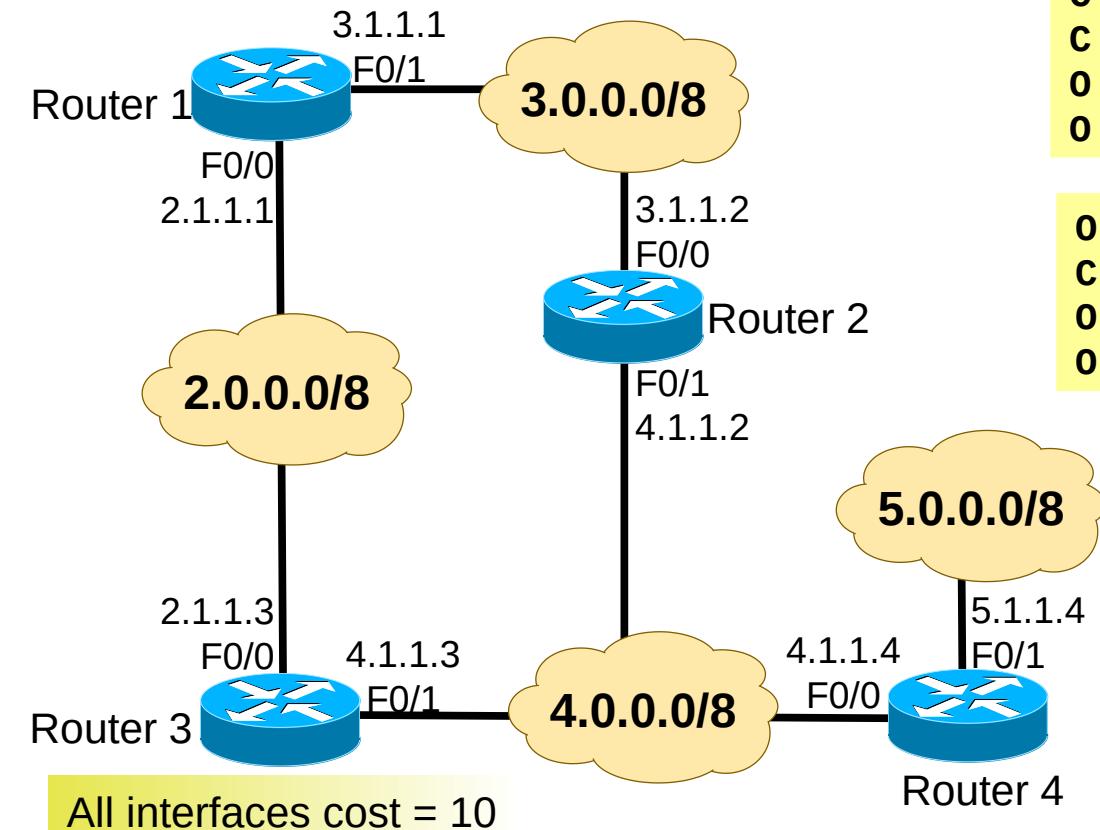


OSPF Path Costs

- Each router link/interface has an associated OSPF cost.
- The total cost between a router and a network is given by the sum of all OSPF costs of the (routers) output interfaces along the path.
 - ◆ Routers to access directly connect networks never use OSPF paths.



OSPF Example



```
C 2.0.0.0/8 is directly connected, F0/0
C 3.0.0.0/8 is directly connected, F0/1
O 4.0.0.0/8 [110/20] via 2.1.1.3, 00:01:18, F0/0
O 5.0.0.0/8 [110/30] via 2.1.1.3, 00:01:00, F0/0
```

```
O 2.0.0.0/8 [110/20] via 3.1.1.1, 00:01:13, F0/0
C 3.0.0.0/8 is directly connected, F0/0
O 4.0.0.0/8 [110/30] via 3.1.1.1, 00:01:13, F0/0
O 5.0.0.0/8 [110/40] via 3.1.1.1, 00:01:10, F0/0
```

Router 1 and Router 2 after disconnecting the F0/1 at Router2

```
C 2.0.0.0/8 is directly connected, F0/0
C 3.0.0.0/8 is directly connected, F0/1
O 4.0.0.0/8 [110/15] via 3.1.1.2, 00:01:13, F0/1
O 5.0.0.0/8 [110/25] via 3.1.1.2, 00:01:10, F0/1
```

Router1, now with the cost of Router2 F0/1 interface equal to 5

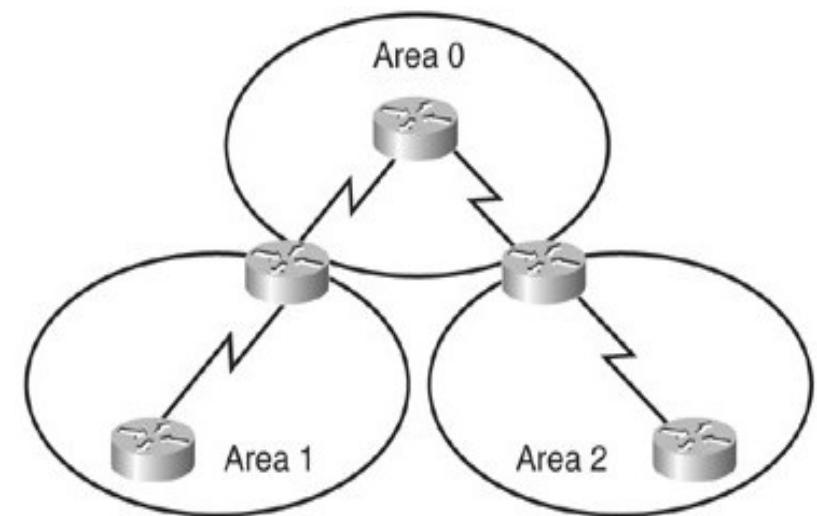
```
C 2.0.0.0/8 is directly connected, F0/0
C 3.0.0.0/8 is directly connected, F0/1
O 4.0.0.0/8 [110/20] via 3.1.1.2, 00:01:13, F0/1
[110/20] via 2.1.1.3, 00:01:31, F0/0
O 5.0.0.0/8 [110/30] via 3.1.1.2, 00:01:10, F0/1
[110/30] via 2.1.1.3, 00:01:10, F0/0
```

Router 1 initial table



OSPF Hierarchical Routing (1)

- In small networks, the web of router links is not complex, and paths to individual destinations are easily deduced.
- In large networks, the resulting web is highly complex, and the number of potential paths to each destination is large.
 - ◆ Dijkstra calculations comparing all of these possible routes can be very complex and can take significant time.
 - Large LSDB. Because the LSDB covers the topology of the entire network, each router must maintain an entry for every network in the area, even if not every route is selected for the routing table.
 - Frequent SPF algorithm calculations. In a large network, changes are inevitable, so the routers spend many CPU cycles recalculating the SPF algorithm and updating the routing table.
 - Large routing table. OSPF does not perform route summarization by default. If the routes are not summarized, the routing tables can become very large, depending on the size of the network.
- Link-state routing protocols usually reduce the size of the Dijkstra calculations by partitioning the network into areas.



OSPF Hierarchical Routing (2)

- Using multiple OSPF areas has several important advantages:
 - ◆ Reduced frequency of SPF calculations.
 - Detailed route information only exists within each area
 - It is not necessary to flood all link-state changes to all other areas.
 - Only routers that are affected by the change need to recalculate the SPF algorithm and the impact of the change is localized within the area.
 - ◆ Reduced updates overhead.
 - Rather than send an update about each network within an area, a router can advertise a single summarized route or a small number of routes between areas, thereby reducing the overhead associated with updates when they cross areas.
 - ◆ Smaller routing tables.
 - Detailed route entries for specific networks within an area can remain in the area.
 - Routers can be configured to summarize the routes into one or more summary addresses.
 - Advertising these summaries reduces the number of messages propagated between areas but keeps all networks reachable.



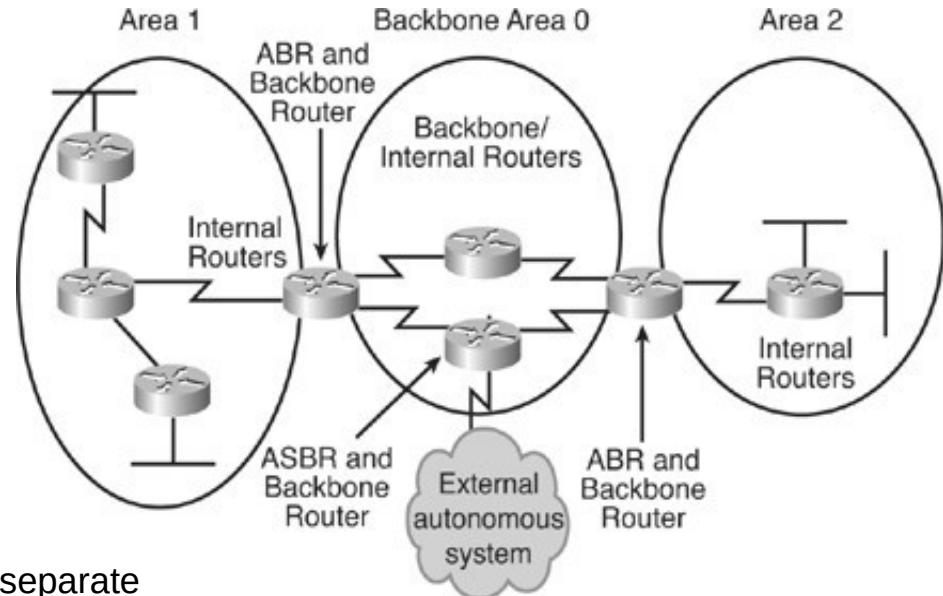
OSPF Two-Layer Area Hierarchy

- OSPF uses a two-layer area hierarchy:
- Backbone area
 - ◆ An OSPF area whose primary function is the fast and efficient movement of IP packets.
 - ◆ The backbone area interconnect with all other OSPF areas.
 - ◆ Generally, end users are not found within a backbone area.
 - ◆ The backbone area is also called OSPF area 0.
 - ◆ Hierarchical networking defines area 0 as the core to which all other areas connect (directly or virtually).
- Regular (non backbone) area
 - ◆ An OSPF area whose primary function is to connect users and resources.
 - ◆ Regular areas (also called normal areas) are usually set up along functional or geographic groupings.
 - By default, a regular area does not allow traffic from another area to use its links to reach other areas.
 - By default, all traffic from other areas must cross backbone area 0.
 - ◆ Regular areas can have several subtypes, including standard area, stub area, totally stubby area, not-so-stubby area (NSSA), and totally stubby NSSA.

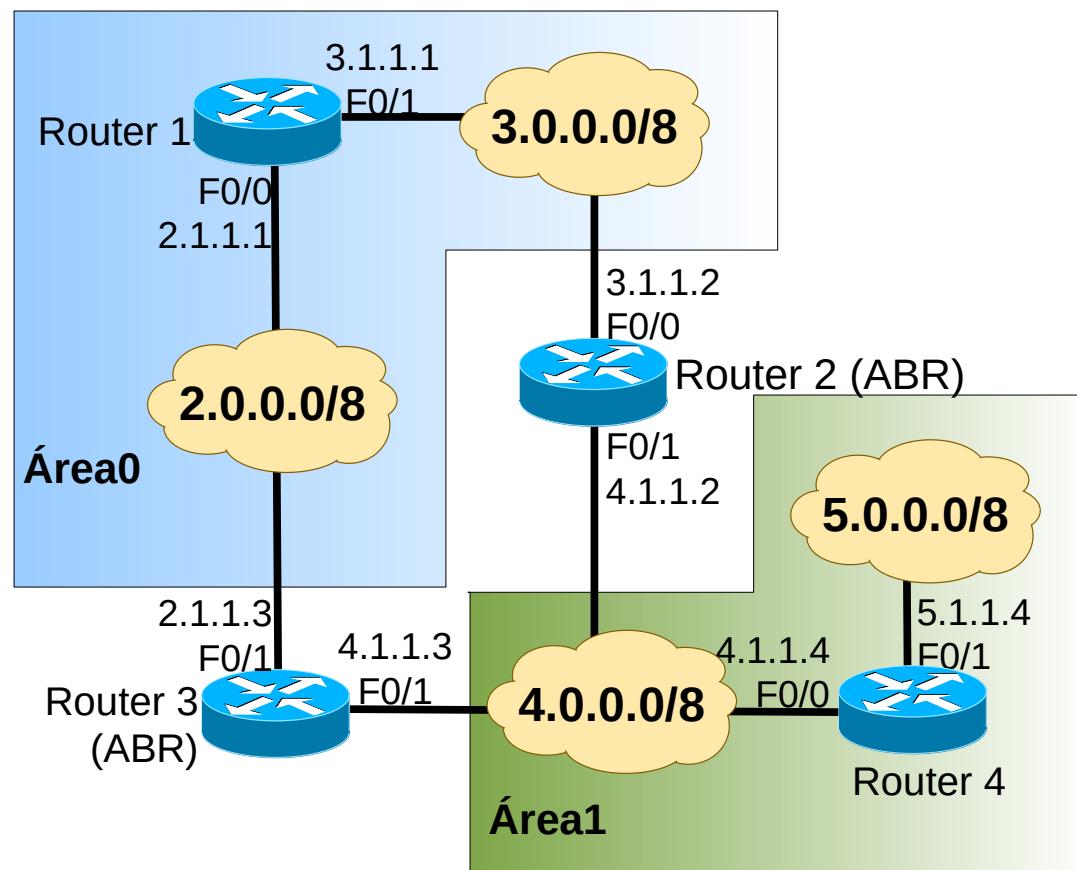


OSPF Routers Types

- Internal router
 - ◆ Routers that have all of their interfaces in the same area.
 - ◆ All routers within the same area have identical LSDBs.
- Backbone router
 - ◆ Routers that sit in the perimeter of the backbone area 0 and that have at least one interface connected to area 0.
 - ◆ Backbone routers maintain OSPF routing information using the same procedures and algorithms as internal routers.
- Area Border Router (ABR)
 - ◆ Routers that have interfaces attached to multiple areas, maintain separate LSDBs for each area to which they connect, and route traffic destined for or arriving from other areas.
 - ◆ Connect area 0 to a non backbone area and are exit points for the area
 - Routing information destined for another area can get there only via the ABR of the local area.
 - ◆ The ideal design is to have each ABR connected to two areas only, the backbone and another area.
 - The recommended upper limit is three areas.
- Autonomous System Boundary Router (ASBR)
 - ◆ Routers that have at least one interface attached to a different routing domain (such as another OSPF autonomous system or a domain using other routing protocol).
 - An OSPF autonomous system consists of all the OSPF areas and the routers within them.
 - ◆ ASBRs can redistribute external routes into the OSPF domain and vice versa.
- A router can be more than one router type.
 - ◆ For example, if a router interconnects to area 0 and area 1, and to a non-OSPF network, it is both an ABR and an ASBR.



OSPF Hierarchical Routing Example



Link State ID: 2.1.1.3	Link State ID: 3.1.1.2
Network Mask: /8	Network Mask: /8
Attached Router: 3.1.1.1	Attached Router: 3.1.1.1
Attached Router: 4.1.1.3	Attached Router: 4.1.1.2

Net Link States from Router 1

Advertising Router: 4.1.1.2	Number of Links: 1
Number of Links: 2	Router Interface address: 3.1.1.2
Router Interface address: 3.1.1.1	TOS 0 Metrics: 10
TOS 0 Metrics: 10	Advertising Router: 4.1.1.3
Router Interface address: 2.1.1.1	Number of Links: 1
TOS 0 Metrics: 10	Router Interface address: 2.1.1.3
	TOS 0 Metrics: 10

Router Link States from Router 1

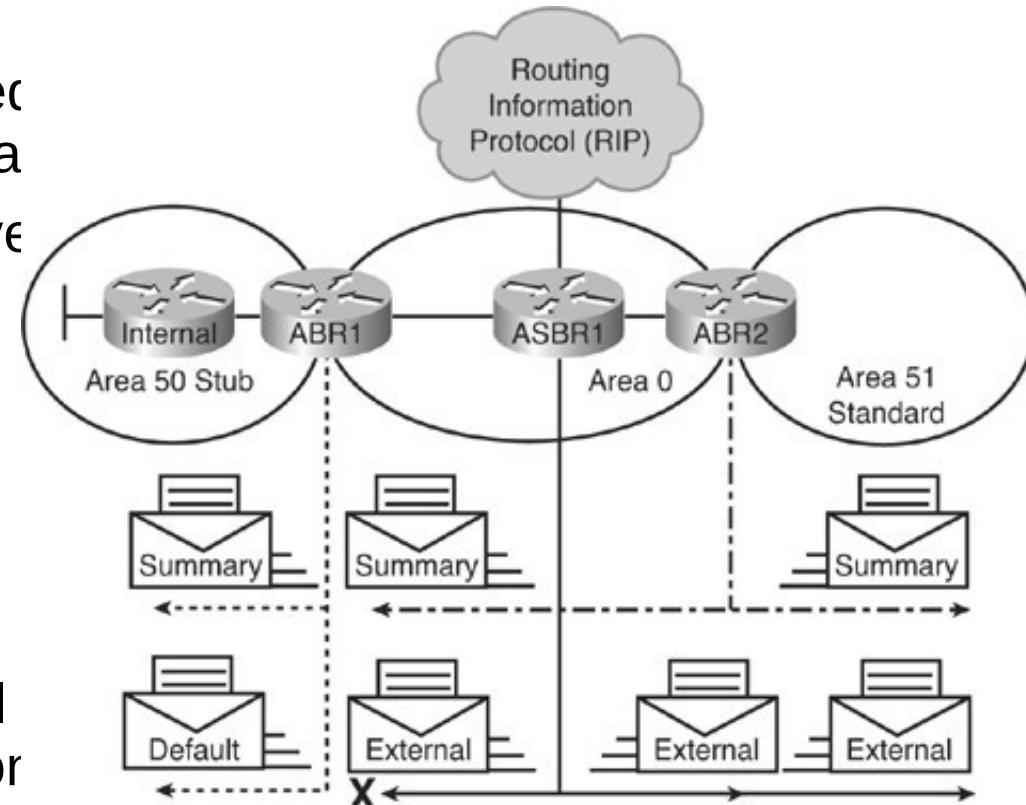
Link State ID: 4.0.0.0	Link State ID: 5.0.0.0
Advertising Router: 4.1.1.2	Advertising Router: 4.1.1.2
Network Mask: /8	Network Mask: /8
TOS: 0 Metric: 10	TOS: 0 Metric: 20
Link State ID: 4.0.0.0	Link State ID: 5.0.0.0
Advertising Router: 4.1.1.3	Advertising Router: 4.1.1.3
Network Mask: /8	Network Mask: /8
TOS: 0 Metric: 10	TOS: 0 Metric: 20

Summary Net Link States from Router 1



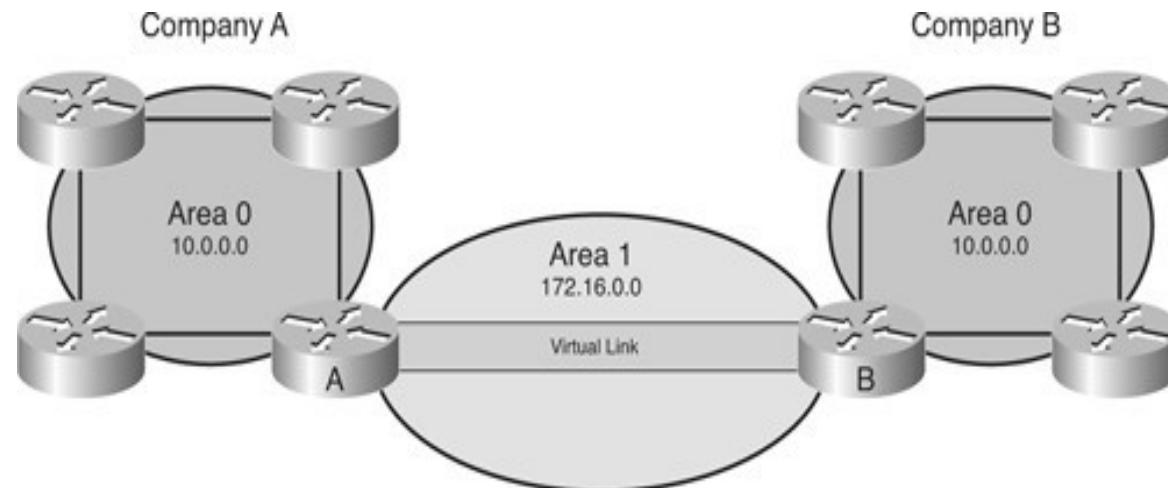
Stub Areas

- Configuring a stub area reduces the size of the LSDB inside an area, resulting in reduced memory requirements for routers in that area
- Routers within the stub area also do not have to run the SPF algorithm as often because they will receive fewer routing updates.
- External network LSAs (type 5), such as those redistributed from other routing protocols into OSPF, are not permitted to flood into a stub area.
- Routing from these areas to a route external to the OSPF autonomous system is based on a default route (0.0.0.0).
 - Stub area ABR when receives an external LSA, sends a 0.0.0.0 LSA to the stub area.
 - If a packet is addressed to a network that is not in the routing table of an internal router, the router automatically forwards the packet to the ABR that originates a 0.0.0.0 LSA.

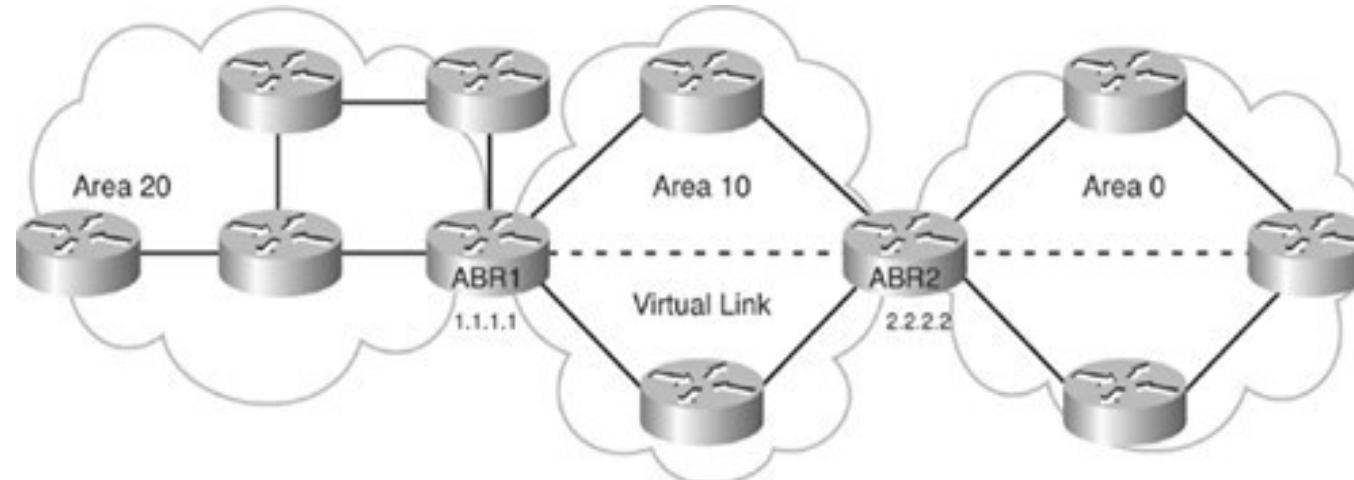


Areas Virtual Links

- Virtual Links can be used to connect a discontiguous Area 0.

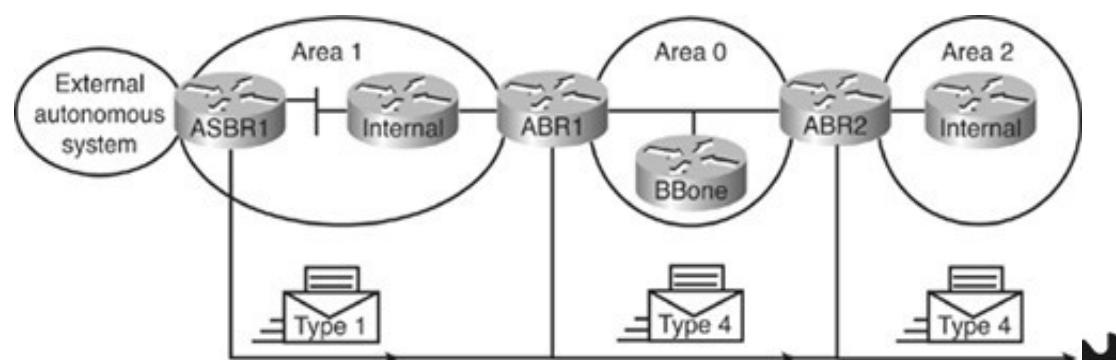
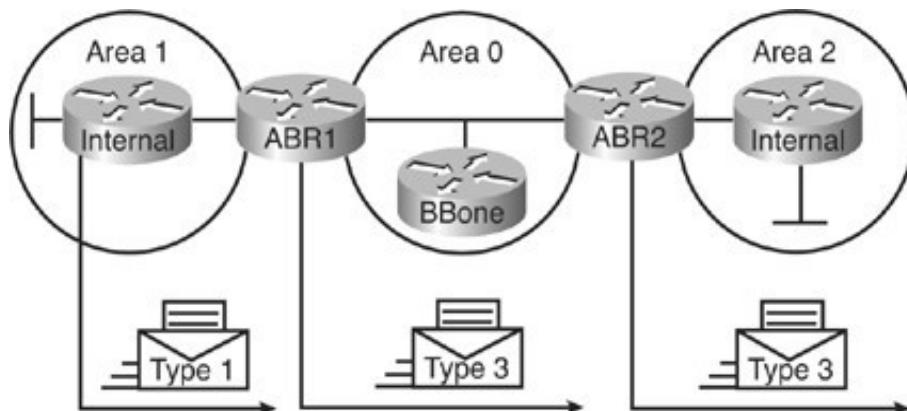


- Virtual Links can be used to connect an area to the backbone Area.



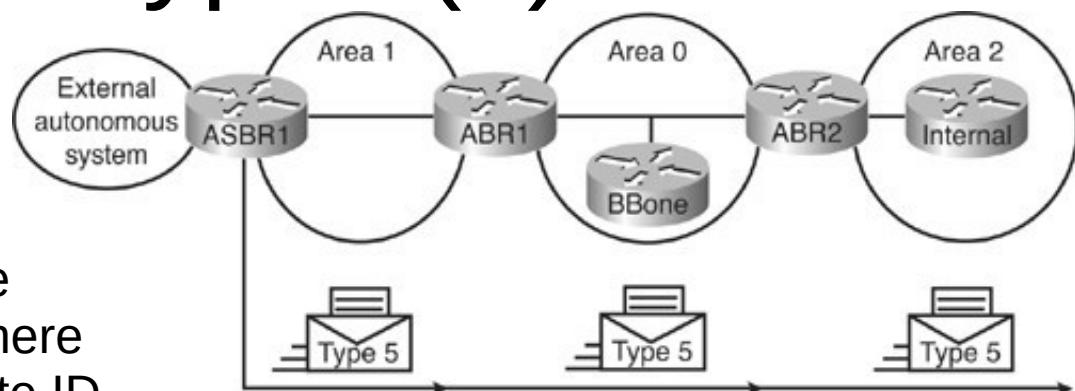
OSPF LSA Types (1)

- Type 1 (Router LSA) - Every router generates router-link advertisements for each area to which it belongs. Router-link advertisements describe the states of the router's links to the area and are flooded only within a particular area. All types of LSAs have 20-byte LSA headers. One of the fields of the LSA header is the link-state ID. The link-state ID of the type 1 LSA is the originating router's ID.
- Type 2 (Network LSA) - DRs generate network link advertisements for multiaccess networks, which describe the set of routers attached to a particular multiaccess network. Network link advertisements are flooded in the area that contains the network. The link-state ID of the type 2 LSA is the DR's IP interface address.
- Types 3 and 4 (Summary LSA) - ABRs generate summary link advertisements. Summary link advertisements describe the following inter-area routes
 - ◆ Type 3 describes routes to the area's networks (and may include aggregate routes) [Inter-Area Prefix LSA].
 - ◆ Type 4 describes routes to ASBRs [Inter-Area Router LSA].



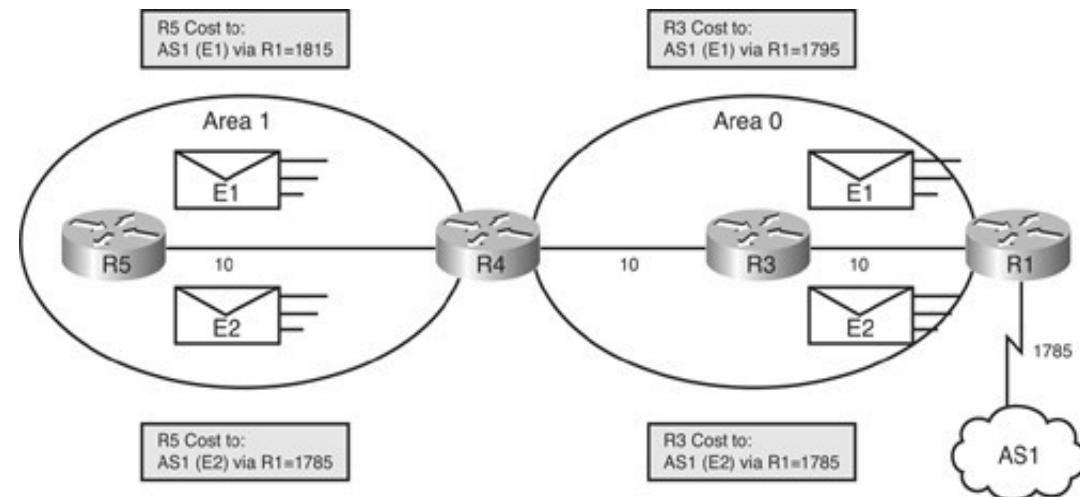
OSPF LSA Types (2)

- Type 5 (AS external LSA) - ASBRs generate autonomous system external link advertisements. External link advertisements describe routes to destinations external to the autonomous system and are flooded everywhere except to any type of stub areas. The link-state ID of the type 5 LSA is the external network number.
- Type 6 (Multicast OSPF LSA) - These LSAs are used in multicast OSPF applications.
- Type 7 (LSAs for NSSAs) - These LSAs are used in NSSAs.
- Type 8 (External attributes LSA for BGP) - These LSAs are used to internetwork OSPF and BGP.
- Types 9, 10, or 11 (Opaque LSAs) - These LSA types are designated for future upgrades to OSPF for distributing application-specific information through an OSPF domain. Standard LSDB flooding mechanisms are used to distribute opaque LSAs. Each of the three types has a different flooding scope.
 - ◆ Type 9 LSAs are not flooded beyond the local network or subnetwork.
 - ◆ Type 10 LSAs are not flooded beyond the borders of their associated area.
 - ◆ Type 11 LSAs are flooded throughout the autonomous system (the same as for Type 5 LSAs). (Opaque LSAs are defined in RFC 5250, The OSPF Opaque LSA Option.)



Types of OSPF Routes

- OSPF intra-area (router LSA) and network LSA
 - ◆ Networks from within the router's area, advertised by way of router LSAs and network LSAs.
- Inter-area (summary LSA)
 - ◆ Networks from outside the router's area but within the OSPF autonomous system, advertised by way of summary LSAs.
- Type 2 external routes (E2)
 - ◆ Networks from outside the OSPF domain, advertised by way of external LSAs.
 - ◆ The cost of OSPF E2 routes is always the external cost only.
 - ◆ Use this type if only one ASBR is advertising an external route to the autonomous system.
 - ◆ This is usually the default for external routes.
- Type 1 external routes (E1)
 - ◆ Networks from outside the OSPF domain, advertised by way of external LSAs.
 - ◆ Calculate the cost by adding the external cost to the internal cost of each link the packet crosses.
 - ◆ Use this type when multiple ASBRs are advertising an external route to the same autonomous system, to avoid suboptimal routing.
 - ◆ Always preferred over Type 2 external routes (E2). Even for higher metrics!



OSPF Area Types

• Standard area

- ◆ This default area type accepts link updates, route summaries, and external routes.

• Backbone area

- ◆ The backbone area is labeled area 0, and all other areas connect to this area to exchange and route information.
- ◆ The OSPF backbone has all the properties of a standard OSPF area.

• Stub area

- ◆ Cannot contain ASBRs (except ABRs that may also be ASBRs).
- ◆ From Area 0 ABR, receives summary routes (LSA Type 3) and automatic default route. External routes (LSA Type 5) are blocked.

• Totally stubby area

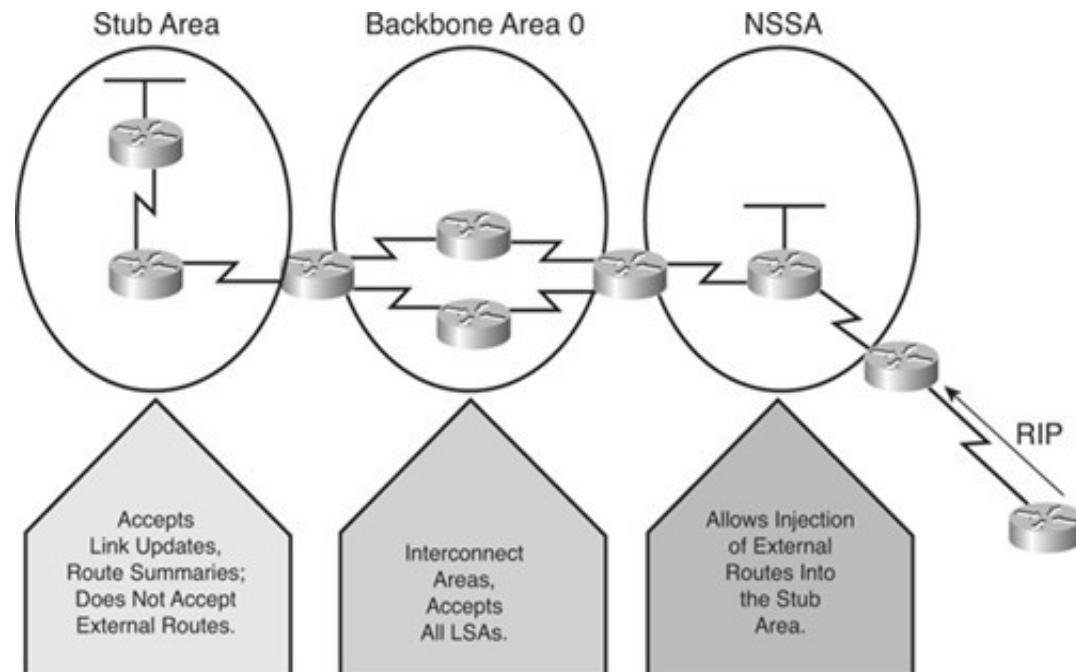
- ◆ Cisco proprietary area type.
- ◆ Cannot contain ASBRs (except ABRs that may also be ASBRs).
- ◆ From Area 0 ABRs, receives automatic default route. Summary routes (LSA Type 3) and external routes (LSA Type 5) are blocked.

• NSSA (Not So Stubby Area)

- ◆ Is an addendum to the OSPF RFC.
- ◆ Contain ASBRs (receives external routes).
- ◆ From Area 0 ABRs, receives summary routes (LSA Type 3). External routes (LSA Type 5) are blocked.
- ◆ No automatic default route is sent to NSSA Area by ABR.
- ◆ Uses a LSA type 7 to announce external routes to Area 0 ABR, ABR transforms the LSA Type 7 into a LSA type 5 and sends it to Area 0.

• Totally stubby NSSA

- ◆ Contain ASBRs (receives external routes).
- ◆ From Area 0 ABRs, receives automatic default route. Summary routes (LSA Type 3) and external routes (LSA Type 5) are blocked.
- ◆ Uses a LSA type 7 to announce external routes to Area 0 ABR, ABR transforms the LSA Type 7 into a LSA type 5 and sends it to Area 0.



Routing - OSPFv3

- Based on OSPFv2, with enhancements:
 - ◆ Uses IPv6 for transport
 - ◆ Distributes IPv6 prefixes
 - ◆ Uses multicast group addresses FF02::5 (OSPF IGP) and FF02::6 (OSPF IGP Designated Routers)
 - ◆ Runs over a link rather than a subnet
 - ◆ Multiple instances per link
 - ◆ Topology not IPv6-specific
 - ✚ Router ID, Area ID, Link ID remain a 4 bytes number
 - ✚ Neighbors are always identified by Router ID (4 bytes)
 - ✚ With an additional table with mapping between IPv6 prefixes and Link IDs
 - ◆ Uses link-local addresses as IPv6 source addresses



OSPFv3 - LSA Types

- Link LSA (Type 8)
 - ◆ Informs neighbors of link local address
 - ◆ Informs neighbors of IPv6 prefixes on link
- Intra-Area Prefix LSA (Type 9)
 - ◆ Associates IPv6 prefixes with a network or router
- Flooding scope for LSAs has been generalized
 - ◆ Three flooding scopes for LSAs
 - ◆ Link-local
 - ◆ Area
 - ◆ AS
- LSA Type encoding expanded to 16 bits
 - ◆ Includes flooding scope



Integrated System-Integrated System (IS-IS) Protocol

- IS-IS was defined in 1992 in the ISO/IEC recommendation 10589.
- IS-IS is a link-state routing protocol.
 - ◆ Provides fast convergence and excellent scalability.
 - ◆ Very efficient in its use of network bandwidth.
- Uses Dijkstra's Shortest Path First algorithm (SPF).
- Types of packets
 - ◆ IS-IS Hello packet (IIH), Link State Packet (LSP), Partial Sequence Number Packet (PSNP) and Complete Sequence Number Packet (CSNP).
- Link States are called LSPs
 - ◆ Contain all information about one router adjacencies, connected IP prefixes, OSI end systems, area addresses, etc.
 - ◆ One LSP per router (plus fragments).
 - ◆ One LSP per LAN network.
- IS-IS has 2 layers of hierarchy
 - ◆ The backbone is called level-2.
 - ◆ Areas are called level-1.
 - ◆ A router can take part in L1 and L2 inter-area routing (or inter-level routing).

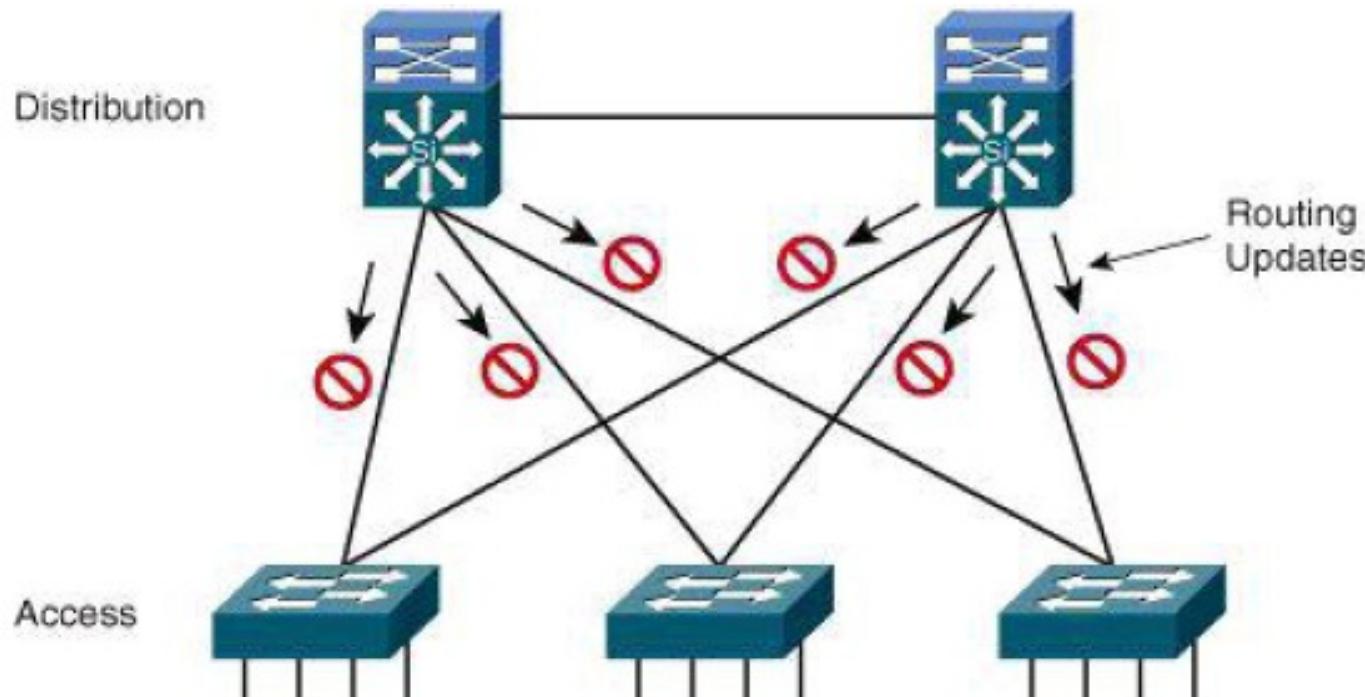


Enhanced Interior Gateway Routing Protocol (EIGRP) Protocol

- EIGRP is a Cisco-proprietary protocol that combines the advantages of link-state and distance vector routing protocols.
- EIGRP has its roots as a distance vector routing protocol and is predictable in its behavior.
- What makes EIGRP an advanced distance vector protocol is the addition of several link-state features, such as dynamic neighbor discovery.
 - ◆ EIGRP Maintains a Neighbor Table, a Topology Table, and a Routing Table.
- EIGRP has Variable-length subnet masking (VLSM) support.
- Has a sophisticated metric that considers five criteria:
 - ◆ Two by default:
 - Bandwidth - The smallest (slowest) bandwidth between the source and destination.
 - Delay - The cumulative interface delay along the path.
 - ◆ Available, are not commonly used, because they typically result in frequent recalculation of the topology table:
 - Reliability - The worst reliability between the source and destination, based on keepalives.
 - Loading - The worst load on a link between the source and destination based on the packet rate and the interface's configured bandwidth.
 - Maximum transmission unit (MTU) - The smallest MTU in the path.
- A significant advantage of EIGRP (and IGRP) over other protocols is its support for unequal metric load balancing that allows administrators to better distribute traffic flow in their networks.



Passive Interfaces on Access Layer

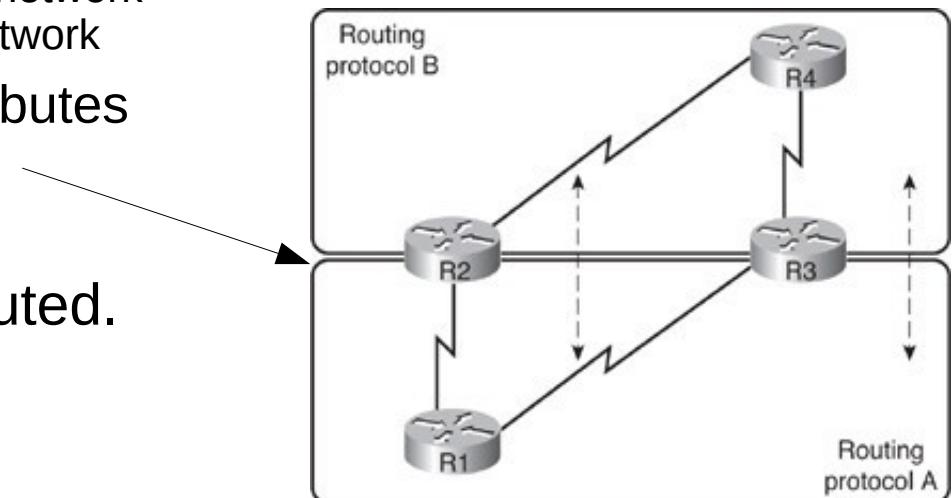
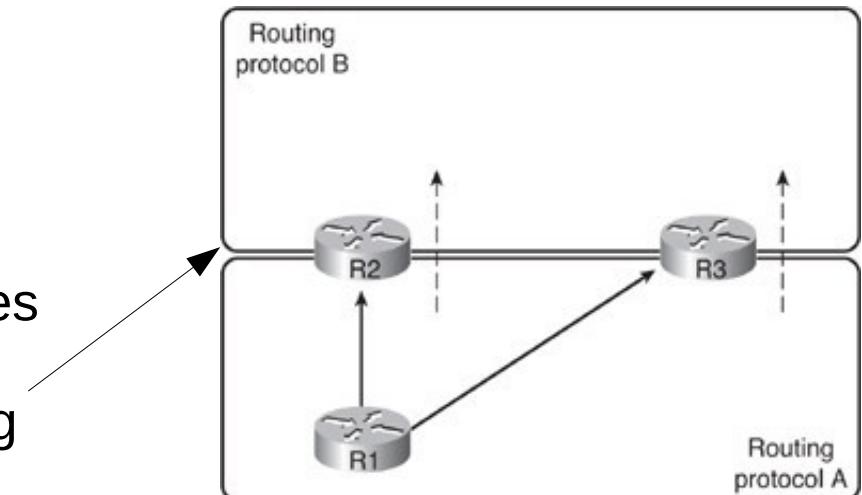


- As a recommended practice, limit unnecessary L3 routing peer adjacencies by configuring the ports toward Layer 2 access switches as passive.
 - Suppress the advertising of routing updates.
 - If a distribution switch does not receive L3 routing updates from a potential peer on a specific interface, it does not form a neighbor adjacency with the potential peer across that interface.



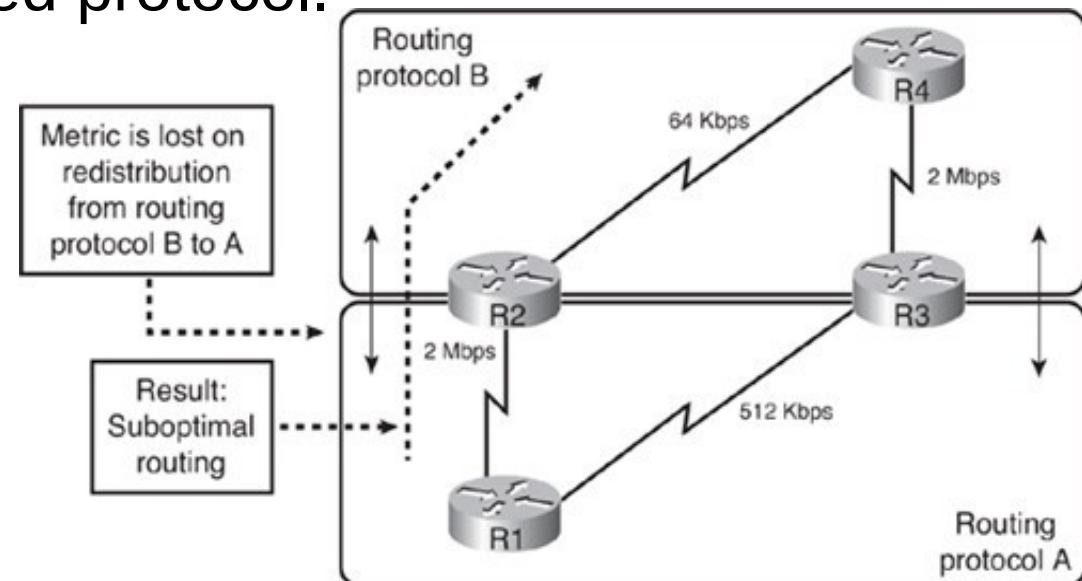
Route Redistribution

- Domains with different routing protocols can exchange routes.
 - ◆ This is called route redistribution.
 - ◆ One-way redistribution - Redistributions only the networks learned from one routing protocol into the other routing protocol.
 - Uses a default or static route so that devices in that other part of the network can reach the first part of the network
 - ◆ Two-way redistribution - Redistributions routes between the two routing processes in both directions
 - ◆ Static routes can also be redistributed.



Redistribution Issues

- Lost metric from redistributed protocol.
 - ◆ It is not possible to achieve an optimal overall routing.

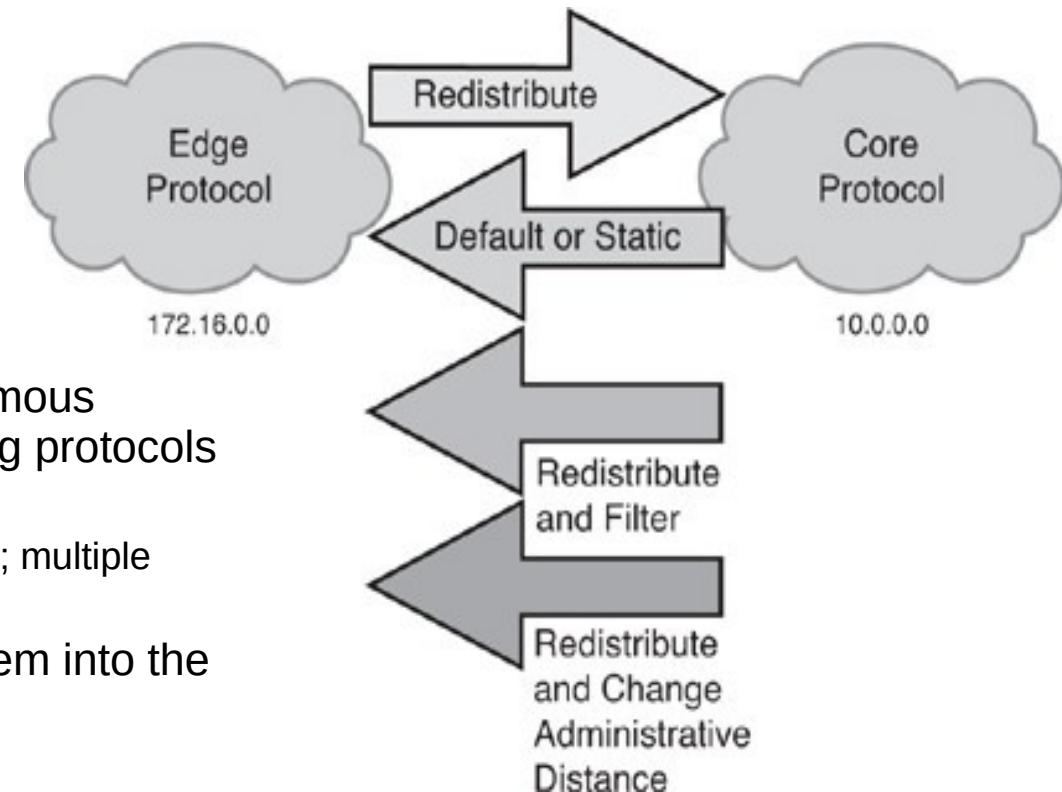


- Preventing Routing Loops in a Redistribution Environment.
 - ◆ Safest way to perform redistribution is to redistribute routes in only one direction, on only one boundary router within the network.
 - ◆ However, that this results in a single point of failure in the network.
 - ◆ If redistribution must be done in both directions or on multiple boundary routers, the redistribution should be tuned to avoid problems such as suboptimal routing and routing loops.



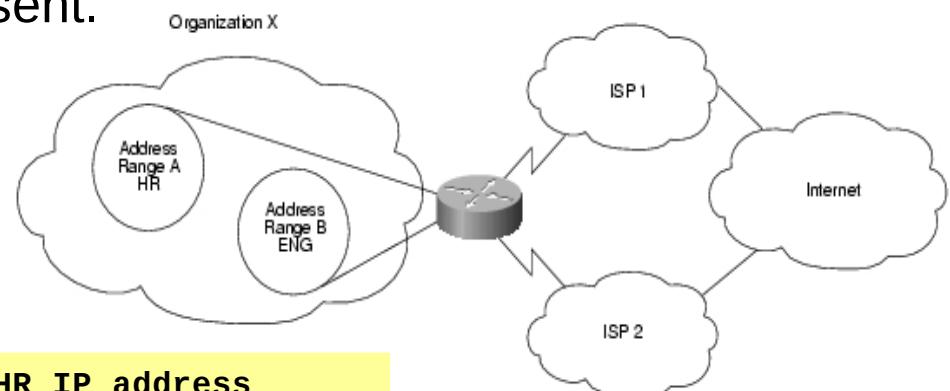
Redistribution Techniques

- Redistribute a default route from the core autonomous system into the edge autonomous system, and redistribute routes from the edge routing protocols into the core routing protocol.
 - ◆ This technique helps prevent route feedback, suboptimal routing, and routing loops.
- Redistribute multiple static routes about the core autonomous system networks into the edge autonomous system, and redistribute routes from the edge routing protocols into the core routing protocol.
 - ◆ This method works if there is only one redistribution point; multiple redistribution points might cause route feedback.
- Redistribute routes from the core autonomous system into the edge autonomous system with filtering to block out inappropriate routes.
 - ◆ For example, when there are multiple boundary routers, routes redistributed from the edge autonomous system at one boundary router should not be redistributed back into the edge autonomous system from the core at another redistribution point.
- Redistribute all routes from the core autonomous system into the edge autonomous system, and from the edge autonomous system into the core autonomous system, and then modify the administrative distance associated with redistributed routes so that they are not the selected routes when multiple routes exist for the same destination.



Policy-Based Routing (PBR)

- PBR allows the operator to define routing policy other than basic destination-based routing using the routing table.
- PBR rules can be used to match source and destination addresses, protocol types, and end-user applications.
- When a match occurs, a set command can be used to define the interface or next-hop address to which the packet should be sent.



```
access-list 1 permit 209.165.200.225  
access-list 2 permit 209.165.200.226  
!  
interface ethernet 1  
 ip policy route-map ChooseISP  
!  
route-map ChooseISP permit 10  
 match ip address 1  
 set ip next-hop 209.165.200.227  
!  
route-map ChooseISP permit 20  
 match ip address 2  
 set ip next-hop 209.165.200.228
```

!From HR IP address
!From ENG IP address

Defines order of the rules

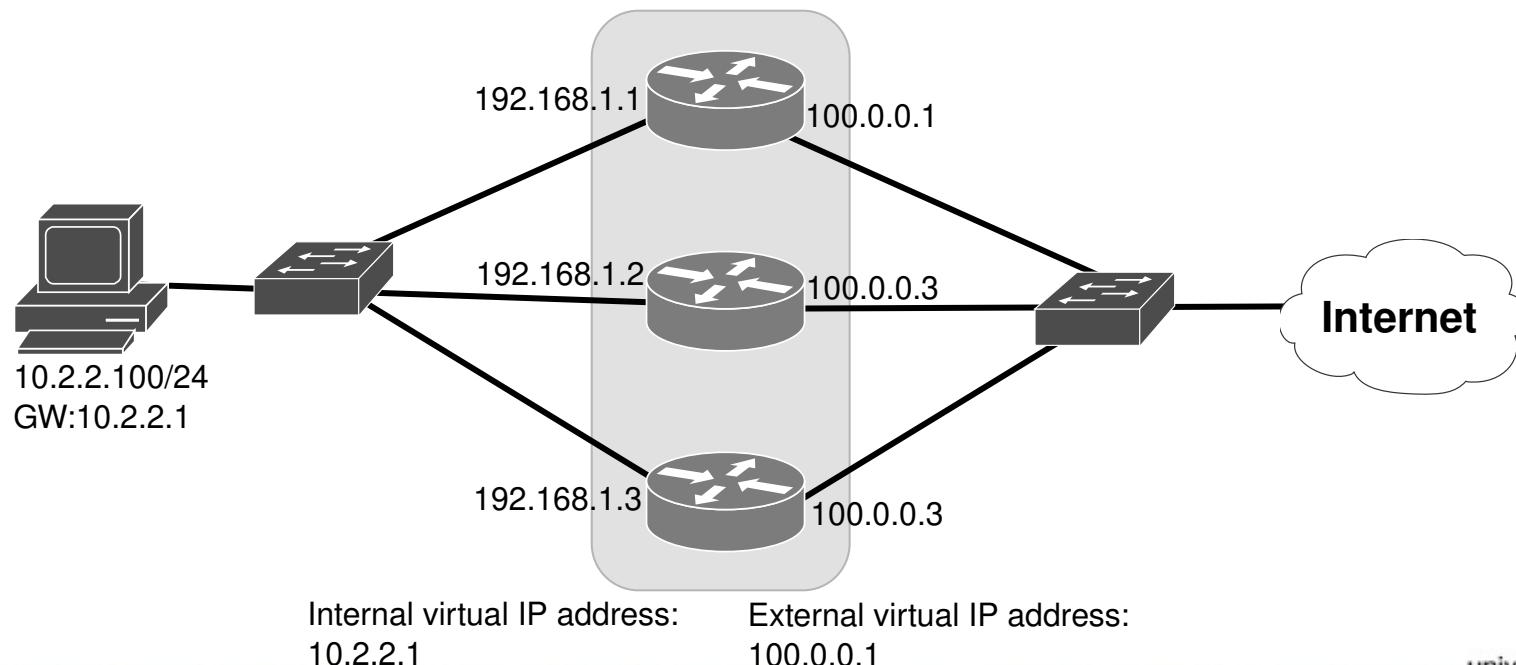
!To ISP2 next-hop

!To ISP1 next-hop



Virtual Router Redundancy Protocol

- VRRP is a standard protocol defined by the IETF in RFC 3768 to create a virtual gateway.
 - ◆ Cisco has HSRP (Hot Standby Routing Protocol) which is very similar.
- A cluster of routers can be handled as a single router using a
 - ◆ Virtual IP address and virtual MAC address.
 - ◆ Default gateway to the clients.
 - ◆ One of the individual routers acts as master (working router), however, upon a failure one of the other individual routers becomes the working router.
 - ◆ Router states are maintained and verified using a multicast group.



Traffic Tunneling & Overlay Networks

Redes de Comunicações II

**Licenciatura em
Engenharia de Computadores e Informática
DETI-UA**

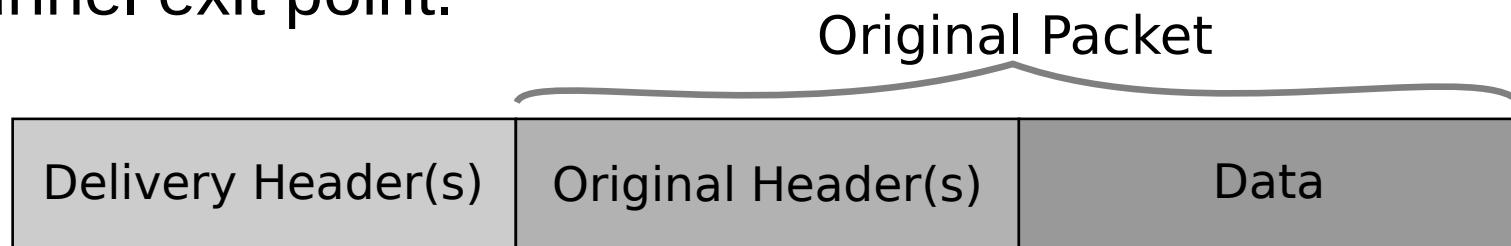


universidade de aveiro

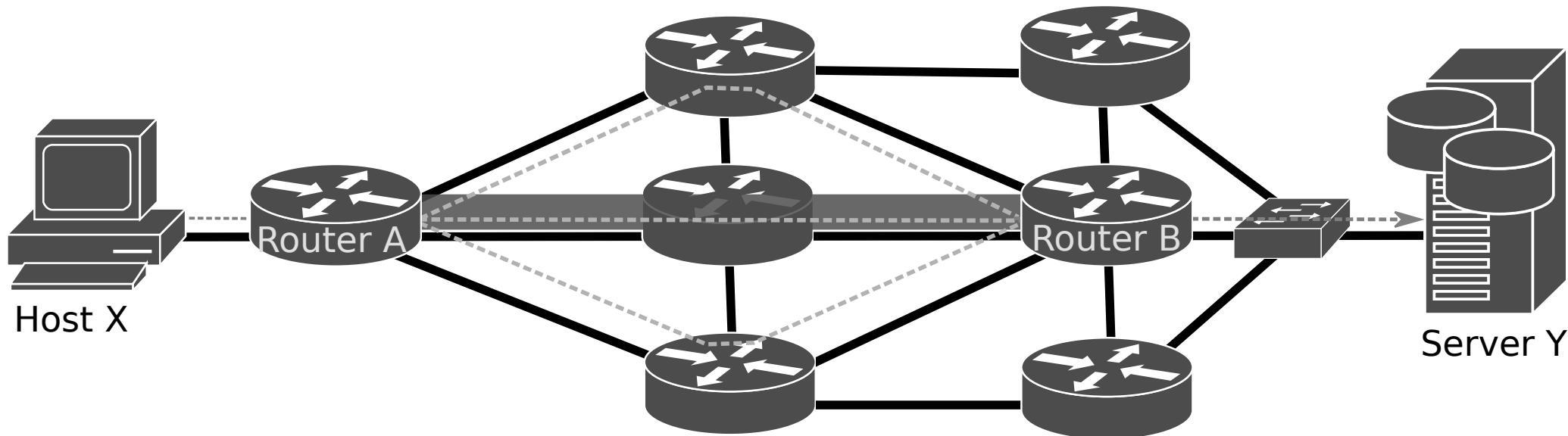
deti.ua.pt

Traffic Tunnel Concept

- Main purposes
 - ◆ Guarantee that a packet that reaches a network node will reach a specific secondary network node independently of the intermediary nodes routing processes,
 - ◆ Guarantee the delivery of a packet to a remote node when the intermediary nodes do not support the original packet network protocol, and,
 - ◆ Define a virtual channel that adds additional data transport features in order to provide differentiated QoS, security requirements and/or optimized routing.
- Achieved by adding, at the tunnel entry point, one or more protocol headers to the original packets to handle their delivery to the tunnel exit point.



Tunnel End-Points



Delivery protocol(s)	Original protocol(s)	Data
Source: A address Destination: B address	Source: X address Destination: Y address	

Virtual Tunnel Interface (VTI)

- Logical construction that creates a virtual network interface that can be handled as any other network interface within a network equipment.
- A tunnel does not require to have any network addresses other the ones already bound to the end-point router.
- However, most implementations impose that a network address must be bound to a tunnel interface in order to enable IP processing on the interface.
 - ◆ The tunnel interface may have a explicitly bound network address or reuse an address of another interface already configured on the router.

```
1 #interface Tunnel 1
2 #ip address 10.1.1.1 255.255.255.252
3 #ipv6 address 2001:A::A:1/64
4 #ip unnumbered FastEthernet0/0
5 #ipv6 unnumbered FastEthernet0/0
6 #ip ospf cost 10
7 #ipv6 ospf 1 area 0
8 #tunnel mode ipip
9 #tunnel source FastEthernet0/0
10 #tunnel destination 200.2.2.2
```



VTI Requirements

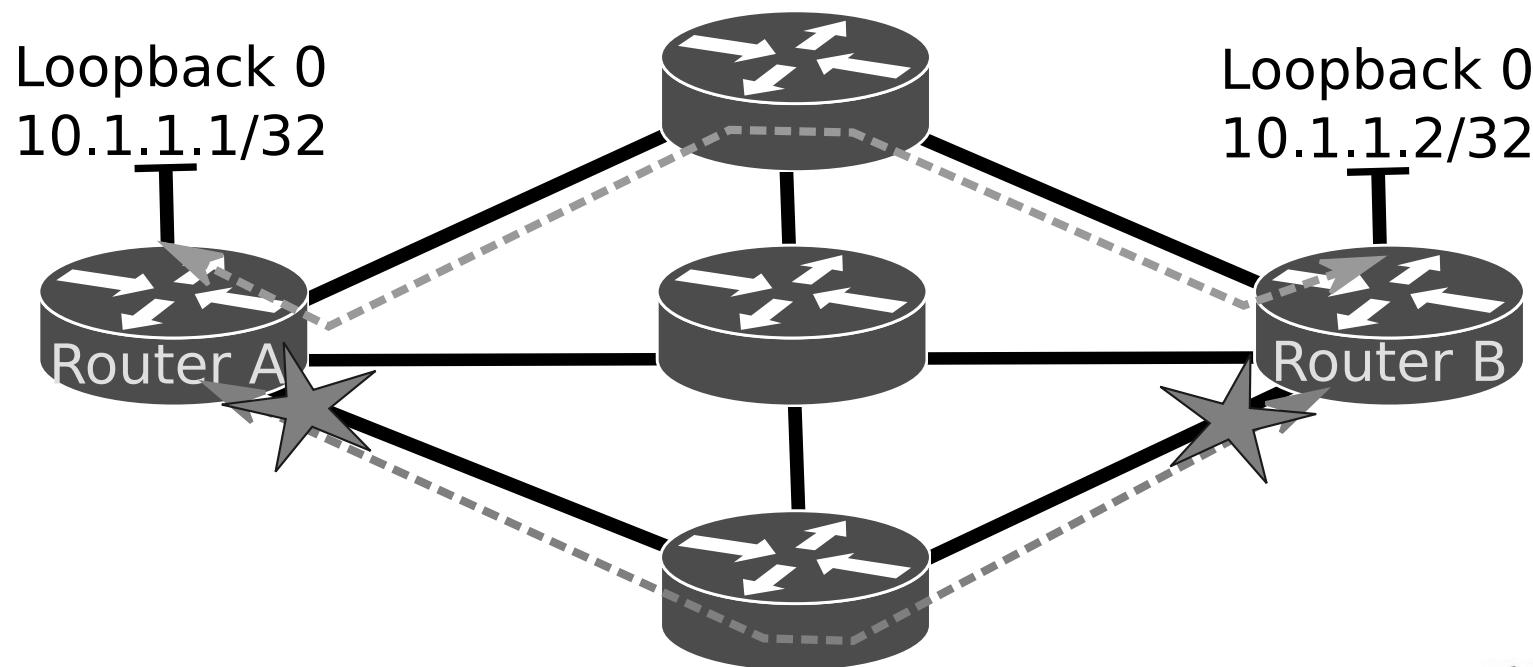
- A numeric identifier,
- A bounded IP address, this will enable IP processing,
 - ◆ Add the tunnel interface to the routing table and allow routing via the interface,
- A defined mode or type of tunnel,
 - ◆ Availability of tunnel models depends on the Router model, operating software and licenses.
- Tunnel source,
 - ◆ Defined as the name of the local interface or IPv4/IPv6 address depending on the type of the tunnel.
- Tunnel destination,
 - ◆ Defined as a domain name or IPv4/IPv6 address depending on the type of the tunnel.
 - ◆ This definition is not mandatory for all types of tunnels because in some cases the tunnel end-point is determined dynamically.
- May optionally have additional configurations for routing, security and QoS purposes.

```
1 #interface Tunnel 1
2 #ip address 10.1.1.1 255.255.255.252
3 #ipv6 address 2001:A::A:1/64
4 #ip unnumbered FastEthernet0/0
5 #ipv6 unnumbered FastEthernet0/0
6 #ip ospf cost 10
7 #ipv6 ospf 1 area 0
8 #tunnel mode ipip
9 #tunnel source FastEthernet0/0
10 #tunnel destination 200.2.2.2
```



Loopback Interfaces as End-Points

- Loopback interface is another logical construction that creates a virtual network interface completely independent from the remaining physical and logical router network interfaces.
- The main propose of a loopback interface is to provide a network address to serve as router identifier in remote network configurations and distribute algorithms.
- The main advantage of using loopback interfaces as tunnel end-points, is the creation of a tunnel not bounded to any individual network card/link that may fail.



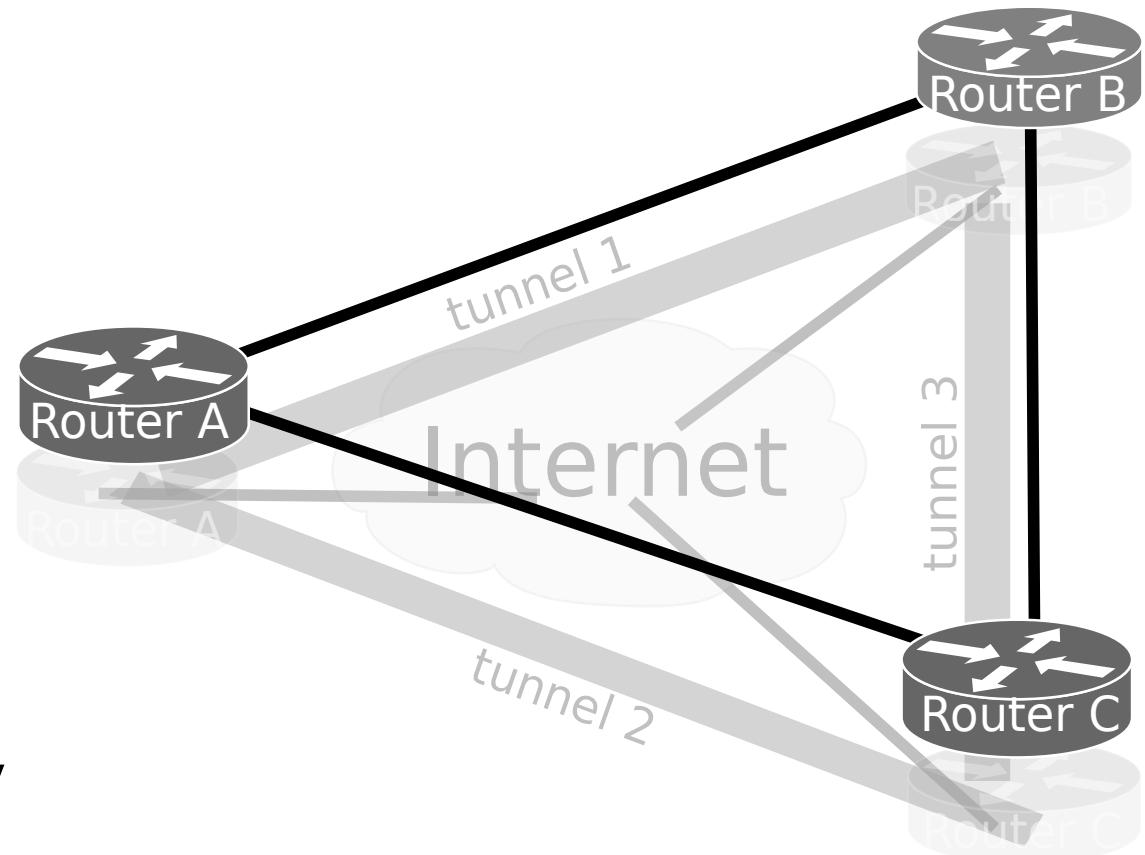
IP Tunnel Types

- IPv4-IPv4
 - ◆ Original IPv4 packets are delivered using IPv4 as network protocol.
- GRE IPv4
 - ◆ Original packets protocol (any network protocol) is defined by GRE header and delivered using IPv4 as network protocol.
- IPv6-IPv6
 - ◆ Original IPv6 packets are delivered using IPv6 as network protocol.
- GRE IPv6
 - ◆ Original packets protocol (any network protocol) is defined by a GRE header and delivered using IPv6 as network protocol.
- IPv6-IPv4
 - ◆ Original IPv6 packets are delivered using IPv4 as network protocol.
- IPv4-IPv6
 - ◆ Original IPv4 packets are delivered using IPv6 as network protocol.



Overlay Network

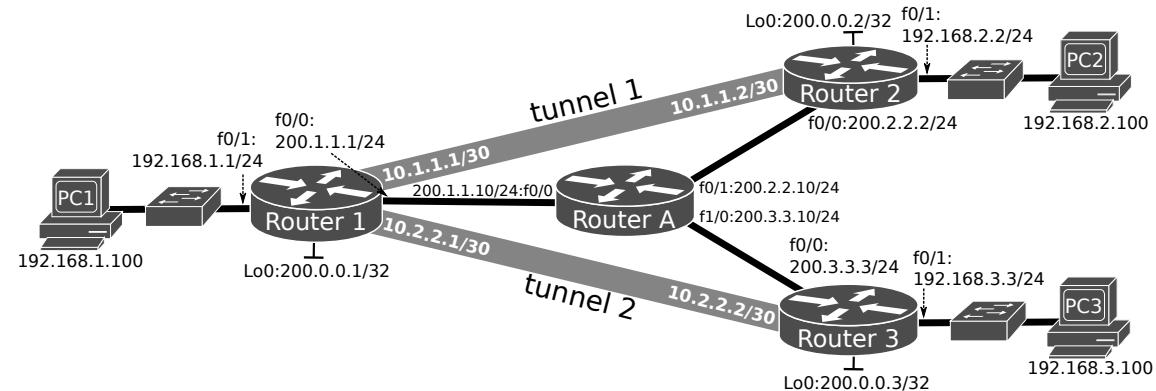
- An overlay network can be defined as a virtual network defined over another network.
 - ◆ For a specific purpose like private transport/routing policies, QoS, security.
- The underlying network can be physical or also virtual.
 - ◆ May result in multiple layers of overlay networks.
- When any level of privacy protocol is present on an overlay network is designated by Virtual Private Network (VPN).



Routing Through/Between Tunnels

- Static Routes

```
1 #ip route 192.168.2.0 255.255.255.0 Tunnel1
2 #ip route 192.168.2.0 255.255.255.0 10.1.1.2
3 #ipv6 route 2001:A:1::/64 Tunnel1
4 #ipv6 route 2001:A:1::/64 2001:0:0::2
5 #ip route 192.168.2.100 255.255.255.255 10.1.1.2
6 #ipv6 route 2001:A:1::100/128 2001:0:0::2
```



- Route-maps

```
1 #access-list 100 permit ip host 192.168.1.100 192.168.2.0 255.255.255.0
2 #route-map routeT1
3 #match ip address 100
4 #set ip next-hop 10.1.1.2
5 #interface FastEthernet0/1
6 #ip policy route-map routeT1
```

- Dynamic Routing

- Multiple (distinct) routing processes.
 - One per overlay network, and
 - One for the underlying network.

```
1 #router ospf 1
2 #network 200.1.1.0 0.0.0.255 area 0
3 #network 200.0.0.1 0.0.0.0 area 0
4 !
5 #router ospf 2
6 #network 10.0.0.0 0.255.255.255 area 0
7 #network 192.168.0.0 0.0.255.255 area 1
```

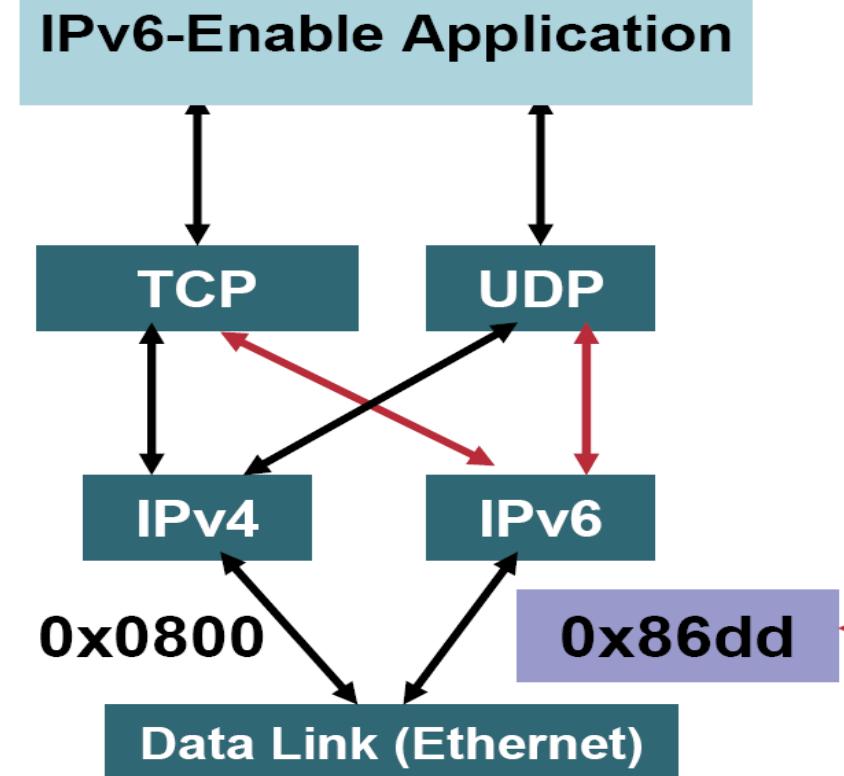
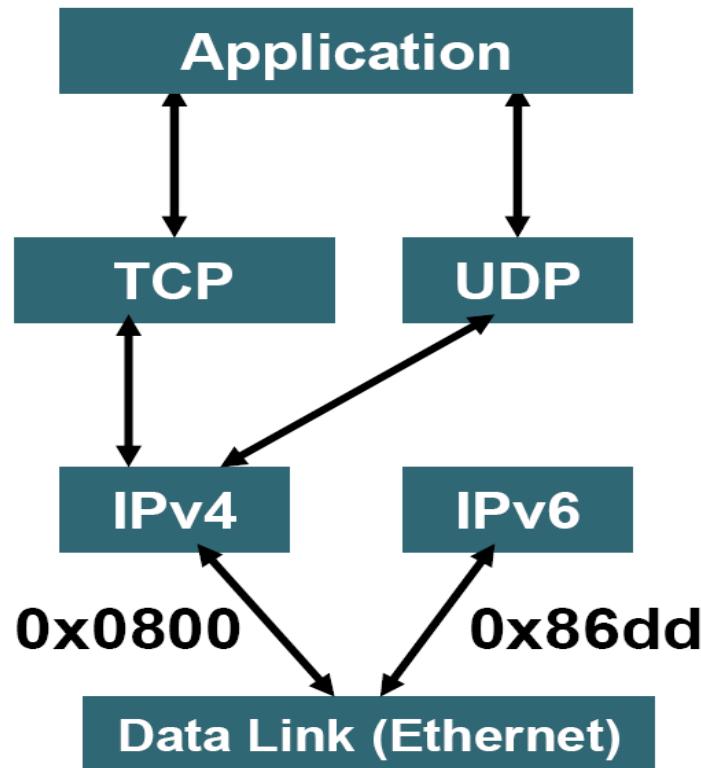


IPv6 Deployment Techniques

- Deploying IPv6 using dual-stack backbones
 - ◆ IPv4 and IPv6 applications coexist in a dual IP layer routing backbone
 - ◆ All routers in the network need to be upgraded to be dual-stack
- IPv6 over IPv4 tunnels
 - ◆ Manually configured
 - With and without Generic Routing Encapsulation (GRE)
 - ◆ Semiautomatic tunnel mechanisms
 - ◆ Fully automatic tunnel mechanisms (IPv4-compatible and 6to4)



Dual Stack



- Applications may talk to both
- Choice of the IP version is based on DNS responses and application preferences



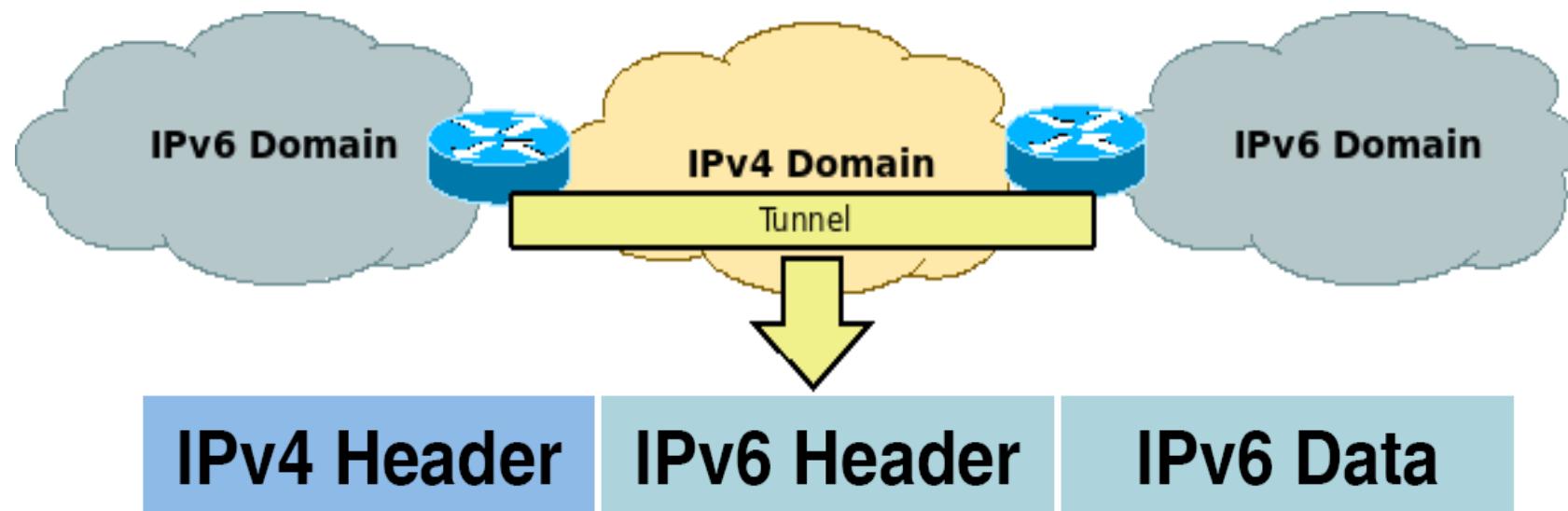
IPv6 Overlay Tunneling

- Manual
 - ◆ IPv6 Manually Configured IPv6 over IPv4
 - ◆ IPv6 over IPv4 GRE Tunnel
- Semi-automatic mechanisms
 - ◆ Tunnel Broker
 - ◆ Teredo
 - ◆ Dual Stack Transition Mechanism (DSTM)
- Automatic mechanisms
 - ◆ Automatic IPv4 Compatible Tunnel (deprecated)
 - ◆ 6to4 Tunnel
 - ◆ ISATAP Tunnels



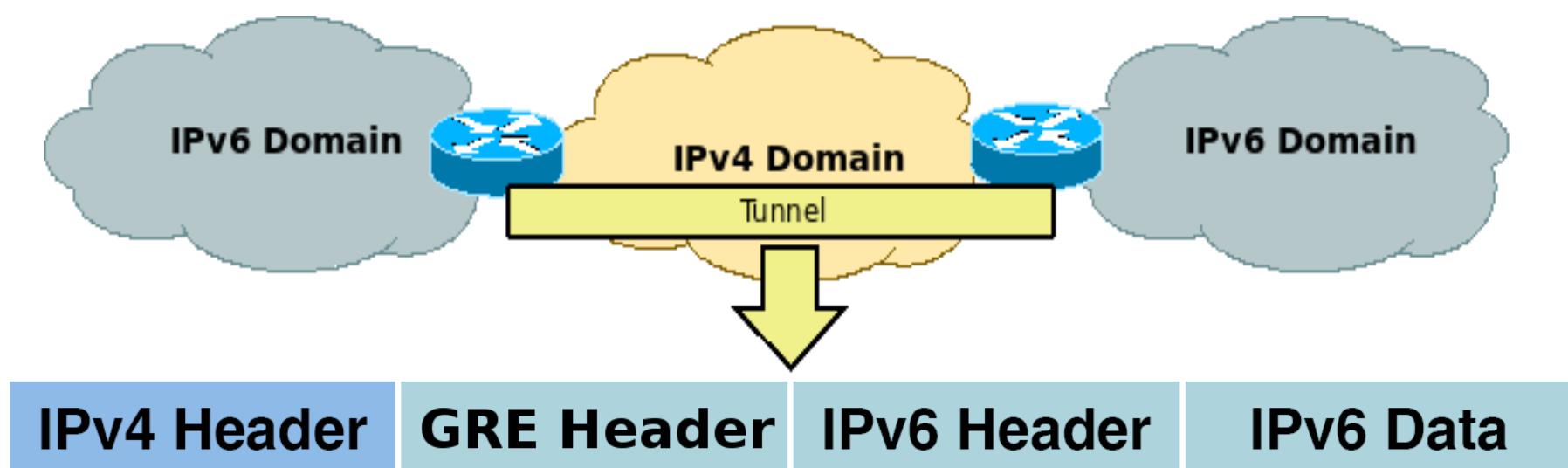
IPv6 Manually Configured

- Permanent link between two IPv6 domains over an IPv4 backbone
- Primary use is for stable connections that require regular secure communication between
 - ◆ Two edge routers, end system and an edge router, or for connection to remote IPv6 networks
- Tunnel between two points
- Complex management



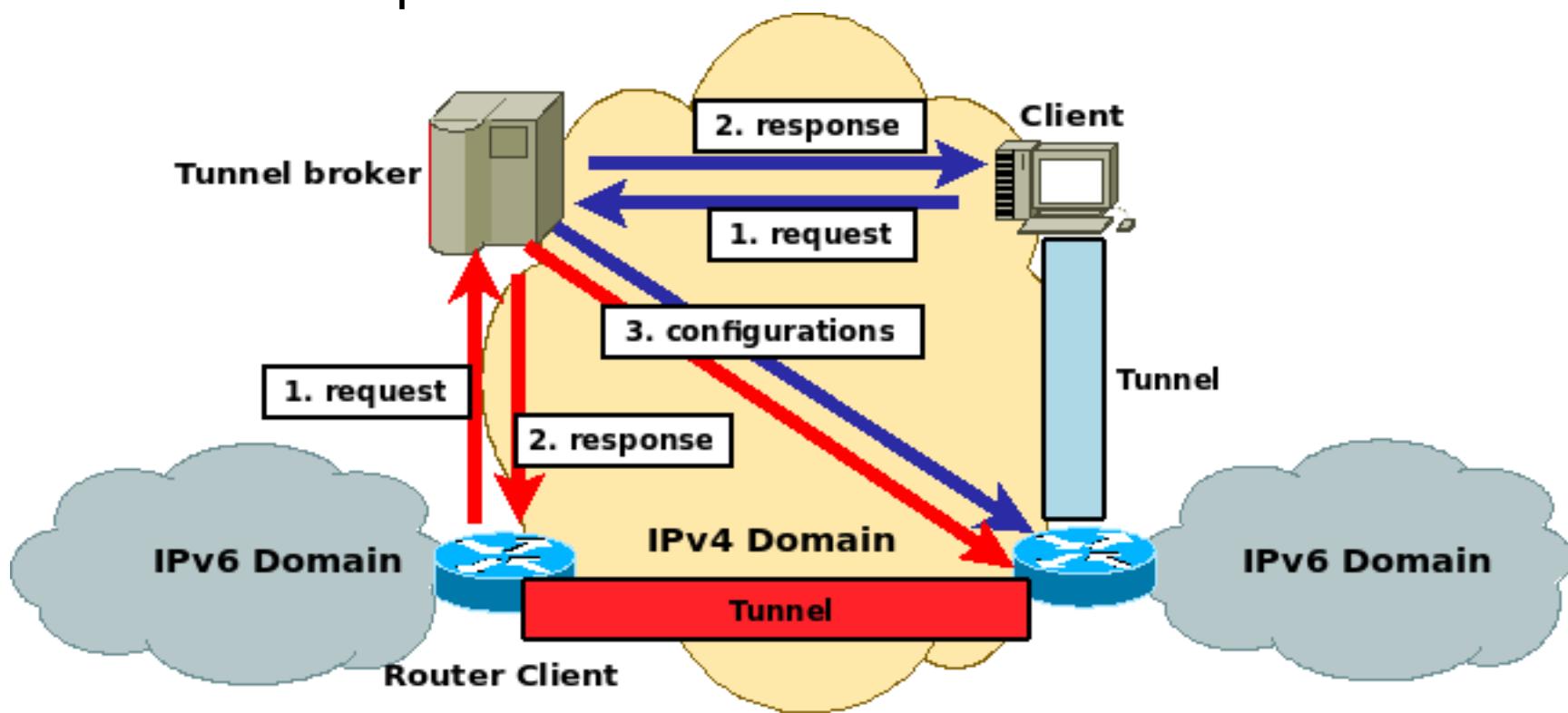
IPv6 over IPv4 GRE Tunnel

- Uses the standard GRE tunneling technique
 - ◆ GRE – Generic Route Encapsulation
- Also must be manually configured
- Primary use is for stable connections that require regular stable communications
- IPv4 over IPv6 also possible



Tunnel Broker

- A tunnel broker service allows IPv6 applications on dual-stack systems access to an IPv6 backbone
- Automatically manages tunnel requests and configuration
- Potential security implications
 - ◆ Broker is a single point of failure
- Most common implementation: Teredo.



Automatic IPv4 Compatible Tunnel

- IPv4 tunnel end-point address is embedded within the destination IPv6 address
- An automatic IPv4-compatible tunnel can be configured between edge routers or between an edge router and an end system.
- Systems must be dual-stack
- Communication only with other IPv4-compatible sites
- This tunneling technique is currently deprecated



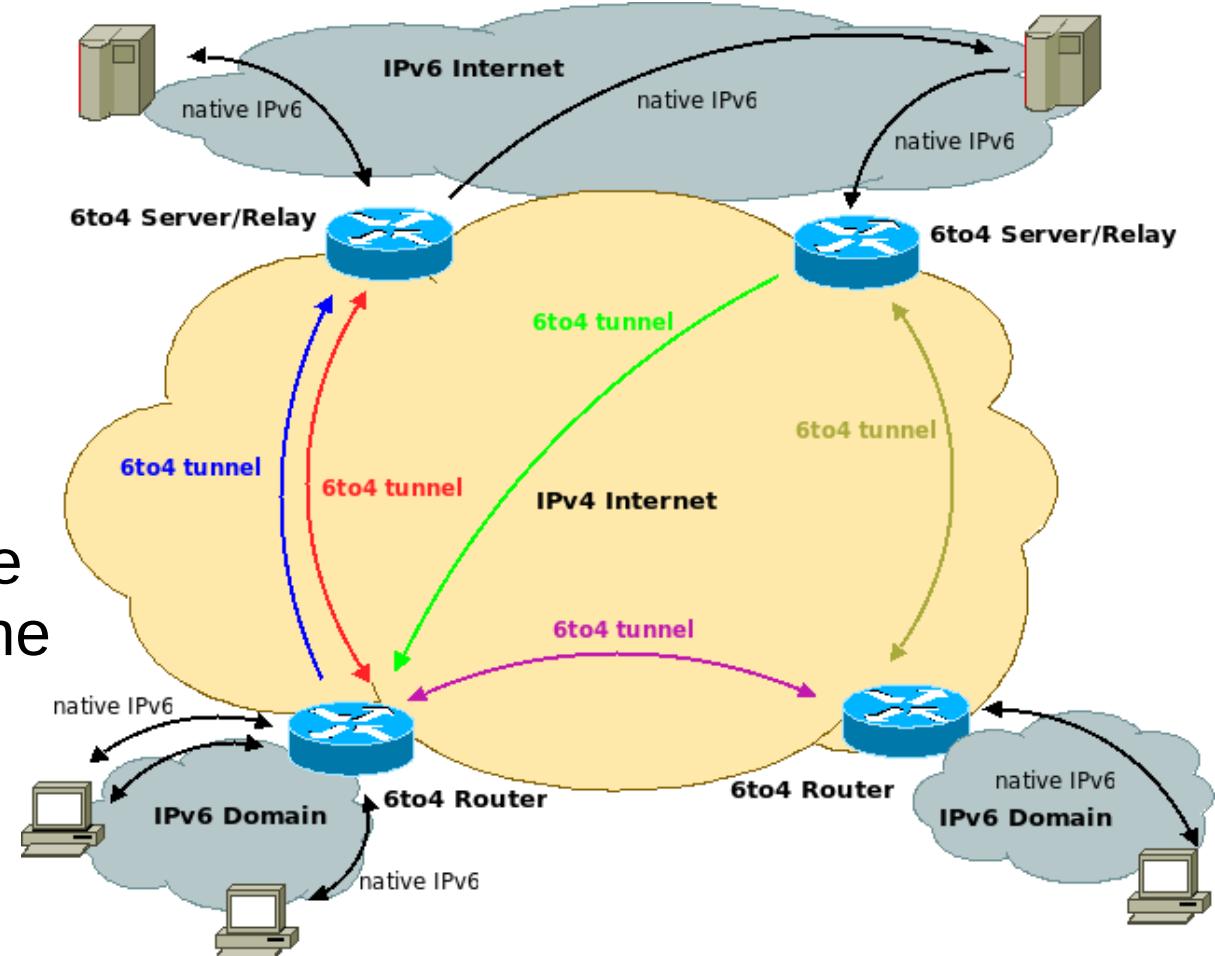
Automatic 6to4 Tunnels

- IPv4 tunnel end-point address is embedded within the destination IPv6 address
- Automatic 6to4 tunnel allows isolated IPv6 domains to connect over an IPv4 network
- Unlike the manually configured tunnels are not point-to-point, they are multipoint tunnels
- 6to4 host/router needs to have a globally addressable IPv4 address
- Cannot be located behind a NAT box
- Unless the NAT box supports protocol 41 packets forwarding
- Address format is:



6to4 Relay Routers

- 6to4 router
- Connects 6to4 hosts from a IPv6 domain and
 - Other 6to4 routers
 - The IPv6 Internet through a 6to4 relay router
- 6to4 relay router
- Connects 6to4 routers on the IPv4 Internet and hosts on the IPv6 Internet.



ISATAP Tunnels

- Intra-site Automatic Tunnel Address Protocol
- Point-to-multipoint tunnels that can be used to connect systems within a site
- Used to tunnel IPv4 within an administrative domain to create a virtual IPv6 network over a IPv4 network
- Scalable approach for incremental deployment
- Encode IPv4 Address in IPv6 Address within the interface ID

64-bit Unicast Prefix

/64

Interface ID
0000:5EFE: **IPv4 Address**



External Routing (BGP and MP-BGP)

Redes de Comunicações II

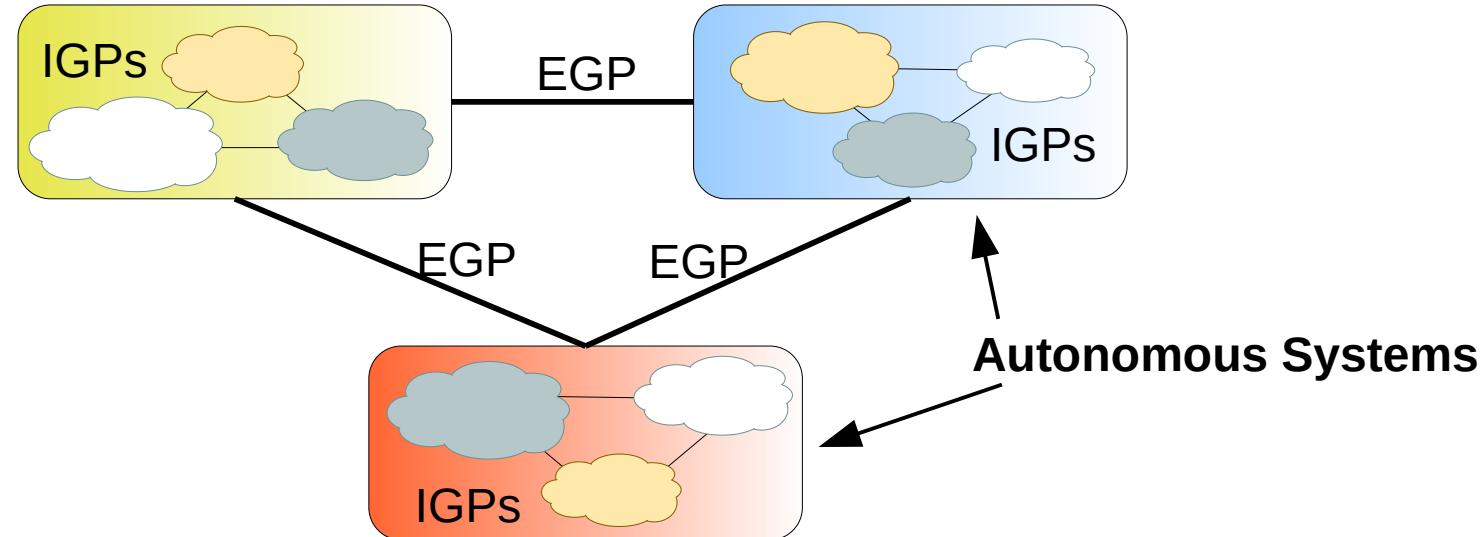
**Licenciatura em
Engenharia de Computadores e Informática
DETI-UA**



universidade de aveiro

deti.ua.pt

Border Gateway Protocol (BGP)



- Border Gateway Protocol Version 4 of the protocol (BGP4) was deployed in 1993 and currently is the protocol that assures Internet connectivity
- BGP is mainly used for routing between Autonomous Systems
- Autonomous System (AS) is a network under a single administration
 - ◆ One or more network operators with a common well defined global routing policy



AS Numbers

- Allocated ID by InterNIC and is globally unique
- RFC 4271 defines an AS number as 2-bytes
 - Private AS Numbers = 64512 through 65535
 - Public AS Numbers = 1 through 64511
 - 39000+ have already been allocated
 - We will eventually run out of AS numbers
- Need to expand AS size from 2-bytes to 4-bytes
- RFC4893 defines BGP support for 4-bytes AS numbers
 - 4,294,967,295 AS numbers
 - As of January 1, 2009, all new Autonomous System numbers issued will be 4-byte by default, unless otherwise requested.
 - The full binary 4-byte AS number is split two words of 16 bits each
 - Notation:
 - <higher2bytes in decimal>.<lower2bytes in decimal>
 - Example1: AS 65546 is represented as “1.10”
 - Example2: AS 50000 is represented as “0.50000”
 - Cannot have a “flag day” solution



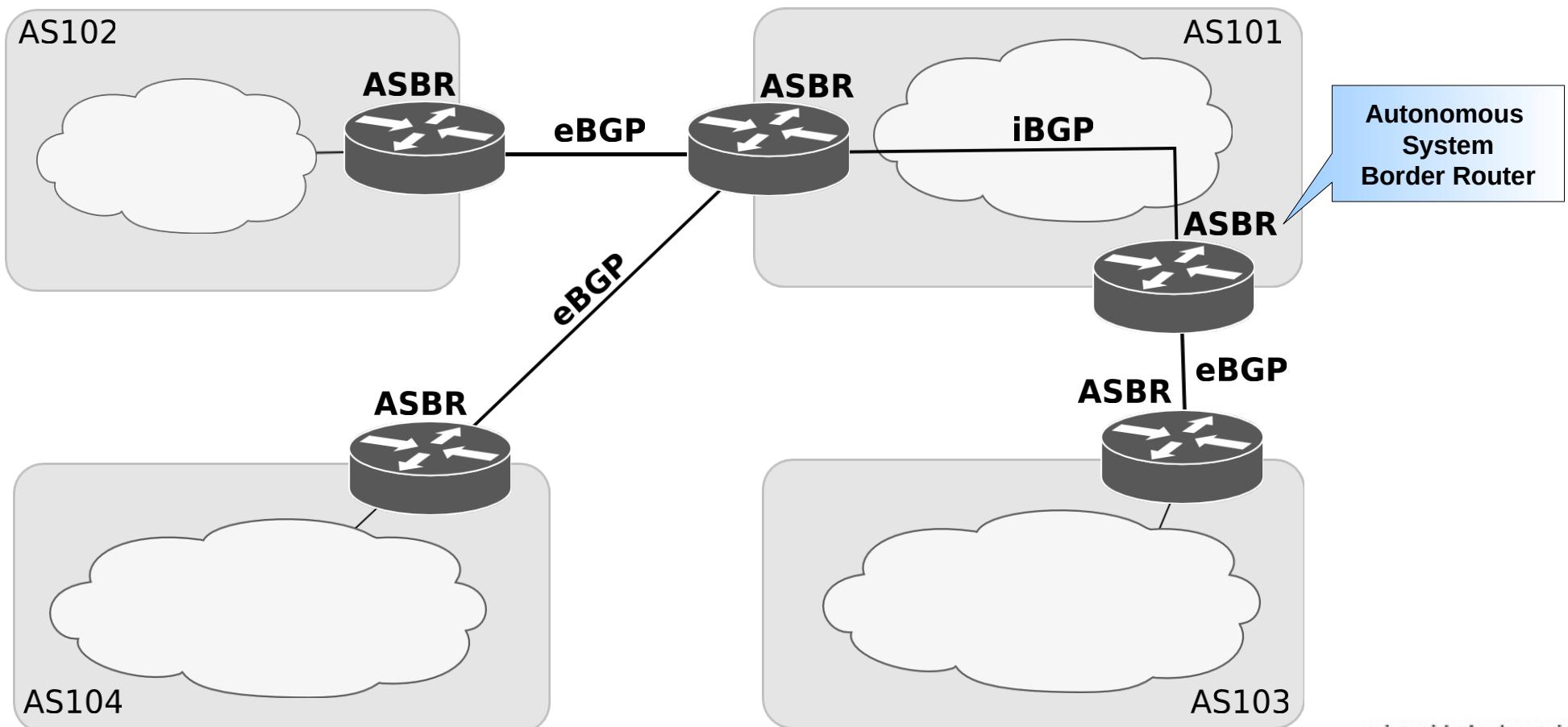
BGP Neighbor Relationships

- Often called peering
 - ◆ Usually manually configured into routers by the administrator
- Each neighbor session runs over TCP (port 179)
 - ◆ Ensures reliable data delivery
- Peers exchange all their routes when the session is first established
- Updates are also sent when there is a topology change in the network or a change in routing policy
- BGP peers exchange session KEEPALIVE messages
 - ◆ To avoid extended periods of inactivity.
 - ◆ Low keepalive intervals can be set if a fast fail-over is required



Internal BGP (iBGP) & External BGP (eBGP)

- Neighbor relations can be established between
 - ◆ Same AS routers (Internal BGP – iBGP).
 - ◆ Different AS routers (External BGP – eBGP).
- Routers that implement neighbor relations are called an Autonomous System Border Router (ASBR).



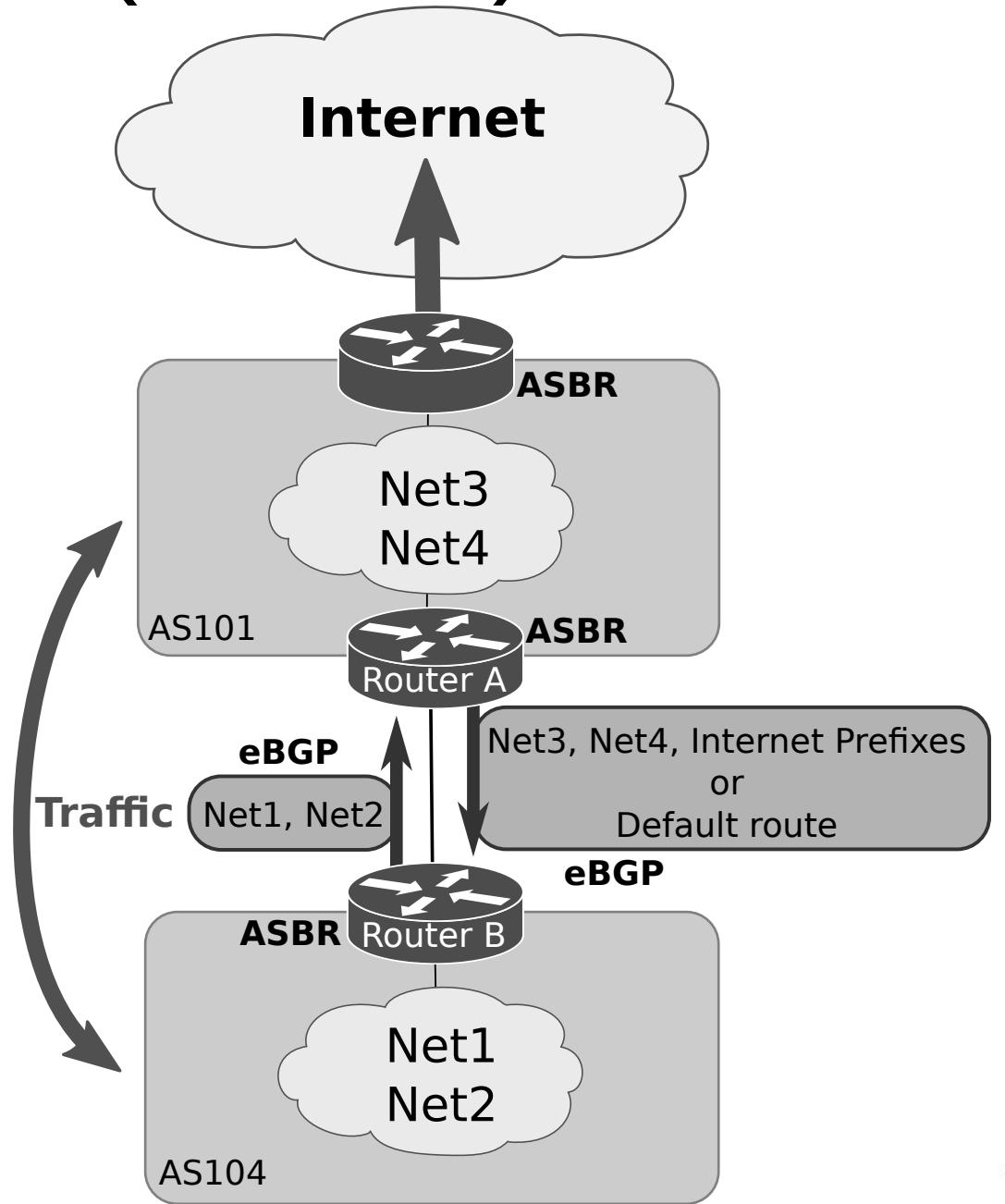
External and Internal BGP

- External BGP (eBGP) is used between AS.
- Internal BGP (iBGP) is used within AS.
- A BGP router never forwards a path learned from one iBGP peer to another iBGP peer even if that path is the best path.
 - ◆ An exception is when a router is configured as route-reflector.
- A BGP forward the routes learned from one eBGP peer to both eBGP and iBGP peers.
 - ◆ Filters can be used to modify this behavior.
- iBGP routers in an AS **must maintain an iBGP session with all other iBGP routers** in the AS (iBGP Mesh).
 - ◆ To obtain complete routing information about external networks.
 - ◆ Most networks also use an IGP, such as OSPF.
 - ◆ Additional methods can be used to reduce iBGP Mesh complexity.
 - ◆ Route reflectors, private AS, ...



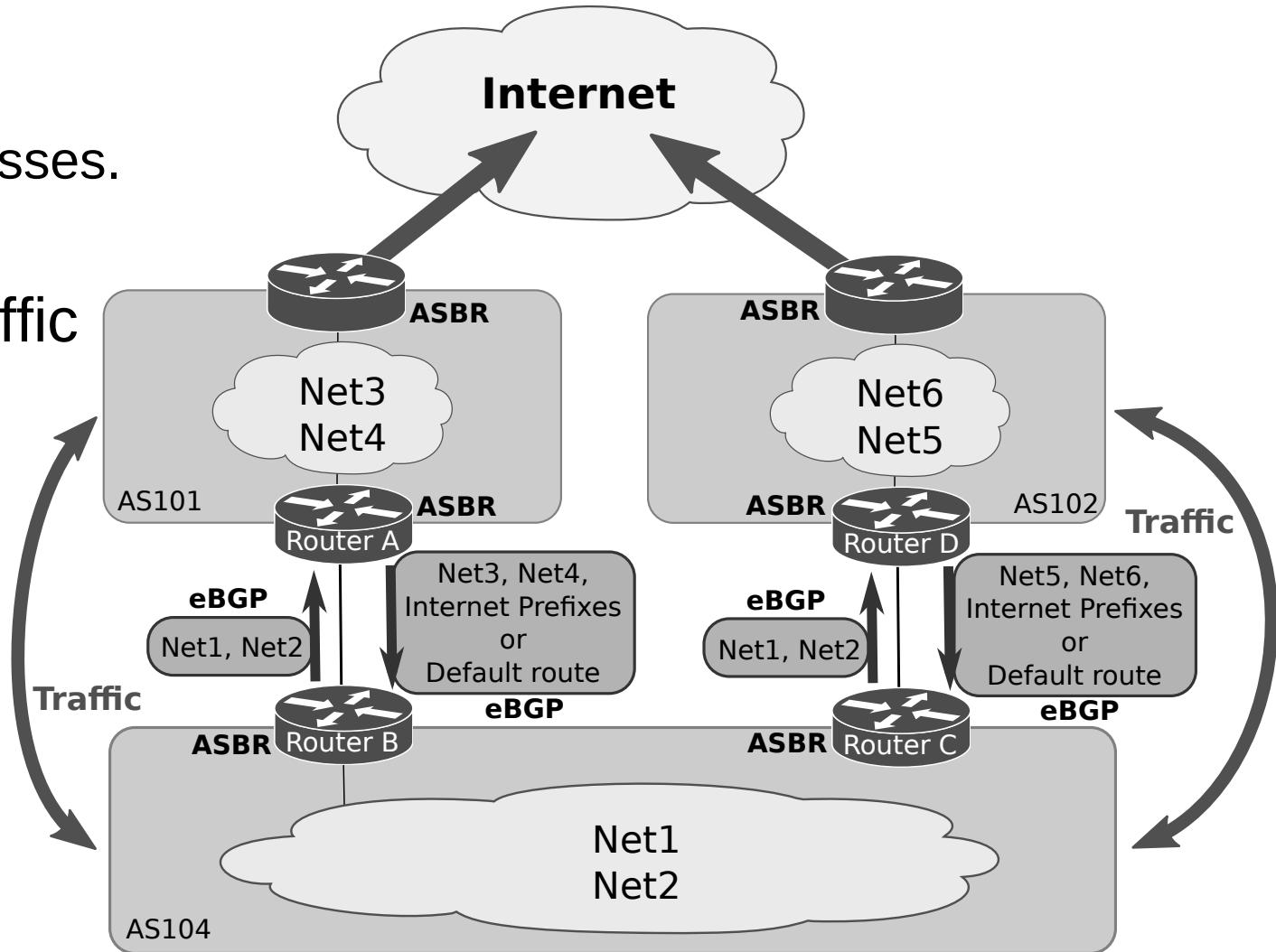
Single-homed (or Stub) AS

- AS has only one border router (ASBR)
 - ◆ Single Internet access.
 - ◆ Single ISP.



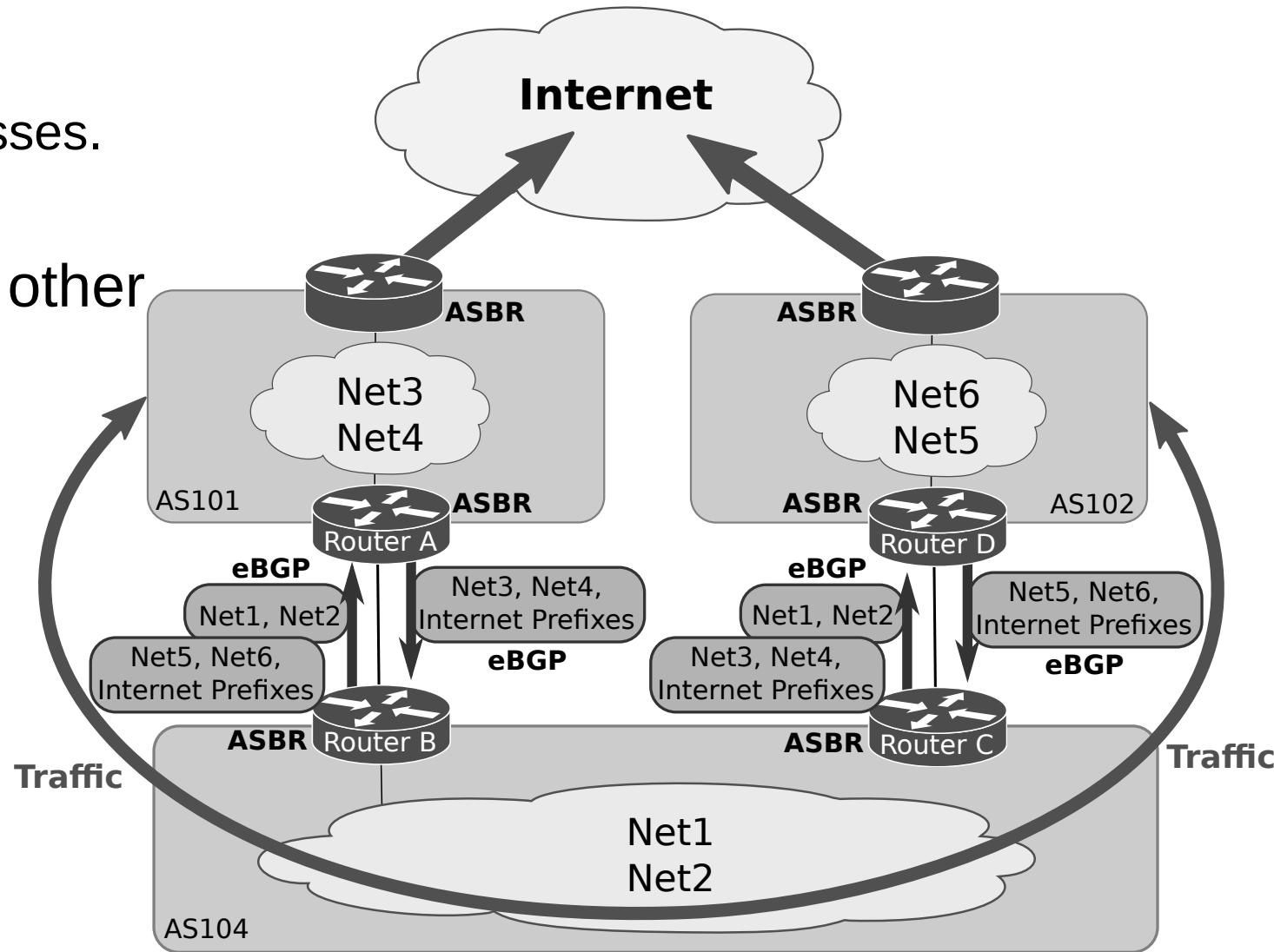
Multi-homed Non-transit AS

- AS has more than one border router (ASBR)
 - ◆ Multiple Internet accesses.
 - ◆ Multiple ISP.
- Does not transport traffic from other AS.



Multi-homed Transit AS

- AS has more than one border router (ASBR).
 - ◆ Multiple Internet accesses.
 - ◆ Multiple ISP.
- Transports traffic from other AS.

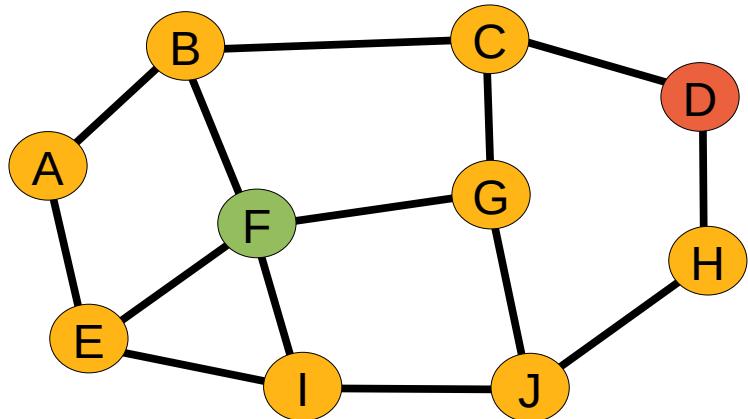


Path-vector

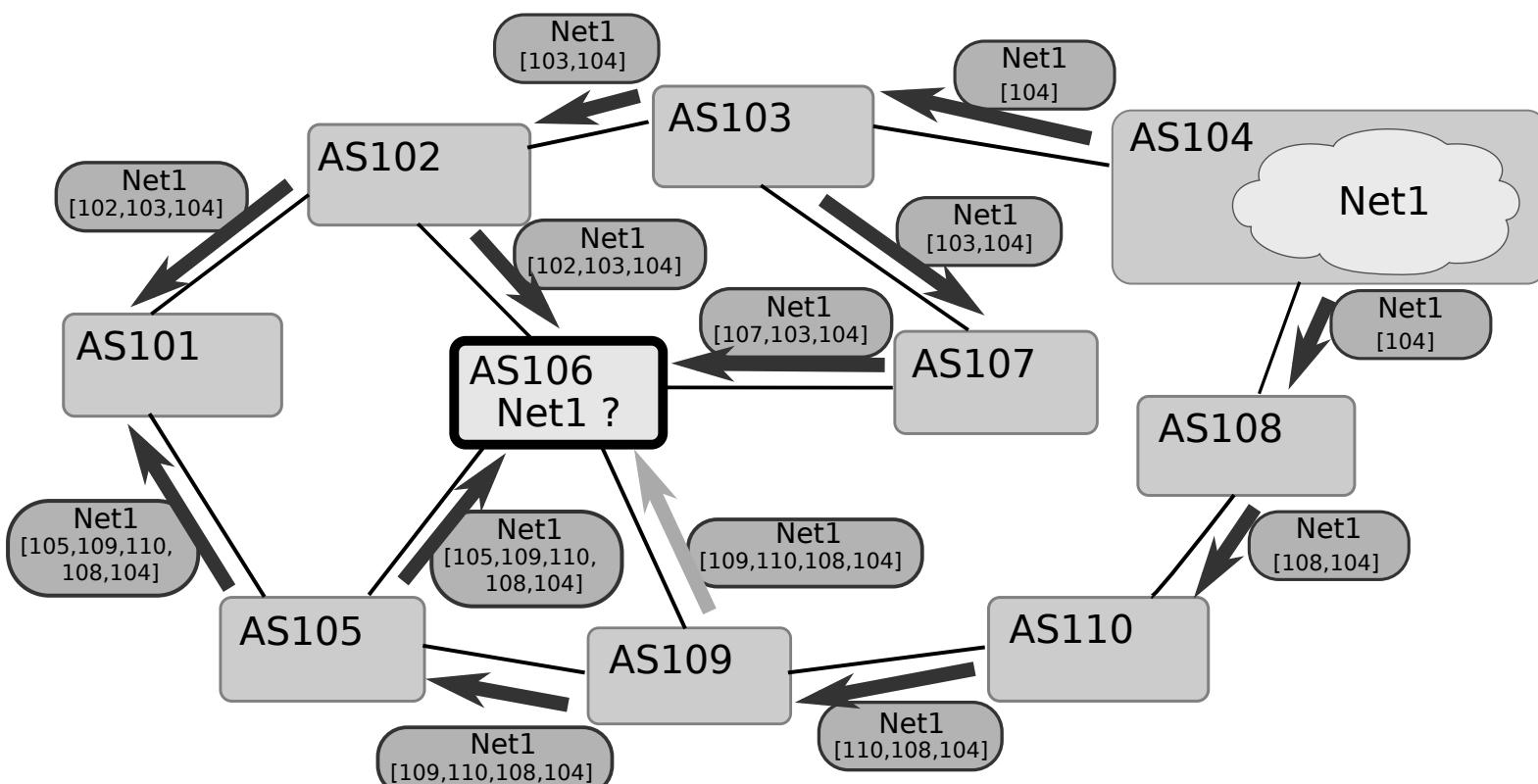
- BGP is a path-vector protocol
- Although it is essentially a distance-vector protocol that carries a list of the AS traversed by the route
 - ◆ Provides loop detection
- An EBGP speaker adds its own AS to this list before forwarding a route to another EBGP peer
- An IBGP speaker does not modify the list because it is sending the route to a peer within the same AS
 - ◆ AS list cannot be used to detect the IBGP routing loops



Path vector

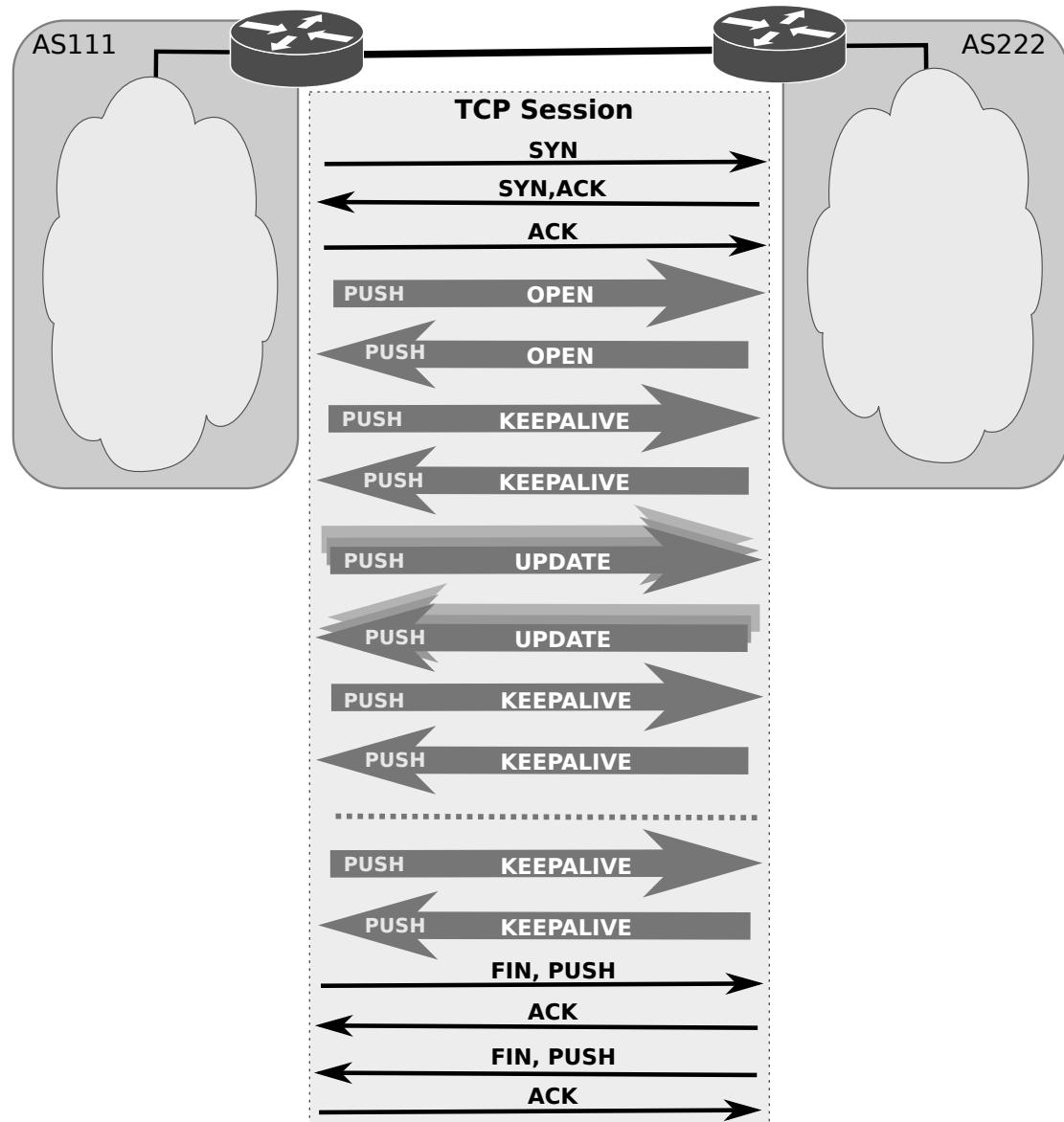


- F receives from its neighbors different paths to D:
 - ◆ De B: "I use BCD"
 - ◆ De G: "I use GCD"
 - ◆ De I: "I use IFGCD"
 - ◆ De E: "I use EFGCD"



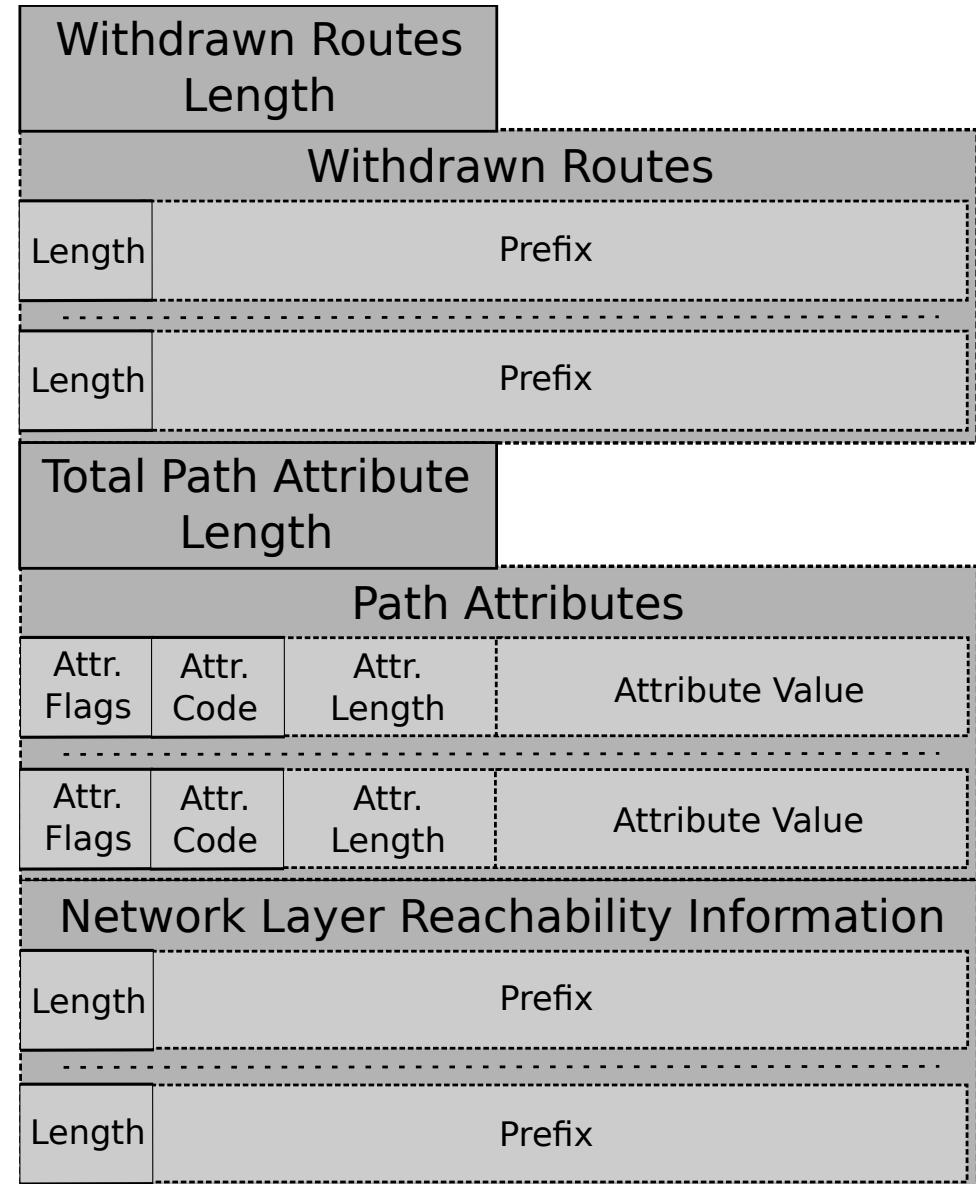
BGP Messages

- OPEN messages are used to establish the BGP session.
- UPDATE messages are used to send routing prefixes, along with their associated BGP attributes (such as the AS-PATH).
- KEEPALIVE messages are exchanged whenever the keepalive period is exceeded, without an update being exchanged.
- NOTIFICATION messages are sent whenever a protocol error is detected, after which the BGP session is closed.

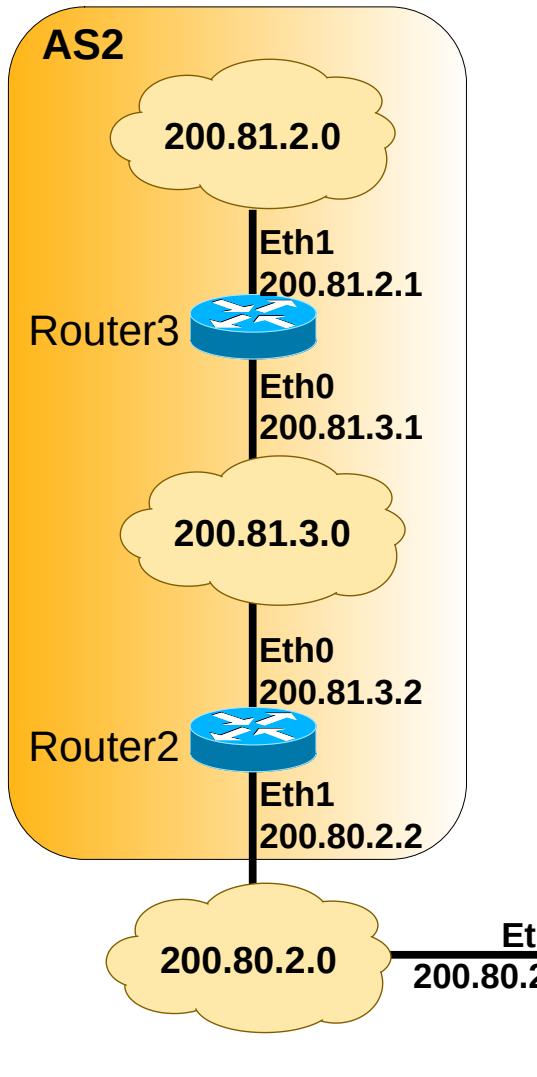


Update Message

- Withdrawn routes – List of IP networks no longer accessible.
- Path attributes – parameters used to define routing and routing policies.
- Network layer reachability information – List of IP networks with connectivity.



Example



C 200.81.3.0/24 is directly connected, Ethernet0
O 200.81.2.0/24 [110/20] via 200.81.3.1, 00:01:12
C 200.80.2.0/24 is directly connected, Ethernet1
B 200.80.1.0/24 [20/0] via 200.80.2.1, 00:00:29

Router 2's routing table

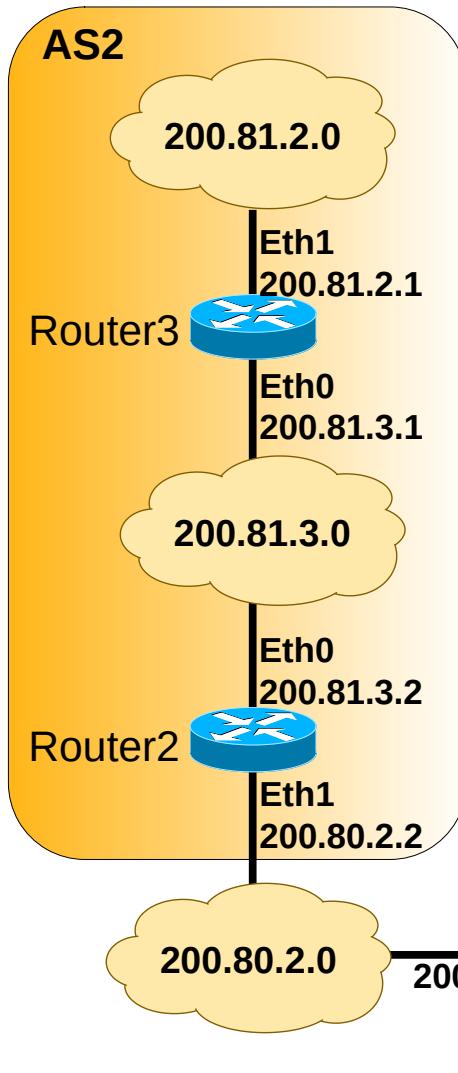
B 200.81.3.0/24 [20/0] via 200.80.2.2, 00:01:58
B 200.81.2.0/24 [20/0] via 200.80.2.2, 00:01:57
C 200.80.2.0/24 is directly connected, Ethernet1
C 200.80.1.0/24 is directly connected, Ethernet0

Router 1's routing table



Example – BGP networks aggregation

Before aggregation



B 200.81.3.0/24 [20/0] via 200.80.2.2, 00:01:58

B 200.81.2.0/24 [20/0] via 200.80.2.2, 00:01:57

C 200.80.2.0/24 is directly connected, Ethernet1

C 200.80.1.0/24 is directly connected, Ethernet0

Router 1

After aggregation

B 200.81.2.0/23 [20/0] via 200.80.2.2, 00:01:06

C 200.80.2.0/24 is directly connected, Ethernet1

C 200.80.1.0/24 is directly connected, Ethernet0

Router 1



BGP Attributes

- A BGP attribute, or path attribute, is a metric used to describe the characteristics of a BGP path.
- Attributes are contained in update messages passed between BGP peers to advertise routes. There are 4+1 categories of BGP attributes.
 - ◆ Well-known Mandatory (included in BGP updates)
 - AS-path, Next-hop, Origin.
 - ◆ Well-known Discretionary (may or may not be included in BGP updates)
 - Local Preference, Atomic Aggregate.
 - ◆ Optional Transitive (may not be supported by all BGP implementations)
 - Aggregator, Community, AS4_Aggregator, AS4_path.
 - ◆ Optional Non-transitive (may not be supported by all BGP implementations)
 - If the neighbor doesn't support that attribute it is deleted
 - Multi-exit-discriminator (MED).
 - ◆ Cisco-defined (local to router, not advertised)
 - Weight



AS-PATH and ORIGIN Attributes

- AS-PATH
 - ◆ When a route advertisement passes through an autonomous system, the AS number is added to an ordered list of AS numbers that the route advertisement has traversed.
- ORIGIN
 - ◆ Indicates how BGP learned about a particular route. Can take three possible values:
 - ◆ IGP (0) value is set if the route is interior to the originating AS, resulting from an explicit inclusion of a network within the BGP routing process by means of manual configuration.
 - ◆ INCOMPLETE (2) value is set if the route is learned by other means, namely, route redistribution from other routing processes into the BGP routing process.
 - ◆ EGP (1) is no longer used in modern networks.

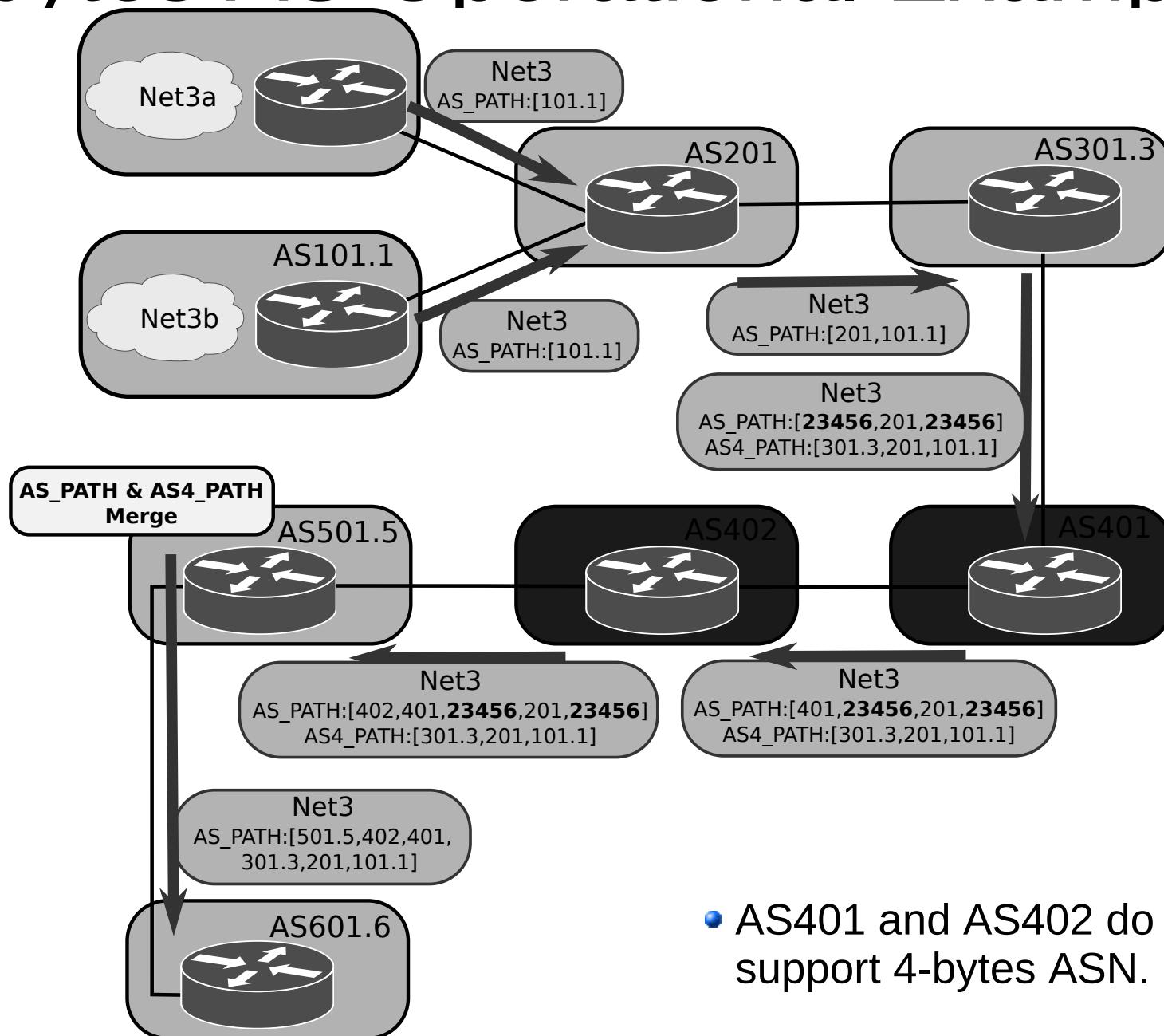


AS4_PATH & AS4_AGGREGATOR

- AS4_PATH attribute has the same semantics as the AS_PATH attribute, except that it is optional transitive, and it carries 4-bytes AS numbers.
- AS4_AGGREGATOR attribute has the same semantics as the AGGREGATOR attribute, except that it carries a 4-bytes AS number.
- 4-byte AS support is advertised via BGP capability negotiation
 - ◆ Speakers who support 4-byte AS are known as NEW BGP speakers
 - ◆ Those who do not are known as OLD BGP speakers
- New Reserved AS number
 - ◆ AS_TRANS = AS 23456
 - ◆ 2-byte placeholder for a 4-byte AS number
 - ◆ Used for backward compatibility between OLD and NEW BGP speakers
- Receiving UPDATEs from a NEW speaker
 - ◆ Decode each AS number as 4-bytes
 - ◆ AS_PATH and AGGREGATOR are effected
- Receiving UPDATEs from an OLD speaker
 - ◆ AS4_AGGREGATOR will override AGGREGATOR
 - ◆ AS4_PATH and AS_PATH must be merged to form the correct as-path
- Merging AS4_PATH and AS_PATH
 - ◆ AS_PATH → [275 250 225 23456 23456 200 23456 175]
 - ◆ AS4_PATH → [100.1 100.2 200 100.3 175]
 - ◆ Merged AS-PATH → [275 250 225 100.1 100.2 200 100.3 175]

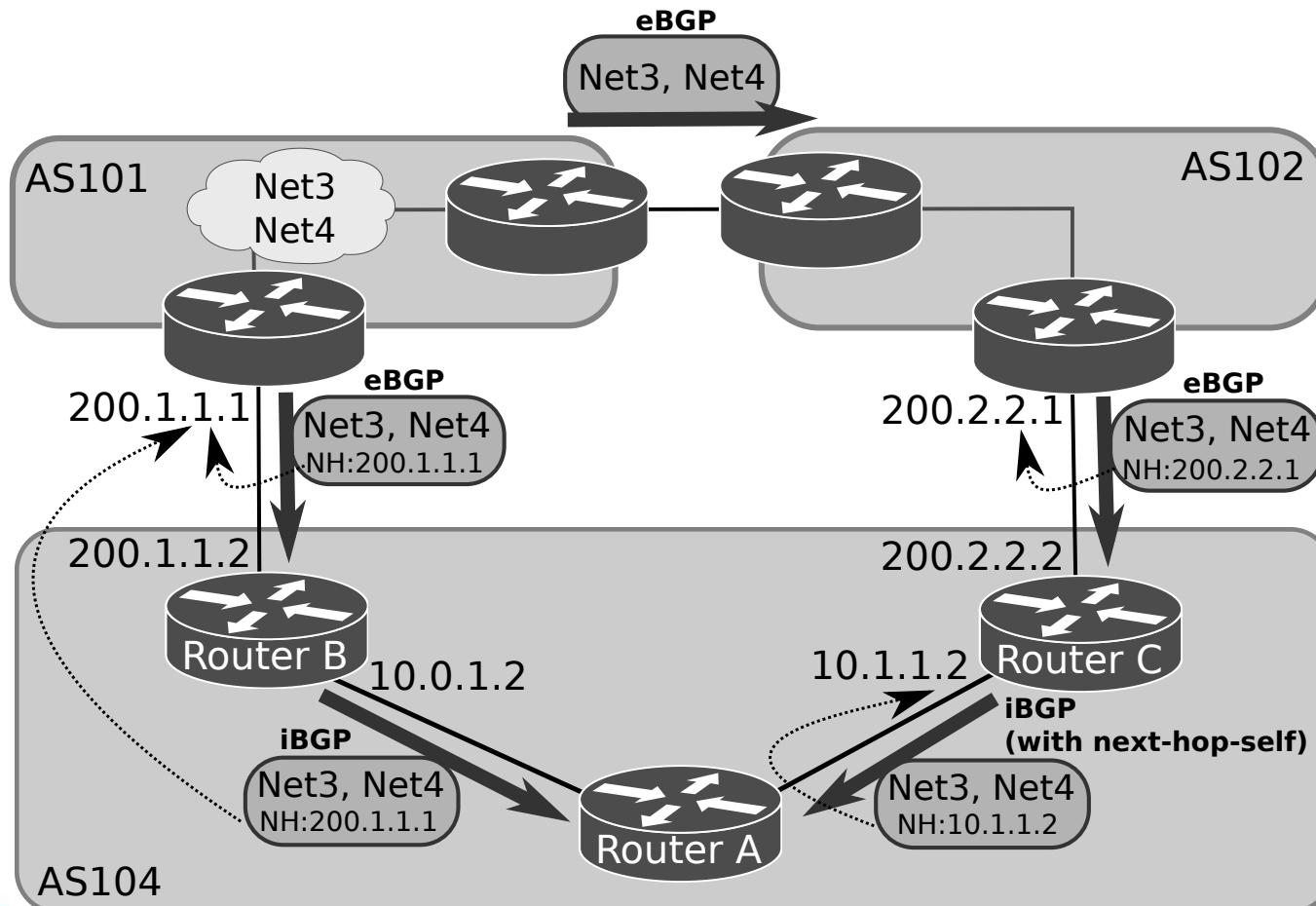


4-bytes AS Operational Example



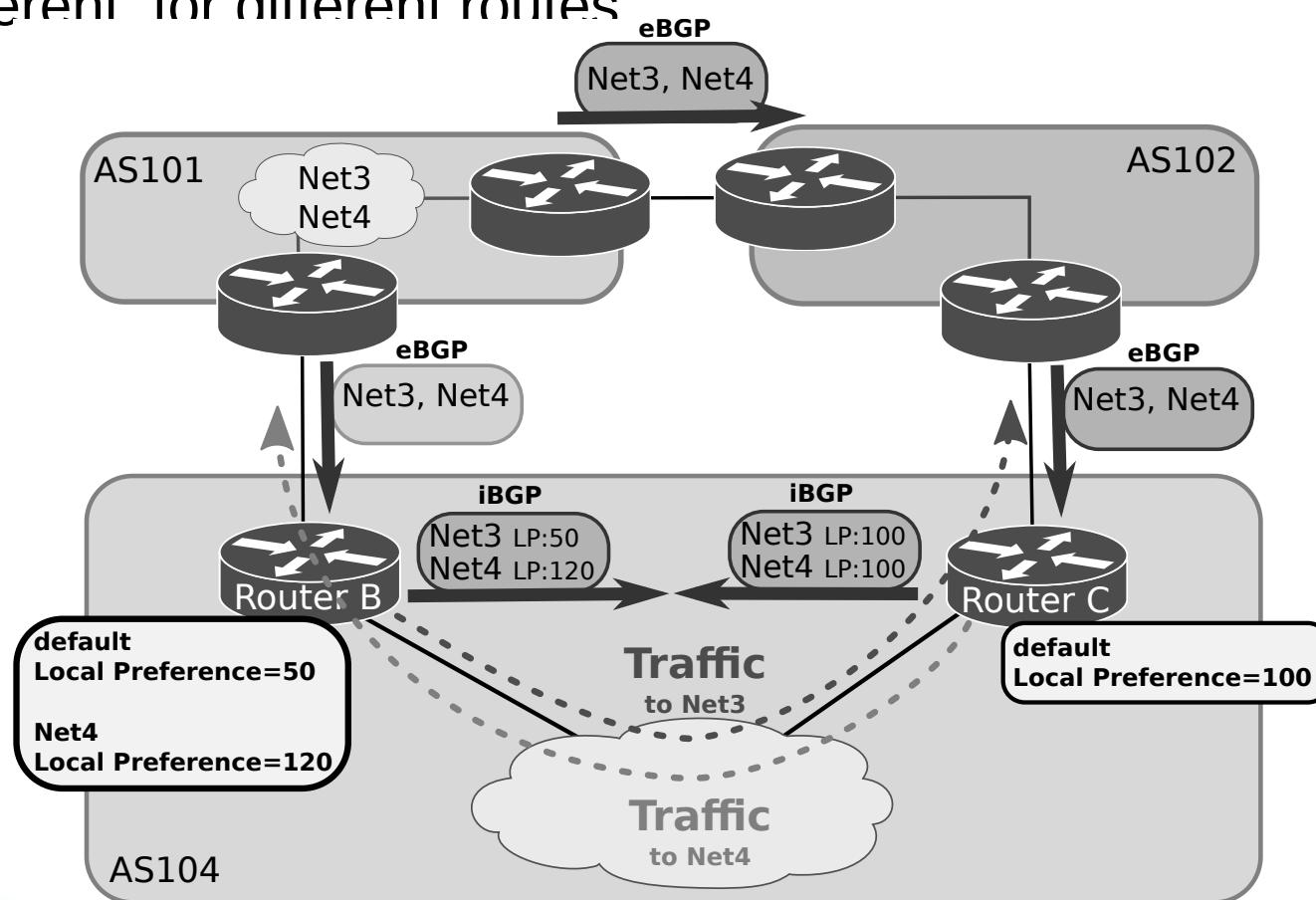
Next-Hop Attribute

- The eBGP next-hop attribute is the IP address that is used to reach the advertising router
- For eBGP, the next-hop address is the IP address of the connection between the peers
- For iBGP, the eBGP next-hop address is carried into the local AS
 - ◆ By configuration the AS border router can be the next-hop to iBGP neighbors



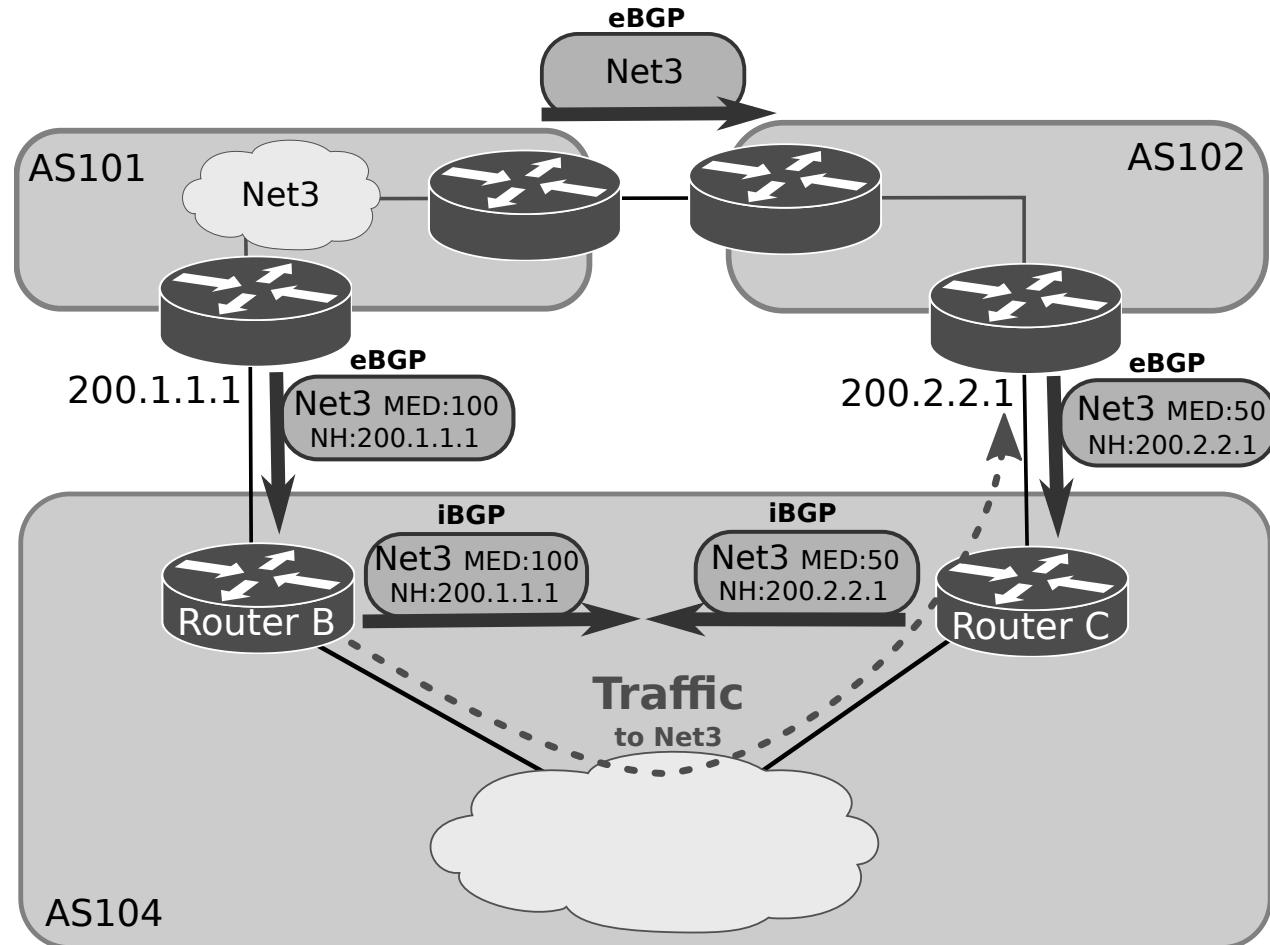
Local Preference Attribute

- The local preference attribute is used to choose an exit point from the local autonomous system (AS).
 - ◆ **Higher value** is preferred.
- The local preference attribute is propagated throughout the local AS.
- Can be different for different routes



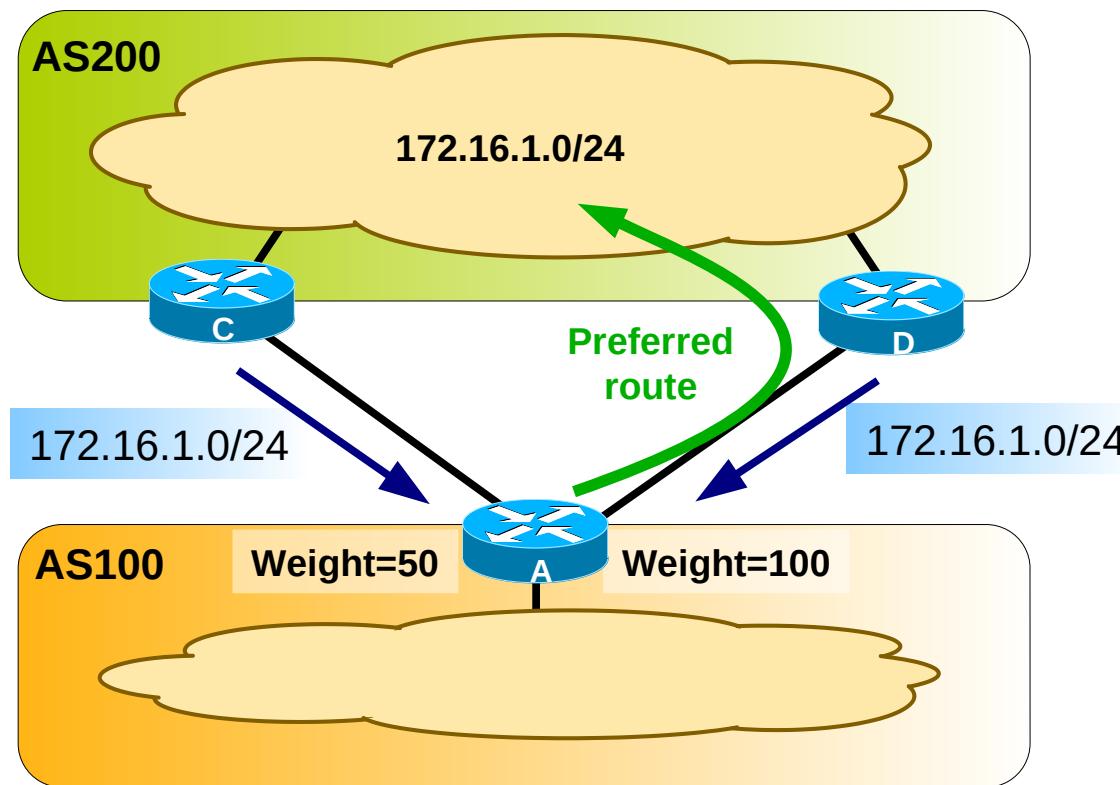
Multi-Exit Discriminator Attribute (MED)

- The multi-exit discriminator (MED) or metric attribute is used as a suggestion to an external AS.
- The external AS that is receiving the MEDs may be using other BGP attributes for route selection.
- The **lower value** of the metric is preferred.
- MED is designed to influence incoming traffic.

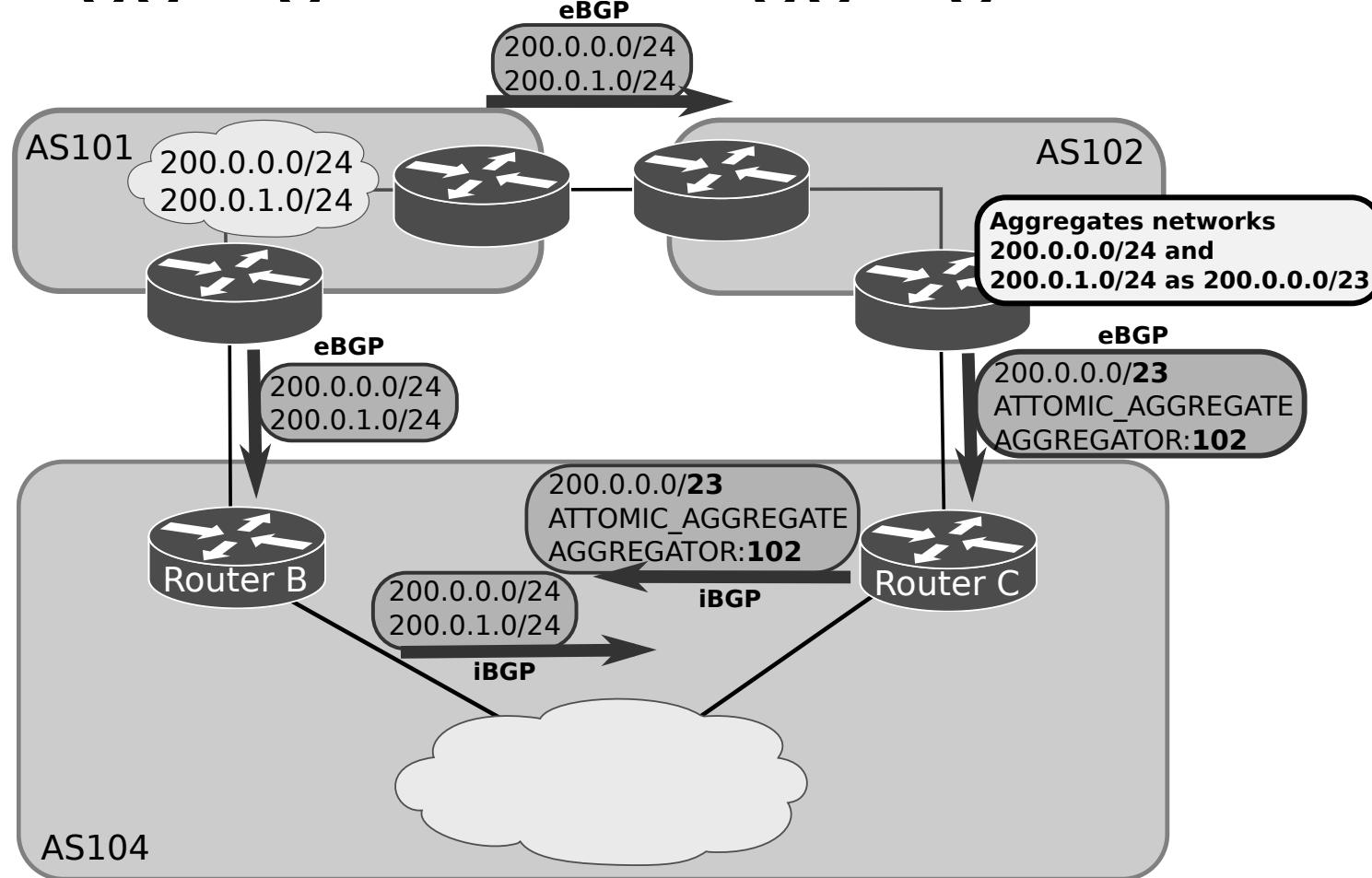


Weight Attribute

- Weight is a Cisco-defined attribute that is local to a router.
- The weight attribute is not advertised to neighboring routers.
- If the router learns about more than one route to the same destination, the route with the **highest weight** will be preferred.



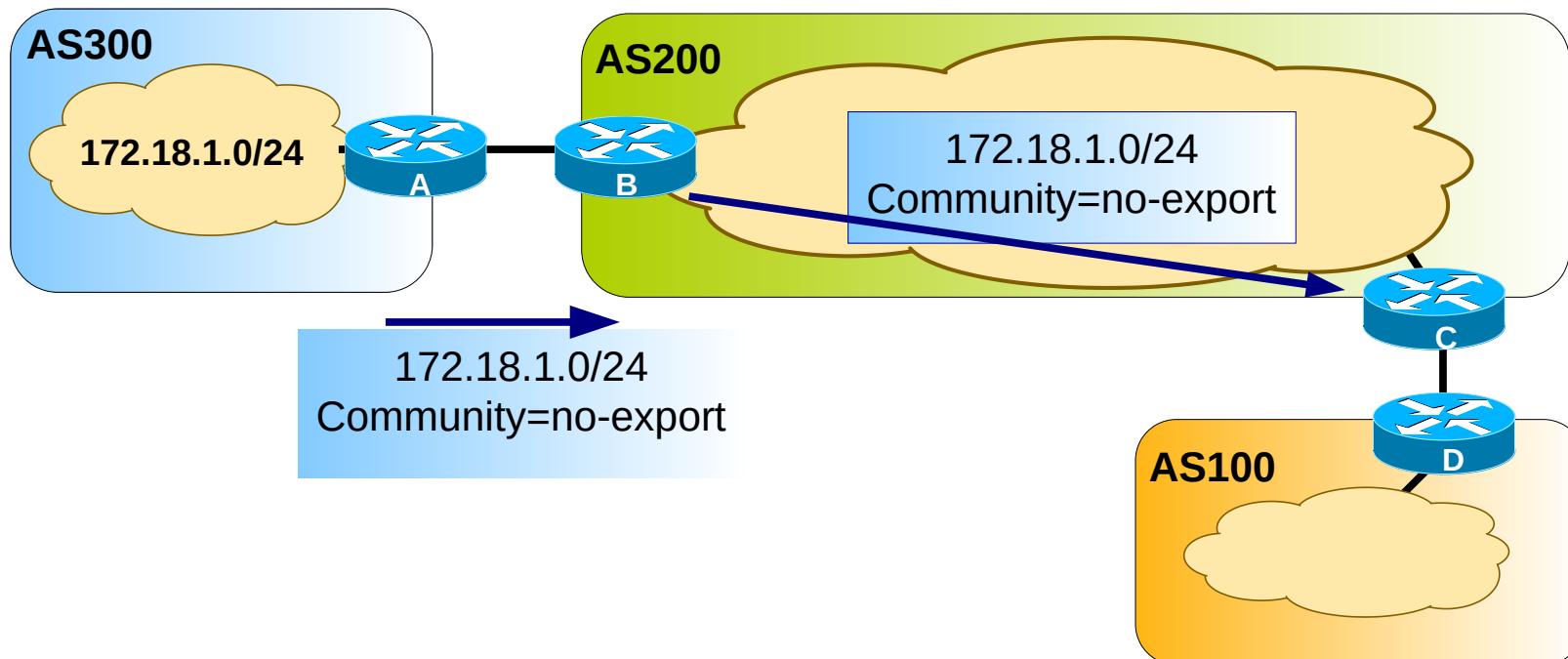
Atomic Aggregate and Aggregator Attributes



- Atomic Aggregate
 - ◆ Is used to alert routers that specific routes have been aggregated into a less specific route.
 - ◆ When aggregation like this occurs, more specific routes are lost.
- Aggregator
 - ◆ Provides information about which AS performed the aggregation.
 - ◆ And the IP address of the router that originated the aggregate.



Community Attribute



- Used to group routes that share common properties so that policies can be applied at the group level
- Predefined community attributes are:
 - no-export - Do not advertise this route to EBGP peers
 - no-advertise - Do not advertise this route to any peer
 - internet - Advertise this route to the Internet community; all routers in the network belong to it
- General communities format is ASnumber:Cnumber
 - e.g. 300:1, 200:38, etc...



BGP Path Selection

- BGP may receive multiple advertisements for the same route from multiple sources.
- BGP selects only one path as the best path.
- BGP puts the selected path in the IP routing table and propagates the path to its neighbors. BGP uses the following criteria, in the order:
 - ◆ Largest weight (Cisco only)
 - ◆ Largest local preference
 - ◆ Path that was originated locally
 - ◆ Shortest path
 - ◆ Lowest origin type (IGP lower than EGP, EGP lower than incomplete)
 - ◆ Lowest MED attribute
 - ◆ Prefer the external path over the internal path
 - ◆ Closest IGP neighbor



Multi-Protocol Border Gateway Protocol (MP-BGP)



MP-BGP Description

- Extension to the BGP protocol
- Carries routing information about other protocols/families:
 - ◆ IPv6 Unicast
 - ◆ Multicast (IPv4 and IPv6)
 - ◆ 6PE - IPv6 over IPv4 MPLS backbone
 - ◆ Multi-Protocol Label Switching (MPLS) VPN (IPv4 and IPv6)
- Exchange of Multi-Protocol Reachability Information (NLRI)



MP-BGP Attributes

- New non-transitive and optional attributes
 - ◆ MP_REACH_NLRI
 - Carry the set of reachable destinations together with the next-hop information to be used for forwarding to these destinations
 - ◆ MP_UNREACH_NLRI
 - Carry the set of unreachable destinations
- Attribute contains one or more triples
 - ◆ Address Family Information (AFI) with Sub-AFI
 - Identifies protocol information carried in the Network Layer Reachability Information
 - ◆ Next-hop information
 - Next-hop address must be of the same family
- Reachability information



MP-BGP Negotiation Capabilities

- MP-BGP routers establish BGP sessions through the OPEN message
 - ◆ OPEN message contains optional parameters
 - ◆ If OPEN parameters are not recognized, BGP session is terminated
 - ◆ A new optional parameter: CAPABILITIES
- OPEN message with CAPABILITIES containing:
 - ◆ Multi-Protocol extensions (AFI/SAFI)
 - ◆ Route Refresh
 - ◆ Outbound Route Filtering



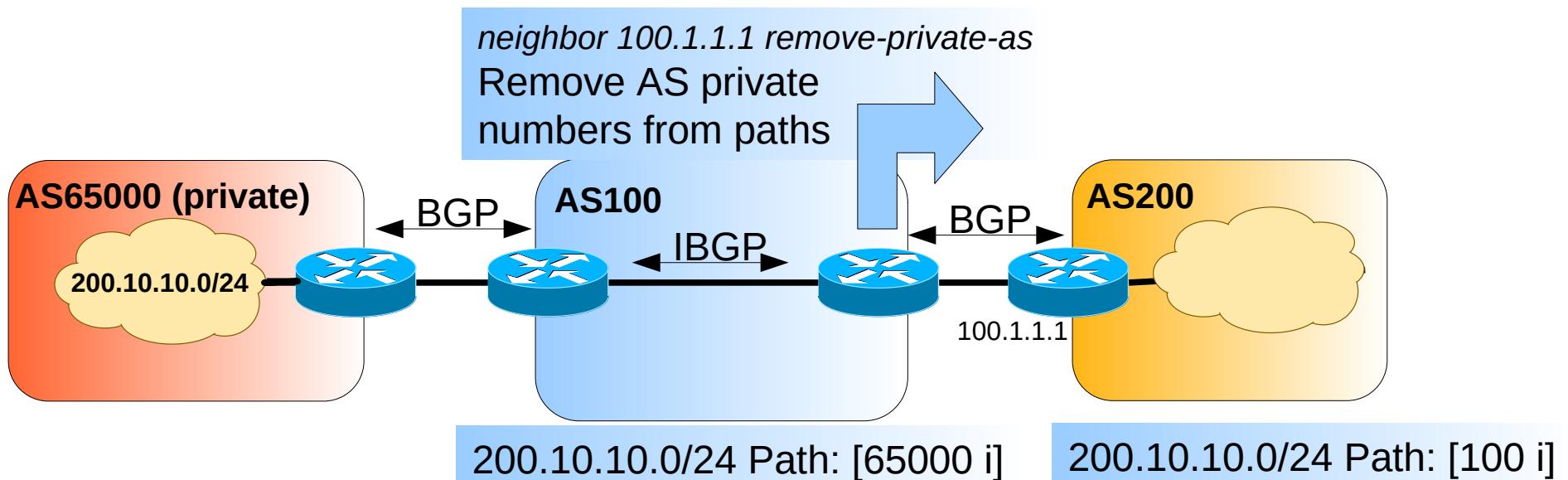
MP-BGP New Features for IPv6

- IPv6 Unicast
 - ◆ MP-BGP enables the creation of IPv6 Inter-AS relations
- IPv6 Multicast
 - ◆ Unicast prefixes for Reverse Path Forwarding (RPF) checking
 - ◆ RPF information is disseminated between autonomous systems
 - ◆ Compatible with single domain Rendezvous Points or Protocol Independent Multicast-Source Specific Multicast (PIM-SSM)
 - ◆ Topology can be congruent or non-congruent with the unicast one
- IPv6 and label (6PE)
 - ◆ IPv6 packet is transported over an IPv4 MPLS backbone
- IPv6 VPN (6VPE)
 - ◆ Multiple IPv6 VPNs are created over an IPv4 MPLS backbone



Private BGP AS

- Private autonomous system (AS) numbers range from 64512 to 65535
- When a customer network is large, the ISP may assign an AS number:
 - ◆ Permanently assigning a **Public** AS number in the range of 1 to 64511
 - ◆ Should have a unique AS number to propagate its BGP routes to Internet
 - ◆ Done when a customer network connects to two different ISPs, such as multihoming
 - ◆ Assigning a **Private** AS number in the range of 64512 to 65535.
 - ◆ It is not recommended that you use a private AS number when planning to connect to multiple ISPs in the future



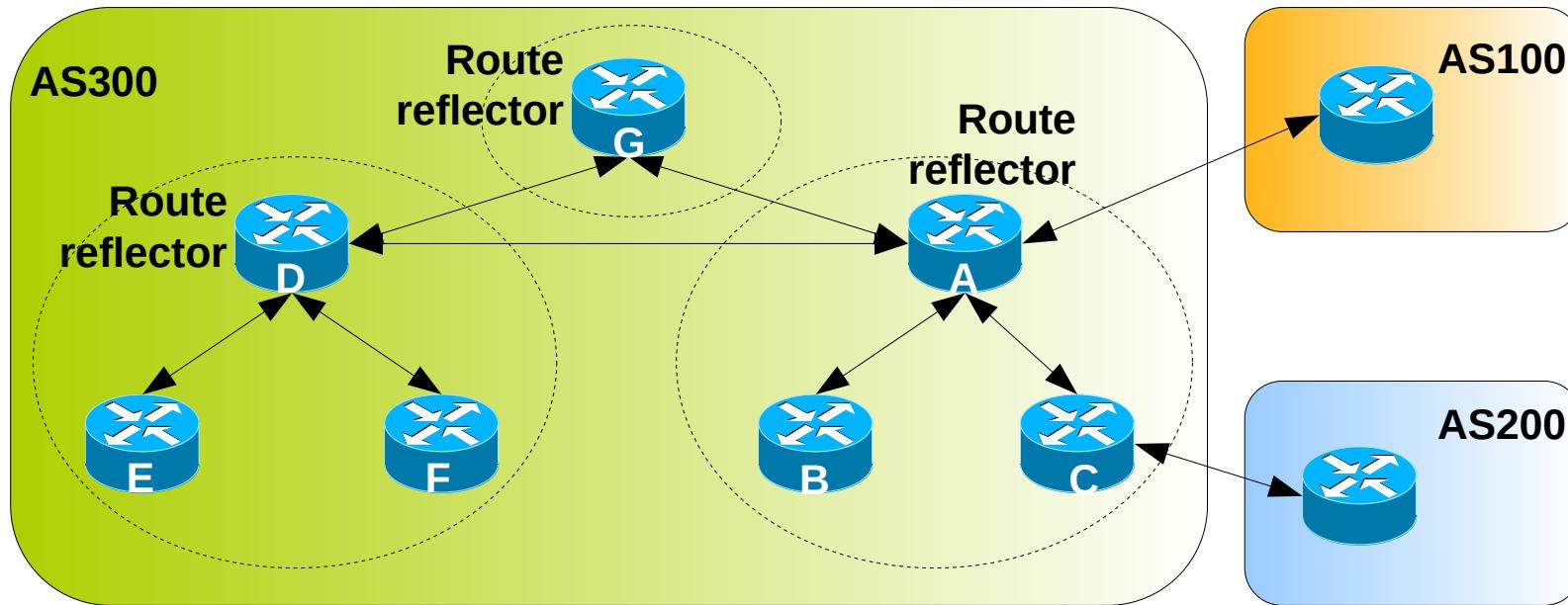
BGP AS Routing Policies

aut-num:	AS15525	
as-name:	PTPRIMENET	
descr:	PT Prime Autonomous System	export: to AS1897 announce RS-PTPRIME # KPNQwest
descr:	Corporate Data Communications Services	export: to AS1930 announce RS-PTPRIME # RCCN
descr:	Portugal	export: to AS3243 announce RS-PTPRIME # Telepac
import:	from AS1930 action pref=100; accept AS-RCCN # RCCN	export: to AS5516 announce {0.0.0.0/0} # INESC
import:	from AS3243 action pref=200; accept AS-TELEPAC # Telepac	export: to AS5533 announce RS-PTPRIME # Via NetWorks Portugal
import:	from AS5516 action pref=100; accept AS5516 # INESC	export: to AS8657 announce RS-PTPRIME # CPRM
import:	from AS5533 action pref=100; accept AS-VIAPT # Via NetWorks Portugal	export: to AS8824 announce RS-PTPRIME # Eastecnica
import:	from AS8657 action pref=300; accept ANY # CPRM	export: to AS8826 announce {0.0.0.0/0} # Siemens
import:	from AS12305 action pref=100; accept AS12305 # Nortenet	export: to AS9186 announce RS-PTPRIME # ONI
import:	from AS1897 action pref=100; accept AS1897 AS9190 AS13134 AS15931 # KPN Qwest	export: to AS12305 announce RS-PTPRIME # Nortenet
import:	from AS13156 action pref=100; accept AS13156 # Cabovisao	export: to AS12353 announce RS-PTPRIME # Vodafone Portugal
import:	from AS8824 action pref=100; accept AS8824 AS15919 # Eastecnica	export: to AS13156 announce RS-PTPRIME # Cabovisao
.....	export: to AS13910 announce ANY # register.com
		export: to AS15931 announce ANY # YASP Hiperbit
		export: to AS24698 announce RS-PTPRIME # Optimus
		export: to AS25005 announce ANY # Finibanco
		export: to AS25253 announce {0.0.0.0/0} # CGDNet
		export: to AS28672 announce ANY # BPN
		export: to AS31401 announce {0.0.0.0/0} # SICAMSERV
		export: to AS39088 announce {0.0.0.0/0} # Santander-Totta
		export: to AS41345 announce RS-PTPRIME # Visabeira
		export: to AS43064 announce RS-PTPRIME # Teixeira Duarte
		export: to AS43643 announce ANY # TAP
	

From RIPE database
<http://www.db.ripe.net>



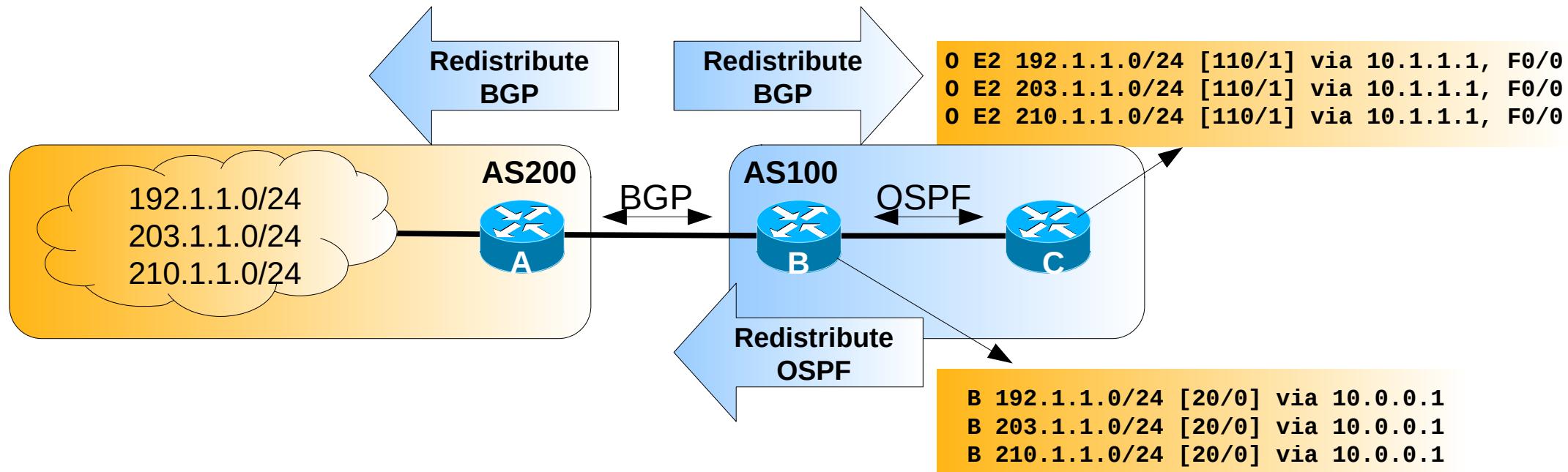
BGP Route Reflectors



- Without a route reflector, the network requires a full iBGP mesh within AS300.
- The route reflector and its clients are called a cluster.
 - Router A is configured as a route reflector, iBGP peering between Routers B and C (and others) is not required.
 - Router D is configured as a route reflector, iBGP peering between Routers E and F (and others) is not required.
- Full IBGP mesh between route reflector Routers.



Routes Redistribution



- Redistributing IGP routes by BGP will:
 - ◆ Simplify BGP configuration (advantage)
 - ◆ And BGP will announce only internal networks with connectivity (advantage)
- Redistributing BGP routes by IGP protocols will:
 - ◆ Make internal routes know all external routes (disadvantage/advantage?)
 - ◆ Increase routing tables size in internal routers (disadvantage)
 - ◆ Decrease routing time, imposes memory requirements, ...
 - ◆ Avoid the usage of internal default routes (disadvantage/advantage?)

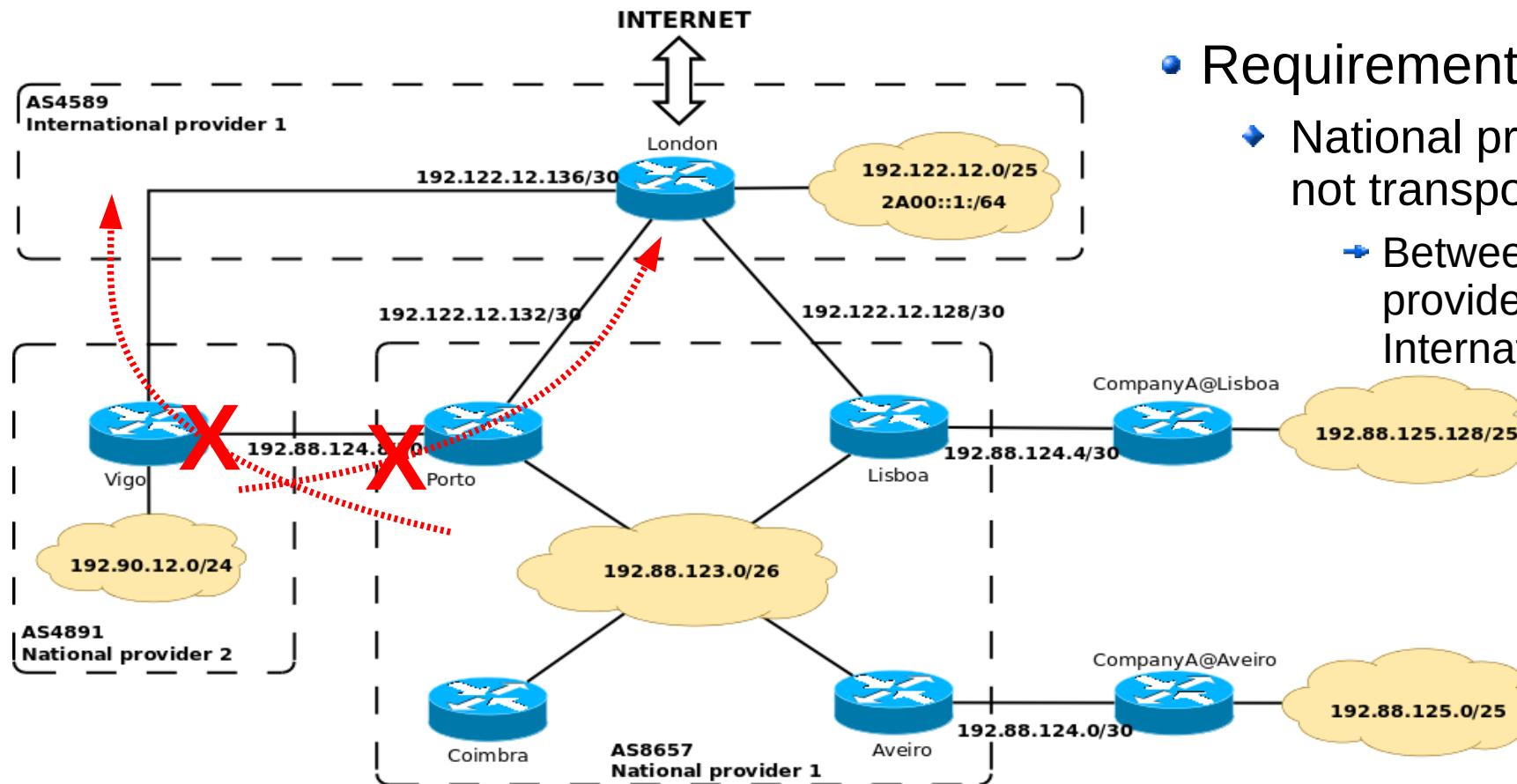


BGP Filtering

- By default BGP processes announce every network path that receives
 - ◆ From other BGP peers, or
 - ◆ Redistributed internal routing processes.
- Sending and receiving BGP updates can be controlled by using a number of different filtering methods.
 - ◆ Route-maps, prefix-lists, distribution-lists.
- BGP updates can be filtered based on:
 - ◆ Route information,
 - ◆ Path information,
 - ◆ Communities.
- Best practices:
 - ◆ Block all IPv4 private networks,
 - ◆ Announce default routes only to peers where a traffic transport contract exists.
 - ◆ Accept default routes only from peers that provide a traffic transport service.



BGP Case Studies



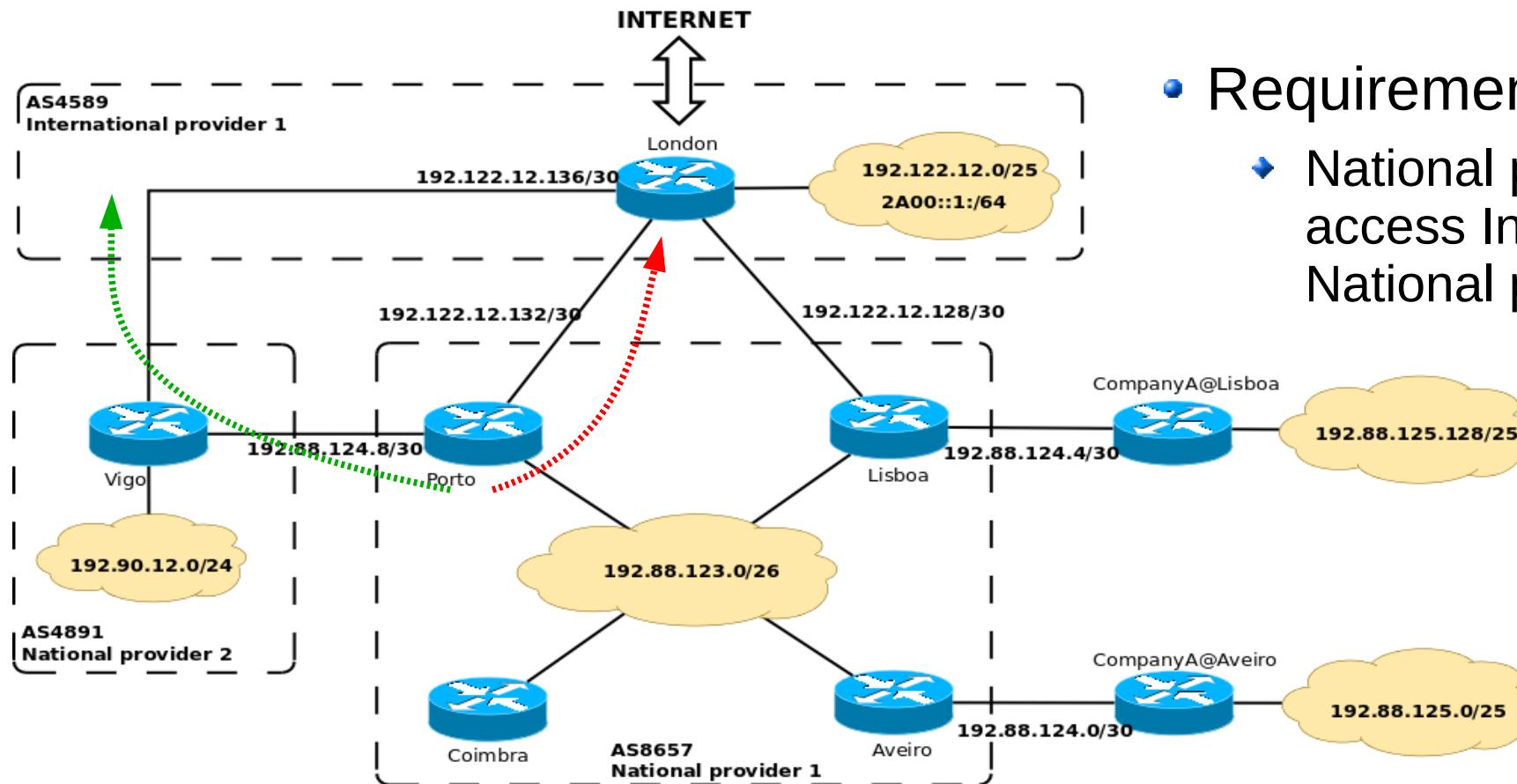
- @Porto, @Lisboa
 - ◆ Route filtering applied to all external BGP announcements
 - ◆ Announce only internal routes/nets
 - Empty path “^\$”

- Requirements
 - ◆ National providers should not transport traffic
 - ◆ Between other national providers and the International provider

- @Vigo
 - ◆ Route-map applied to all external BGP announcements
 - ◆ Announce only internal routes/nets
 - Empty path “^\$”



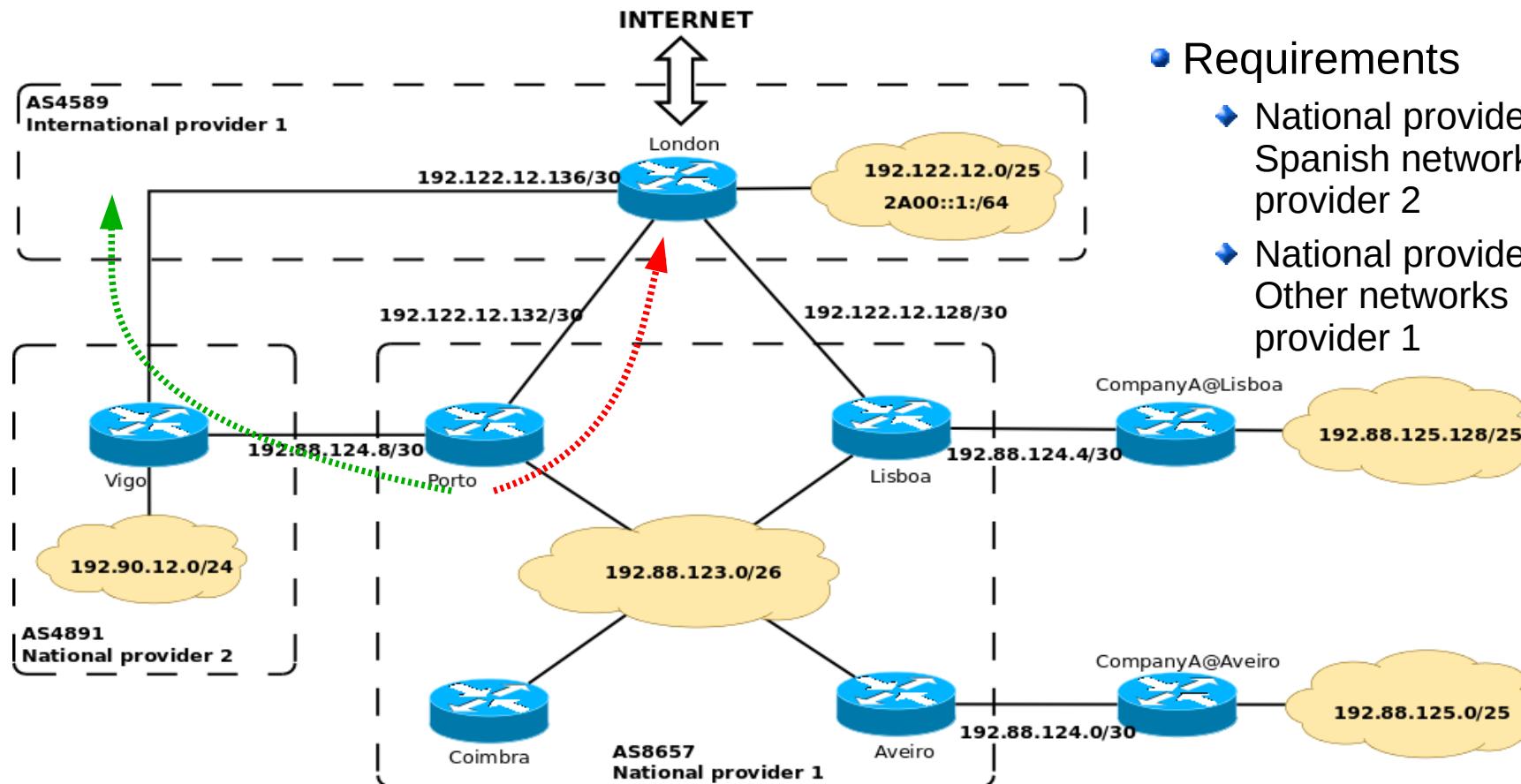
BGP Case Studies



- Requirements
 - ◆ National provider 1 should access Internet using National provider 2
- @Porto, @Lisboa
 - ◆ Route filtering applied to all BGP announcements received
 - ◆ If Path contains “4891” → **Local-preference 200**
 - ◆ If Path does not contain “4891” → **Local-preference 100**



BGP Case Studies



- Requirements

- National provider 1 should access Spanish networks using National provider 2
- National provider 1 should access Other networks using International provider 1

- @Porto, @Lisboa

- Route filtering applied to all BGP announcements received
 - E.g. known Spanish operators AS: 4891, 7654, 9876 and 3352
- If Path starts (from right to left) with "4891\$ or 7654\$ or 9876\$ or 3352\$" and ends in "^4891" → **Local-preference 200**
- If Path does not start with "4891\$ or 7654\$ or 9876\$ or 3352\$" and ends in "^4891" → **Local-preference 50**
- Assuming default Local-preference 100.



Applications Models

Redes de Comunicações II

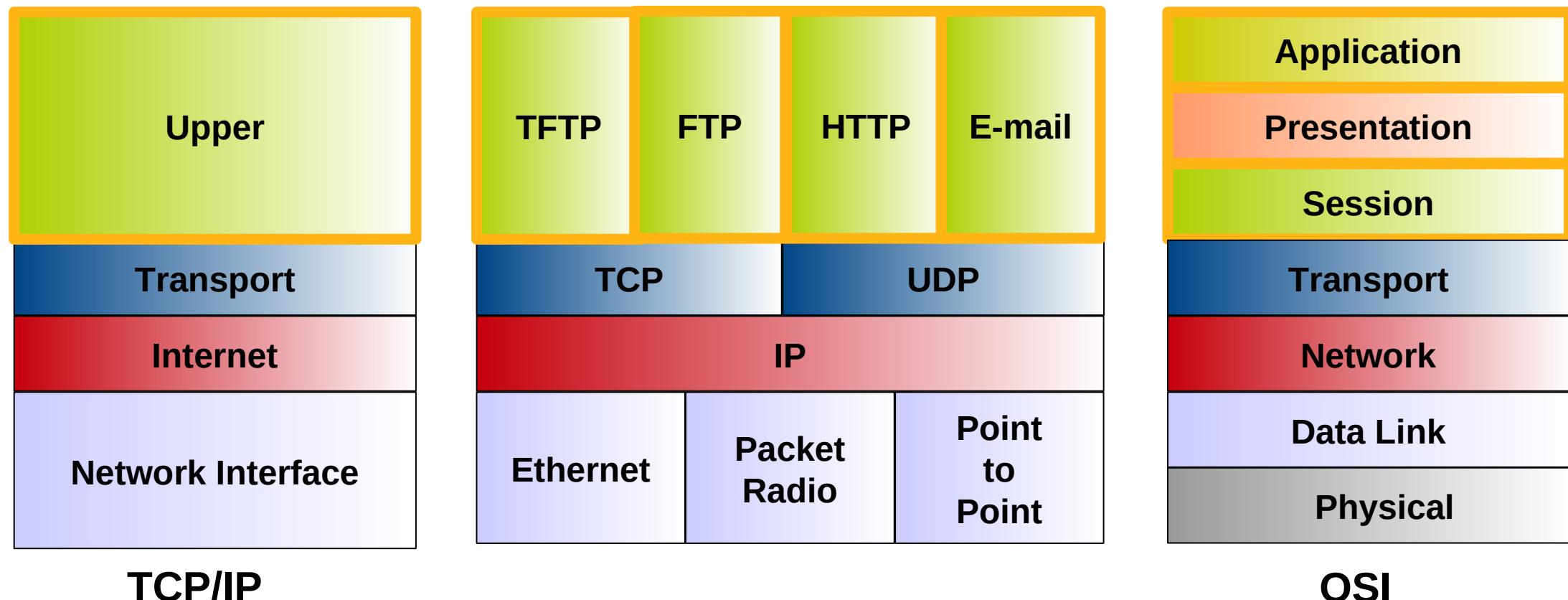
**Licenciatura em
Engenharia de Computadores e Informática
DETI-UA**



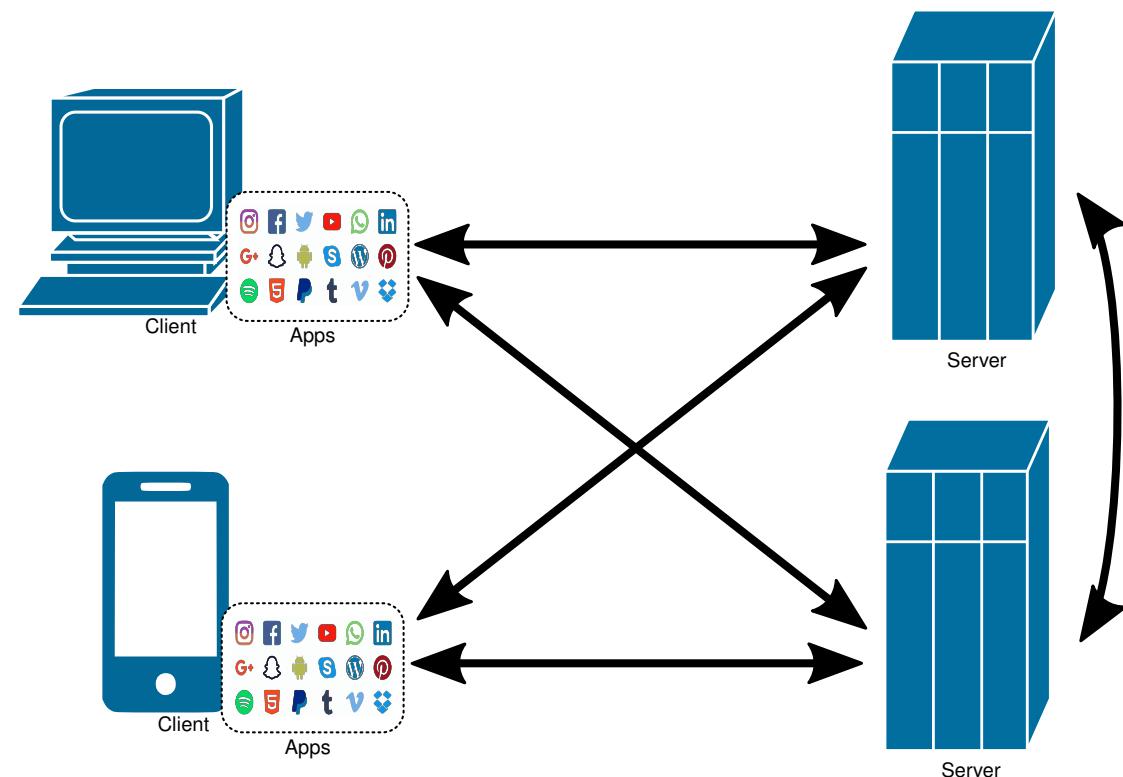
universidade de aveiro

deti.ua.pt

TCP/IP Reference Model



Client-Server Model



Servers:

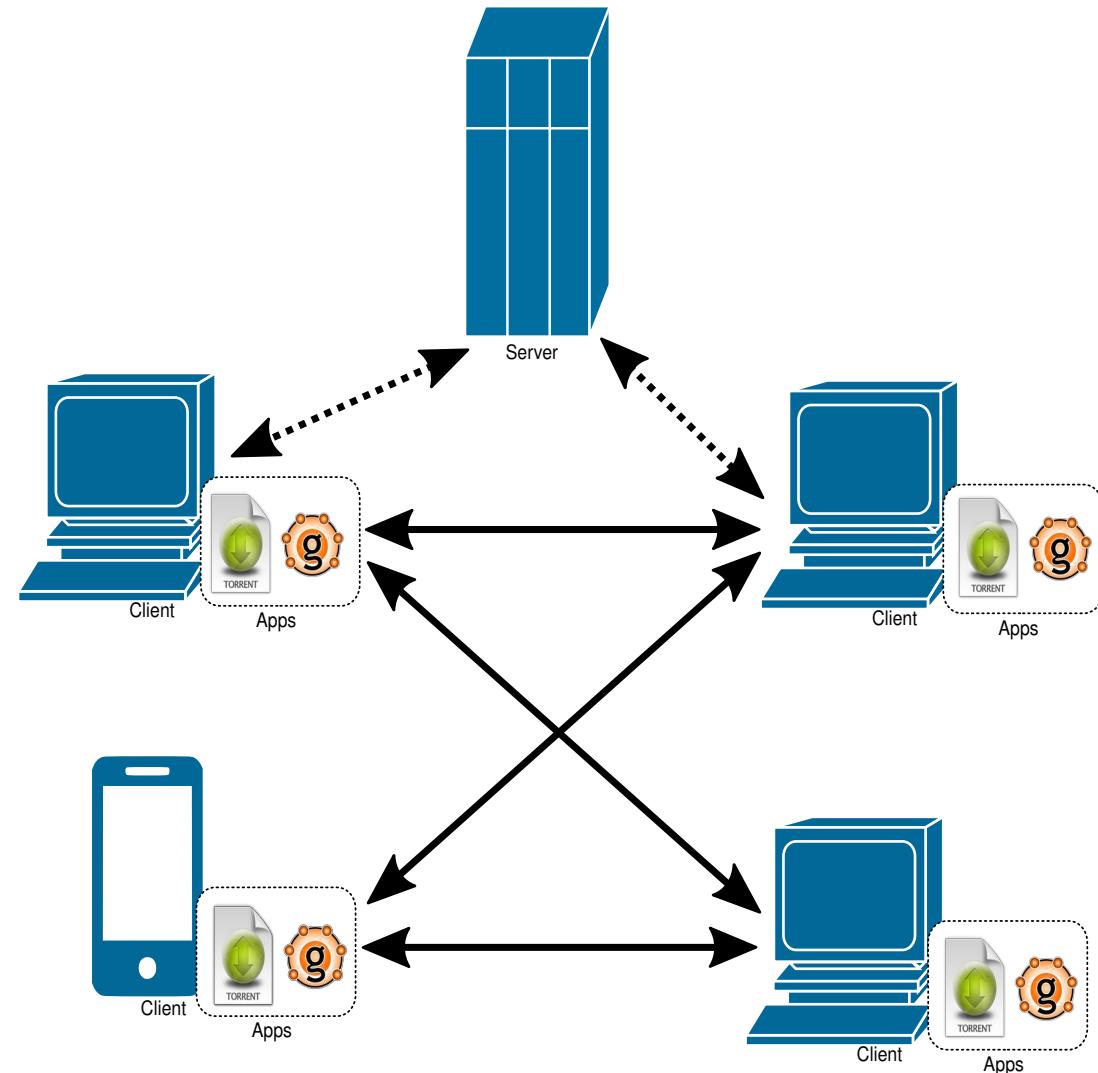
- ◆ Always ON.
- ◆ IP address is always the same or exists a static association between a name and a dynamic IP address.
- ◆ May communicate between them.
 - May act as client.

Clients:

- ◆ Communicate with servers.
- ◆ Can be ON only when in operation.
- ◆ May have dynamic addresses.
- ◆ Within this model, they do not communicate between themselves.
 - P2P is another communication model.



P2P Model



Clients:

- ◆ Communicate between themselves.
- ◆ Can be ON only when in operation.
- ◆ May have dynamic addresses.
- ◆ Peer discovery may be done within the P2P network or using central servers.

Servers:

- ◆ May exist only to bootstrap P2P network.



VoIP Voice (and Video) over IP

Voice over IP

- Network loss: IP datagram lost due to network congestion (router buffer overflow).
- Delay loss: IP datagram arrives too late for playout at receiver.
 - Delays: processing, queueing in network; end-system (sender, receiver) delays.
 - Typical maximum tolerable delay: 400 ms.
- Loss tolerance: depending on voice encoding, packet loss rates between 1% and 10% can be tolerated.
- Speaker's audio: alternating talk/speech with silent periods.
 - 64 kbps during talk/speech.
 - Packets generated only during talk/speech.
 - 20 msec chunks at 8 Kbytes/sec: 160 bytes data.
- Requires session establishment.
- VoIP protocols/frameworks:
 - Session Initiation Protocol (SIP)
 - Session Description Protocol (SDP)
 - H.323
- VoIP and PSTN interoperability in large/ISP scalable scenarios require complex control frameworks:
 - Media Gateway Controller Protocol (MGCP);
 - H.248/Megaco.



Session Initiation Protocol (SIP)

- Defined by RFC 3261.
- Designed for creating, modifying and terminating sessions between two or more participants.
 - Not limited to VoIP calls.
- Is a text-based protocol similar to HTTP.
 - Transported over UDP or TCP protocols.
 - Security at the transport and network layer provided with TLS (requires TCP) or IPSec.
- Offers an alternative to the complex H.323 protocols.
- Due to its simpler nature, the protocol is becoming more popular than the H.323 family of protocols.
- SIP is a peer-to-peer protocol. The peers in a session are called user agents (UAs):
 - User-agent client (UAC) - A client application that initiates the SIP request.
 - User-agent server (UAS) - A server application that contacts the user when a SIP request is received and that returns a response on behalf of the user.
- A SIP endpoint is capable of functioning as both UAC and UAS.



SIP Functionality

- SIP supports five facets of establishing and terminating multimedia communications:
 - User location - determination of the end system to be used for communication;
 - User availability - determination of the willingness of the called party to engage in communications;
 - User capabilities - determination of the media and media parameters to be used;
 - Session setup - "ringing", establishment of session parameters at both called and calling party;
 - Session management - including transfer and termination of sessions, modifying session parameters, and invoking services.



SIP Clients and Servers

• SIP Clients

- Phones (software based or hardware).
- Gateways
- User Agents
- A User Agent acts as a
 - Client when it initiates a request (UAC),
 - Server when it responds to a request (UAS).

• SIP Servers

- Proxy server
 - Receives SIP requests from a client and forwards them on the client's behalf.
 - Receives SIP messages and forward them to the next SIP server in the network.
 - Provides functions such as authentication, authorization, network access control, routing, reliable request retransmission, and security.
- Redirect server
 - Provides the client with information about the next hop or hops that a message should take and then the client contacts the next-hop server or UAS directly.
- Registrar server
 - Processes requests from UACs for registration of their current location.
 - Registrar servers are often co-located with a redirect or proxy server.



SIP Messages

- SIP used for Peer-to-Peer Communication though it uses a Client-Server model.
- SIP is a text-based protocol and uses the UTF-8 charset.
- A SIP message is either a **request** from a client to a server, or a **response** from a server to a client.
 - A request message consists of a Request-Line, one or more header fields, an empty line indicating the end of the header fields, and an optional message-body;
 - A response message consists of a Status-Line, one or more header fields, an empty line indicating the end of the header fields, and an optional message-body.
 - All lines (including empty ones) must be terminated by a carriage-return line-feed sequence (CRLF).



SIP Requests

- Requests are also called “Methods”.
- SIP uses SIP Uniform Resource Indicators (URI) to indicate the user or service to which a request is being addressed.
- The general form of a SIP Request-URI is:
 - `sip:user:password@host:port;uri-parameters`
 - `sip:John@doe.com`
 - `sip:+14085551212@company.com`
 - `sip:alice@atlanta.com;maddr=239.255.255.1;ttl=15`
 - Proxies and other servers route requests based on Request-URI.
- Requests are distinguished by starting with a Request-Line.
 - A Request-Line contains a **Method** name, a **Request-URI**, and **SIP-Version** separated by a single space (SP) character.
 - Request-Line = Method SP Request-URI SP SIP-Version CRLF
 - RFC 3261 defines six methods: INVITE, ACK, OPTIONS, BYE, CANCEL, and REGISTER.
 - SIP extensions provide additional methods: SUBSCRIBE, NOTIFY, PUBLISH, MESSAGE, ...
 - SIP-Version should be “SIP/2.0”.
 - Example:
 - Request-Line: INVITE sip:2001@192.168.56.101 SIP/2.0
- The remaining of a request message is one or more header fields, an empty line indicating the end of the header fields, and an optional message-body.

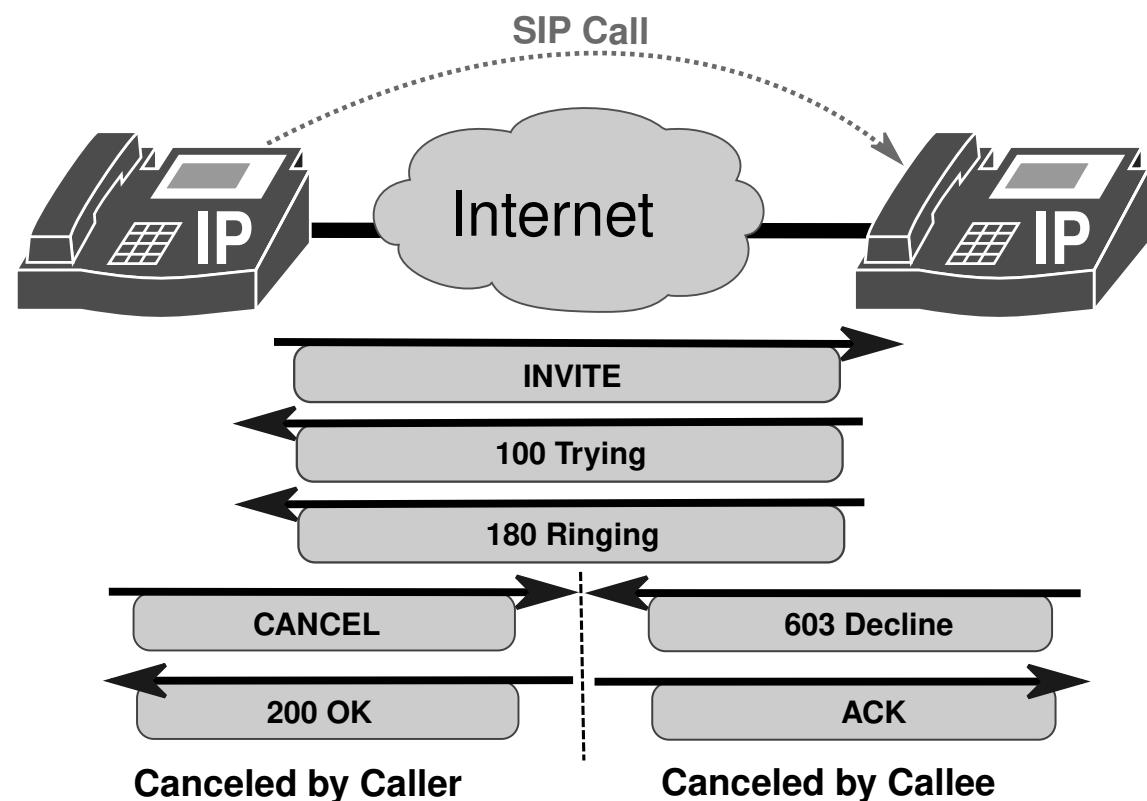
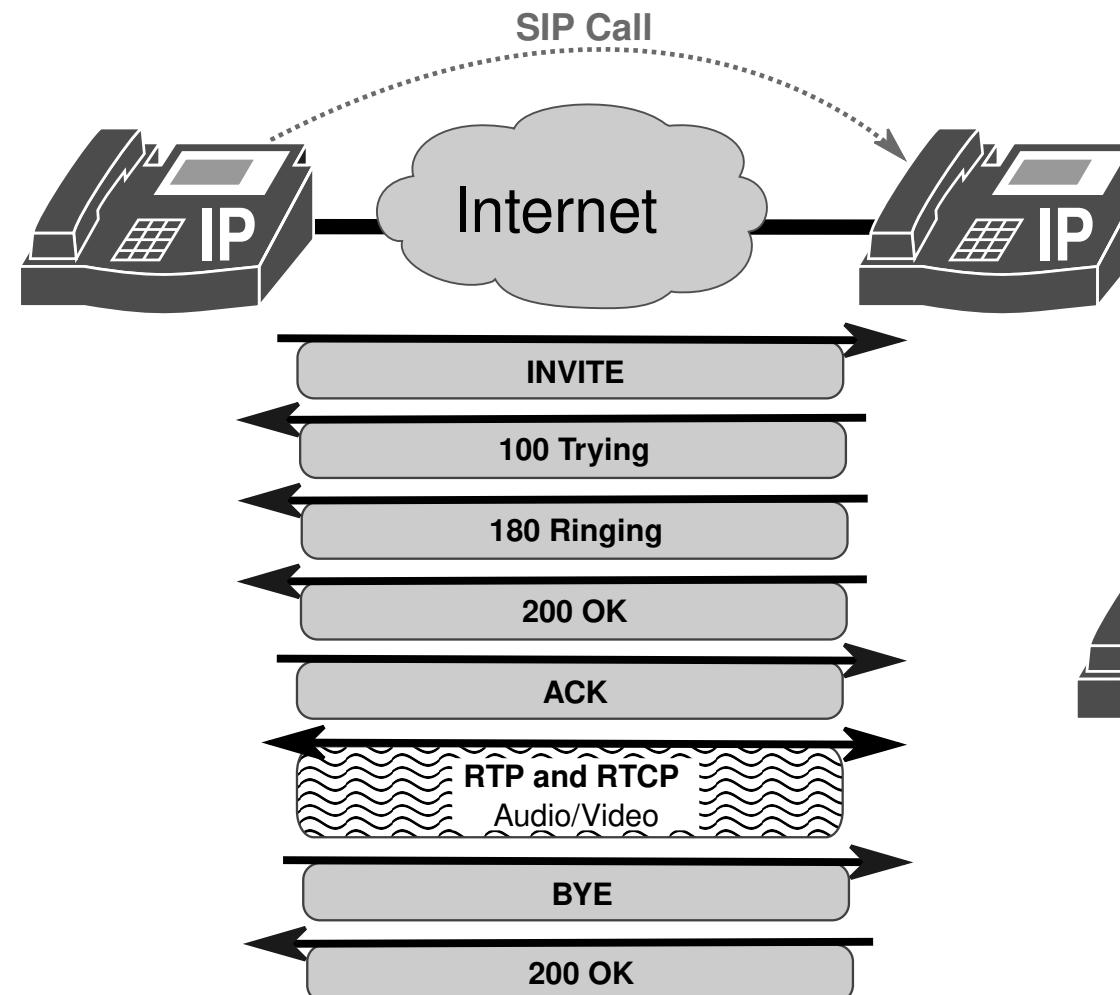


Session Description Protocol (SDP)

- SIP carries (encapsulates) SDP messages.
- When initiating multimedia teleconferences, VoIP calls, streaming video, or other sessions, is required to transmit to participants media details, transport addresses, and other session description metadata.
- SDP (RFC 4566) provides a standard representation for such information, irrespective of how that information is transported.
 - SDP is purely a format for session description.
 - SDP is intended to be general purpose so that it can be used in a wide range of network environments and applications.
 - SDP does not support negotiation of session content or media encodings.



SIP Signaling – Direct Call



Canceled by Caller

Canceled by Callee

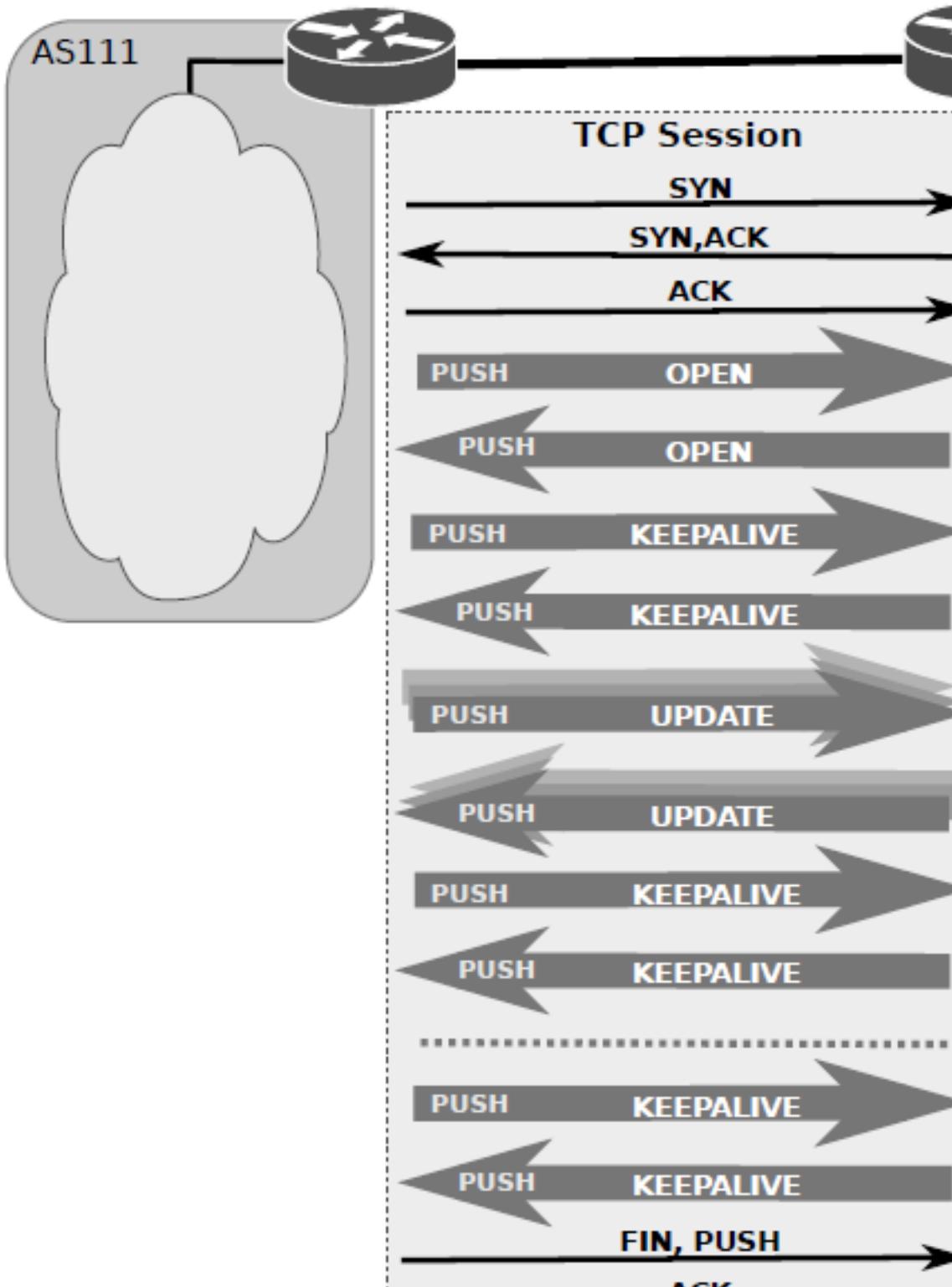


messages are used to
the BGP session.

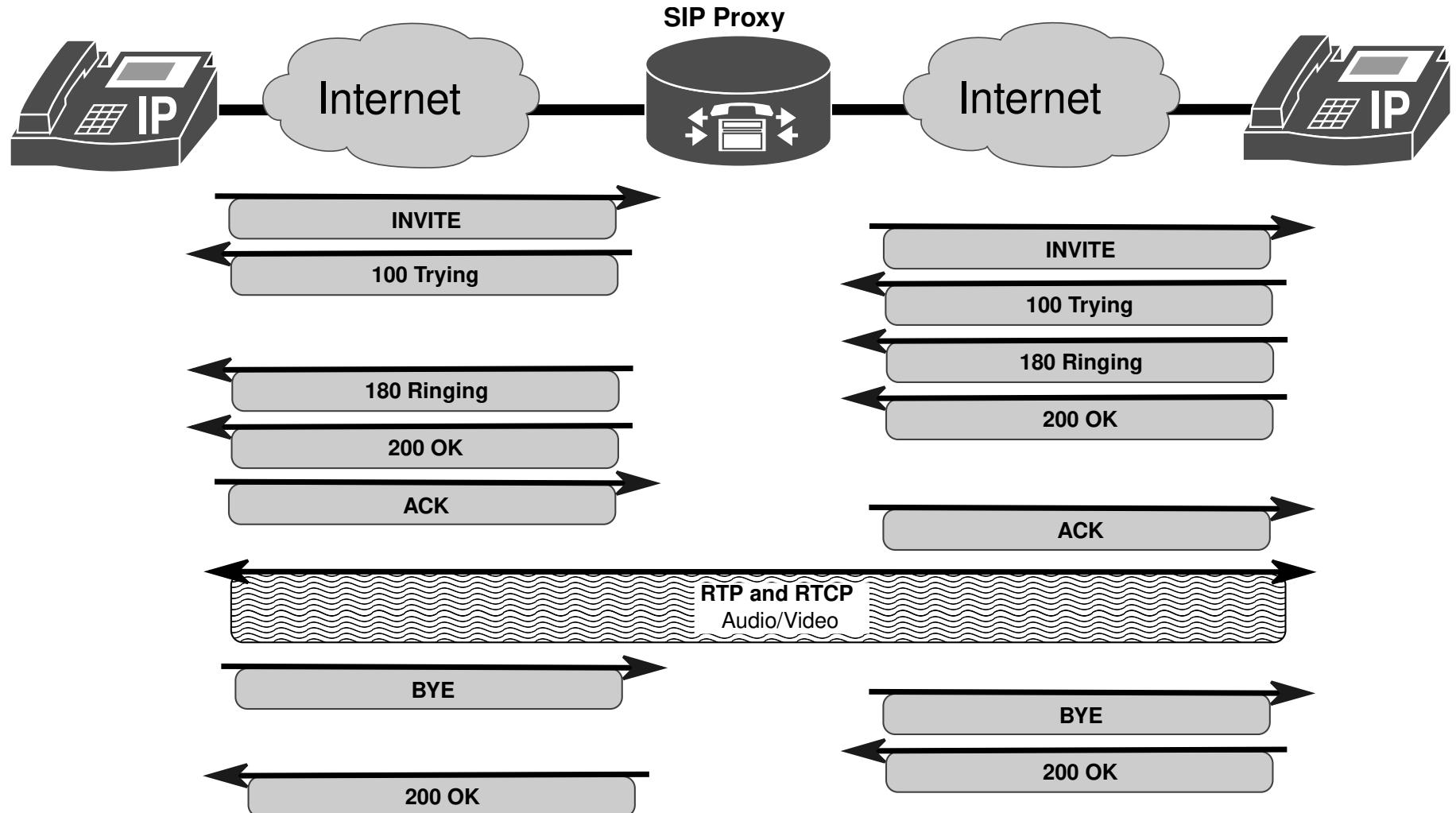
NE messages are used to
iting prefixes, along with
ociated BGP attributes
(the AS-PATH).

LIVE messages are
ed whenever the
e period is exceeded,
an update being
ed.

CATION messages are
enever a protocol error is
, after which the BGP



SIP Proxy Server

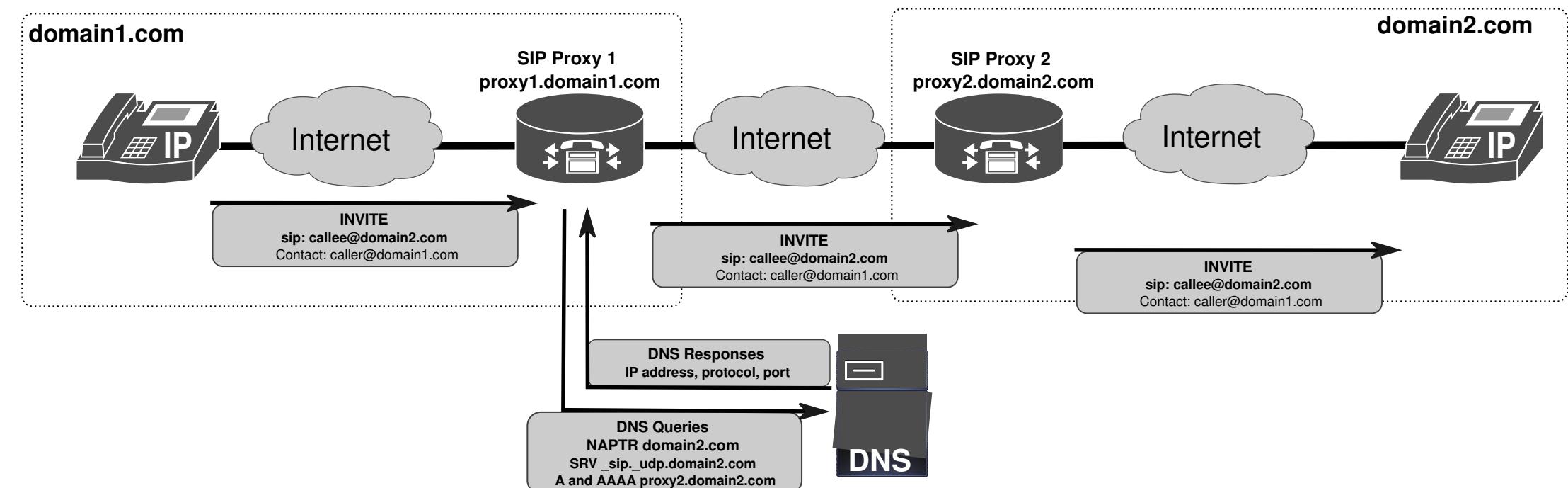


Locating SIP Servers

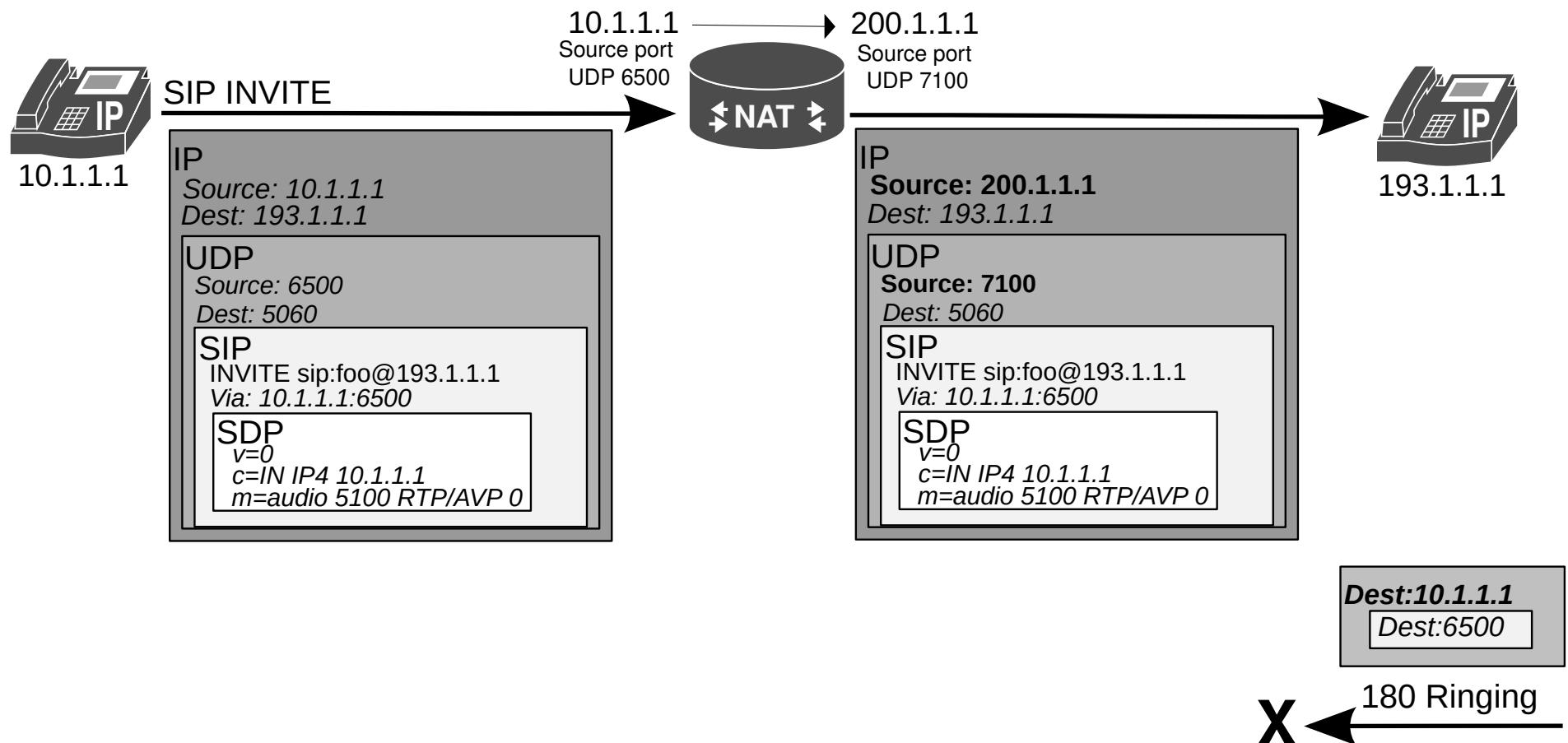
- RFC 3263 defines a set of DNS procedures to locate SIP Servers.
- SIP elements need to send requests/responses to a resource identified by a SIP URI.
 - The SIP URI may identify the desired target resource or a intermediate hop towards that resource.
 - Requires **Transport protocol, IP address and Port**.
 - If the URI specifies any of them, then it should be used.
 - Otherwise, must be retrieved from a DNS server.
 - Using **Service (SRV)** and **Name Authority Pointer (NAPTR)** DNS records.
- NAPTR records provide a mapping from a domain name to:
 - A SRV record (that contains the resource responsible server name),
 - And, the specific transport protocol.
- Example:
 - A client/server that wishes to resolve “sip:user@example.com”,
 - Performs a NAPTR query for domain “example.com”,
 - IN NAPTR 100 50 "s" "SIP+D2U" "" _sip._udp.example.com.
 - Has UDP as possible transport protocol, performs a SRV query for “_sip._udp.example.com”
 - IN SRV 0 1 5060 server1.example.com
 - IN SRV 0 2 5060 server2.example.com
 - Has two possible servers, performs A and AAAA queries for the chosen server.



SIP Proxy Forwarding



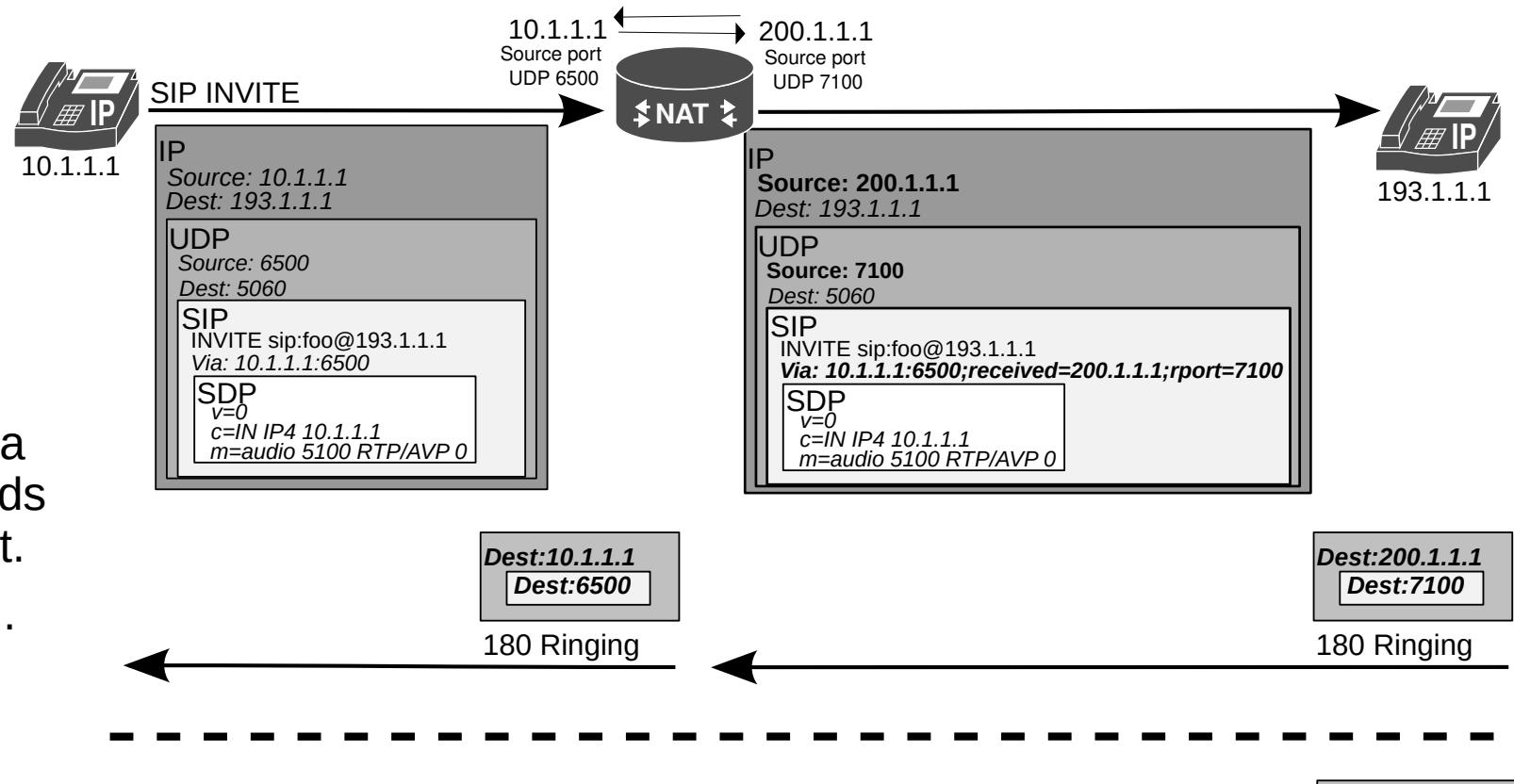
SIP and NAPT



SIP NAPT Traversal

- Symmetric Response Routing (RFC 3581).

- SIP payload is also “translated”, by adding a **received** and **rport** fields with public address/port.
- SDP remains unchanged.

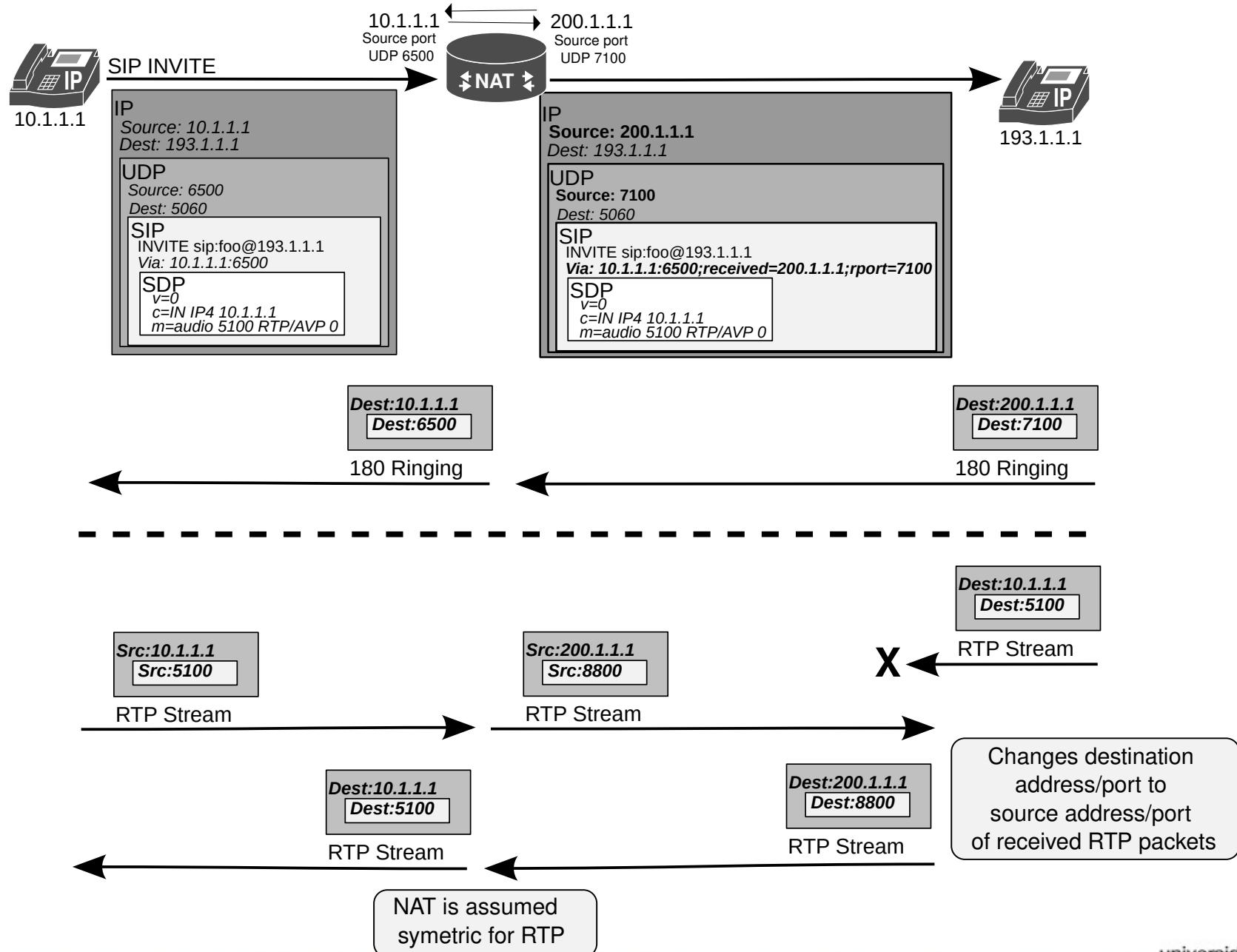


- Media traversal (RTP/RTCP) is still a problem.

- SDP contents mismatch with public address/port.
- Possible solutions
 - Let clients (on private network) find out their public address/port and rewrite SDP payload.
 - Manual configuration (when NAT uses static translations).
 - Automatic discovery (when NAT is dynamic) using STUN protocol.
 - Symmetric (RTP/RTCP) NAT (RFC 4961).
 - NAT SIP Application Layer Gateway (ALG).

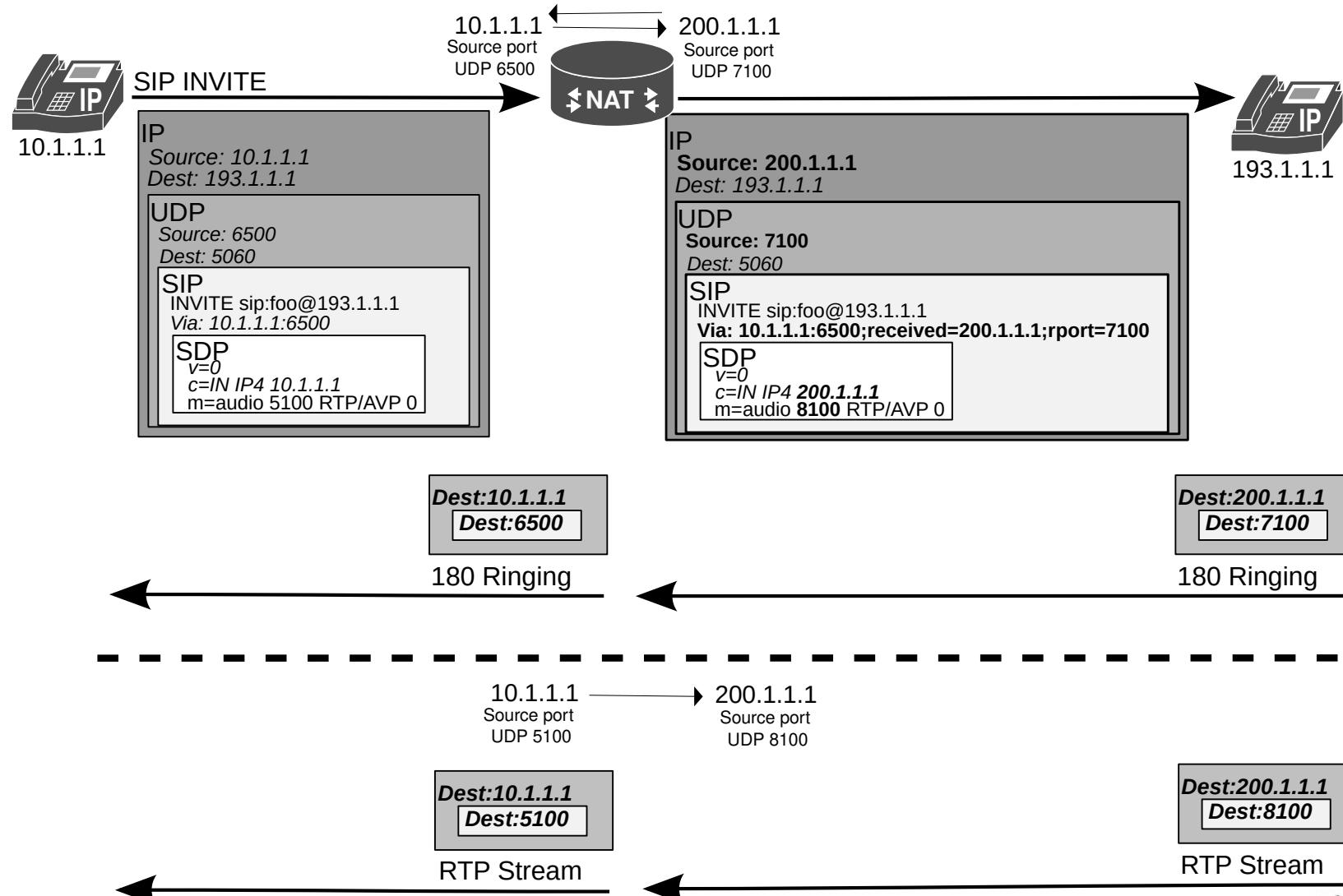


Symmetric (RTP/RTCP) NAT



NAT SIP Application Layer Gateway (ALG)

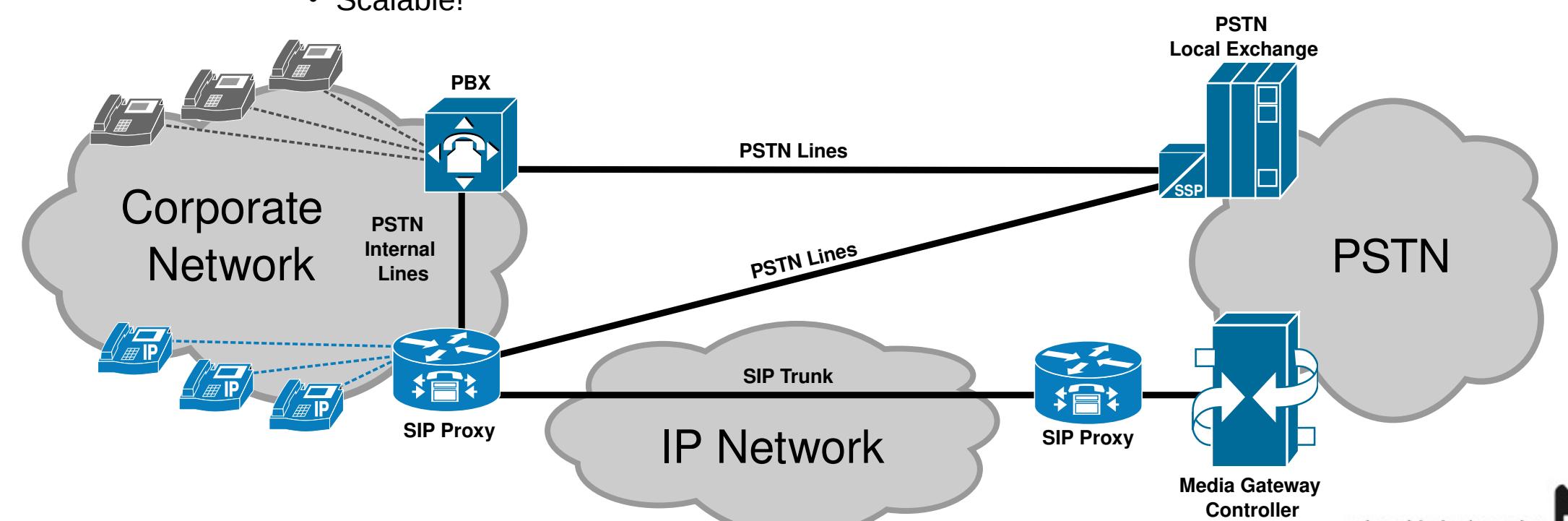
- Required to translate SDP payloads.
- Heavy on NAT gateway.



VoIP and PSTN Connectivity

- SIP proxy.

- With PSTN interface (to ISP or local PBX).
 - Requires multiple PSTN Lines.
 - Not scalable.
- With SIP trunk to remote SIP proxy.
 - Remote proxy/gateway interfaces with PSTN network.
 - Remote proxy/gateway owned by PSTN ISP or by a third-party entity.
 - Usually TCP/IP transport with a TLS security layer.
 - Scalable!

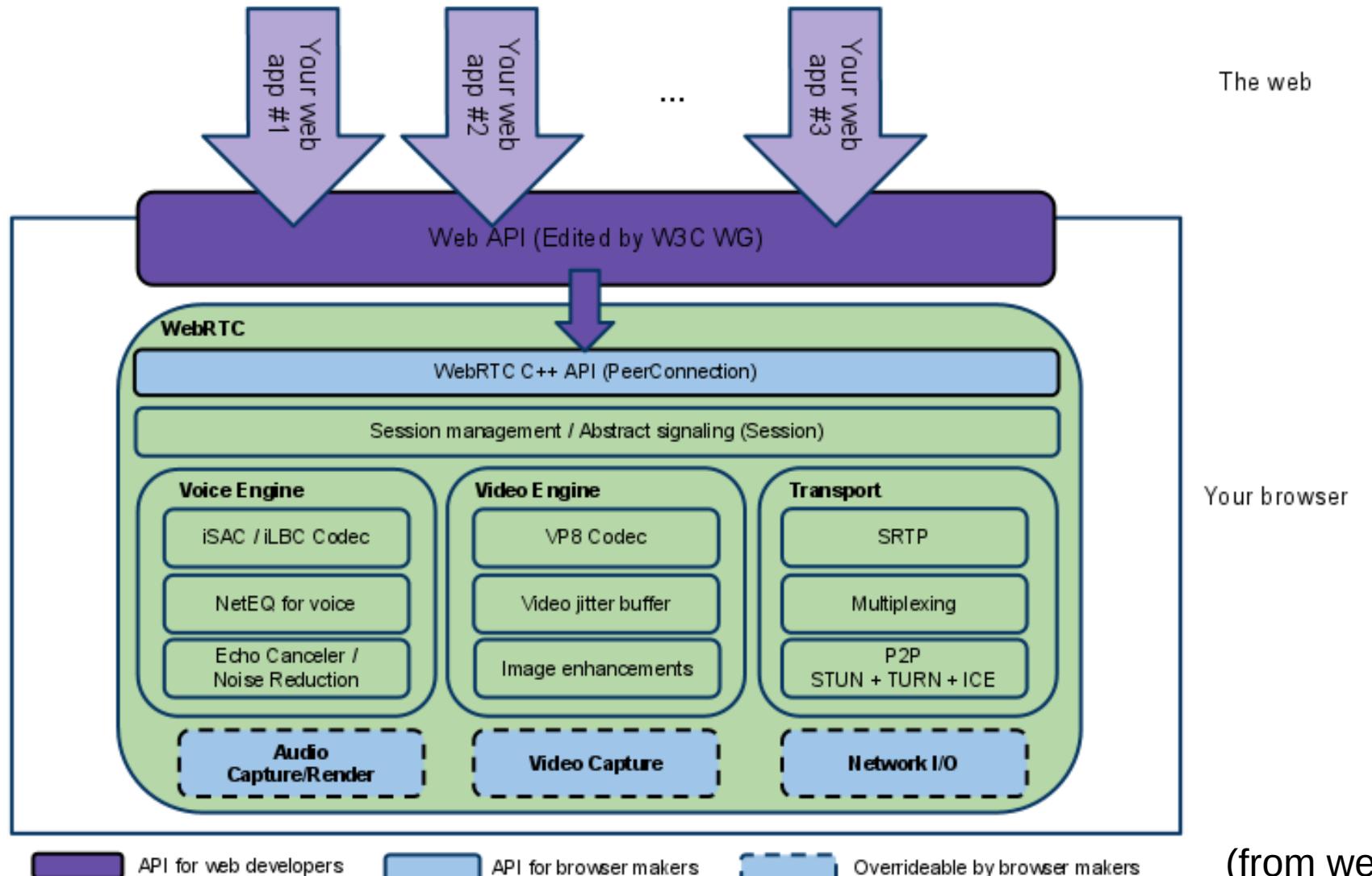


WebRTC

- WebRTC (Web Real Time Communications) is an open source communication technology.
- Typically used for real-time audio and video communications.
- Provides:
 - Peer-to-peer connections.
 - An instance allows an application to establish peer-to-peer communications with another instance in another browser, or to another endpoint implementing the required protocols.
 - RTP Media transport.
 - Allow a web application to send and receive media stream over a peer-to-peer connection.
 - Peer-to-peer Data transport.
 - Allows a web application to send and receive generic application data over a peer-to-peer connection.
 - Peer-to-peer DTMF.



WebRTC Architecture



Multicast Routing

Redes de Comunicações II

**Licenciatura em
Engenharia de Computadores e Informática
DETI-UA**

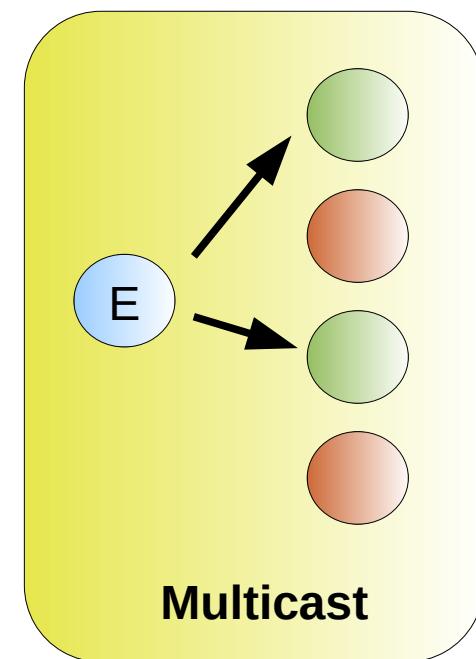
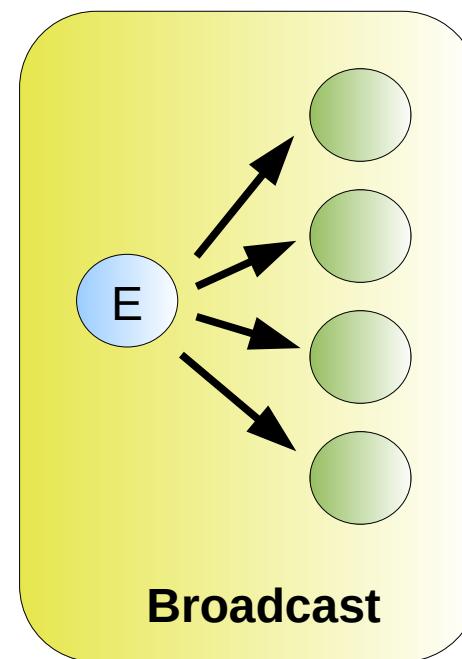
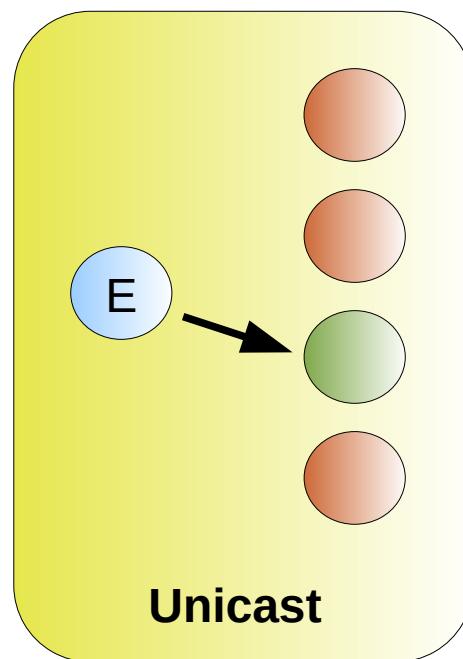


universidade de aveiro

deti.ua.pt

Multicast

- Multicast communications refer to one-to-many and many-to-many communications.
 - ◆ Multicast at application level.
 - ◆ Multicast at network level.



Multicast Abstraction

- Information transmitted by one origin application is received by multiple destination applications in different stations.
- Alternative 1: TCP/IP protocol stack of the sender station establishes point-to-point connections with all destinations and send multiple copies, one for each destination.
- Advantages
 - ◆ Allows the use of networks without multicast capabilities.
 - ◆ Allows the use of TCP protocol with all its advantages.
- Disadvantages
 - ◆ Requires that the sender application specifies the destination address list.
 - ◆ Results in an inefficient use of network resources.



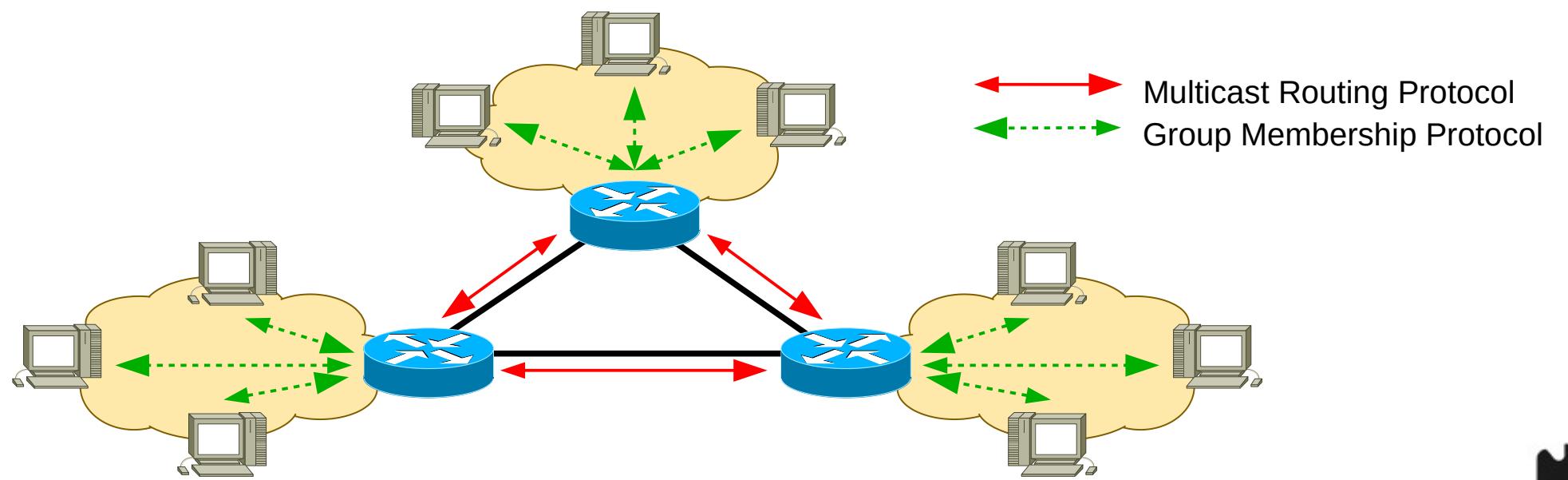
Multicast Abstraction

- Alternative 2: The station sends each IP packet only once and the network is responsible for the copy of the packet to multiple destinations.
- Disadvantages
 - ◆ Requires that the network has multicast capability.
 - ◆ Requires the use of only UDP with all its disadvantages.
- Advantages
 - ◆ It is possible to have a better usage of network resources.
- Issues
 - ◆ How do the sender stations specify the destination stations?!
 - ◆ How do the routers implement multicast capabilities?!



Multicast Abstraction

- IP multicast networks are not “*connectionless*” networks as unicast networks.
- It is required to establish multicast paths between the routers for them to know how to route the multicast packets.
- Then, signalling is required (between the stations and the routers) and routing protocols (between the routers) to establish the required multicast paths.



Identification of destination stations

- Explicit identification may not be desirable.
- It makes use of **IP addresses of class D** (starting by 1110) in IPv4, and addresses of the type FF00::/16 in IPv6.
- The participating stations agree on the use of an address that identifies the session.
- The destination stations announce to the routers their participation in the multicast session identified by the agreed address.
 - ◆ IGMP (v4), MLD (v6).
- The routers route the IP packets sent with this agreed address to all networks in which there are participating stations.
 - ◆ Routing protocols DVMRP, MOSPF, PIM, etc...



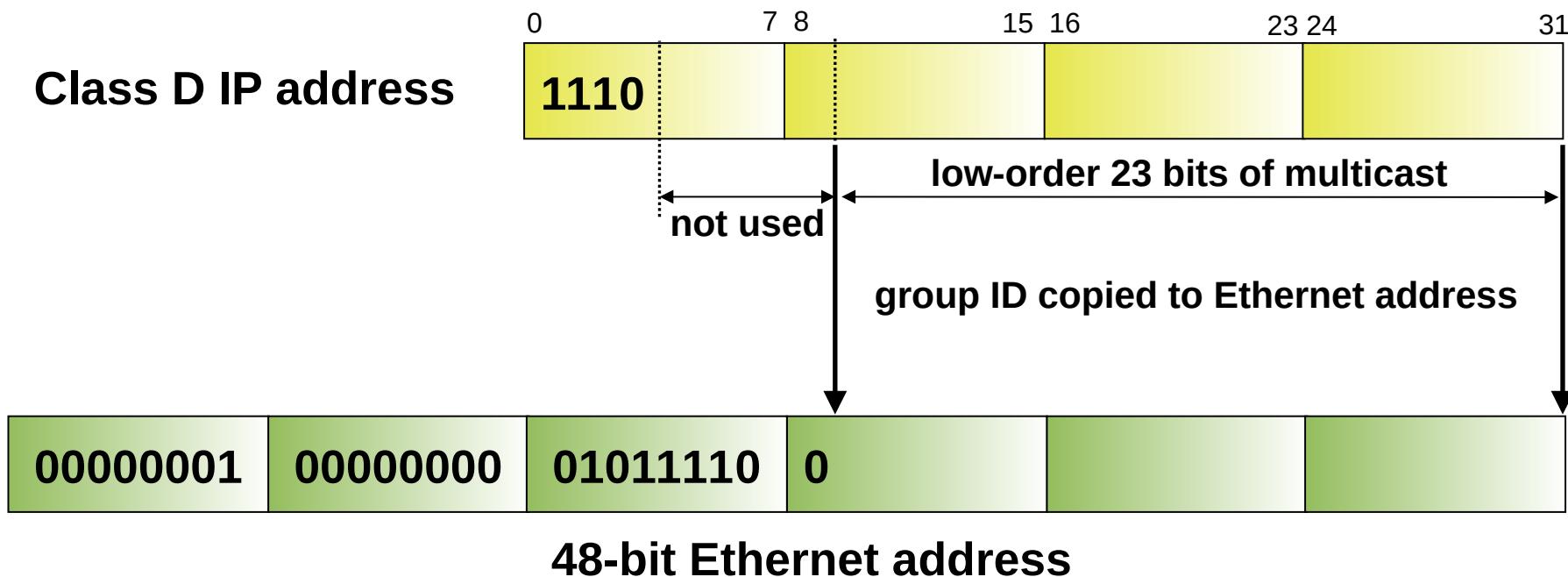
Classe D Addresses



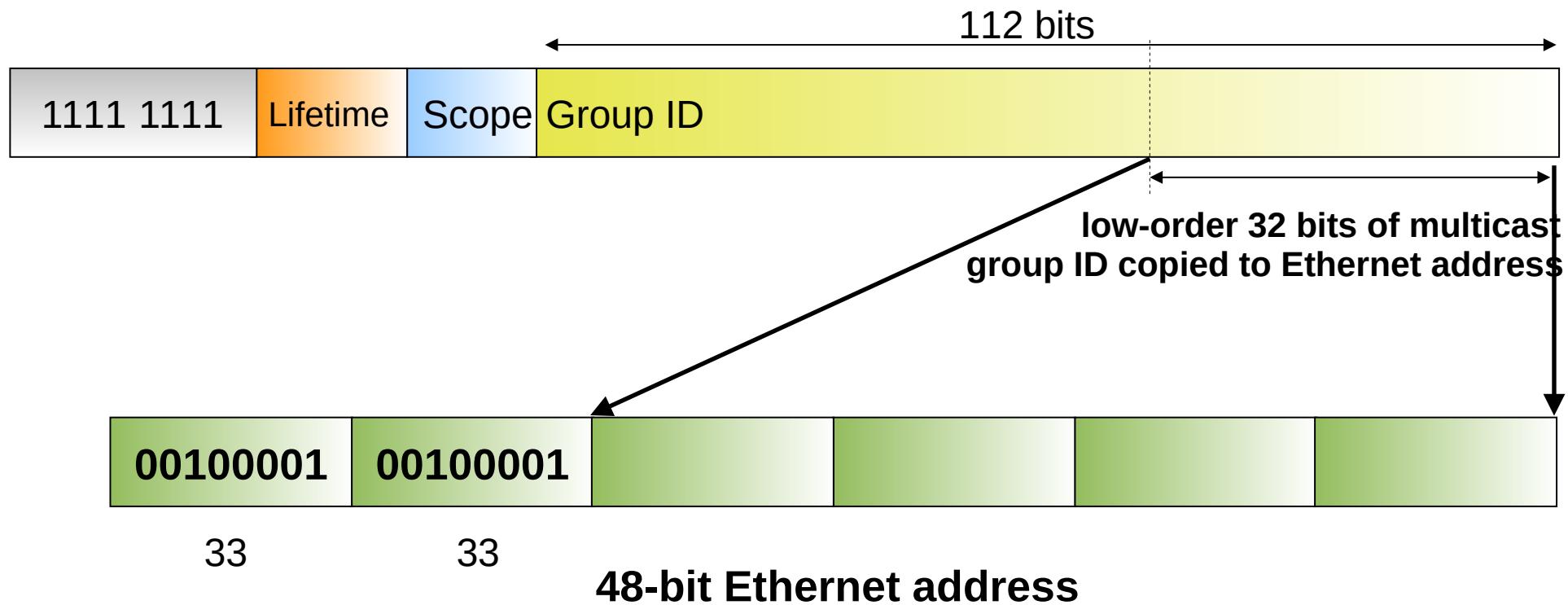
- Some addresses are reserved by the Internet Assigned Numbers Authority (IANA).
- Some addresses have their own meaning. For example, addresses on the form 224.0.0.X:
 - ◆ 1 : All Hosts
 - ◆ 2 : All Multicast Routers
 - ◆ 4 : All DVMRP Routers
 - ◆ 5 : All OSPF routers
 - ◆ 6 : OSPF designated routers
 - ◆ 13 : All PIM routers
- 224.0.0.1 to 224.0.0.225 reserved to routing protocols and other discovery/maintenance protocols..
- The range 239.0.0.0 to 239.255.255.255 is destined to private networks.



Conversion of class D IPv4 address to IEEE 802 address

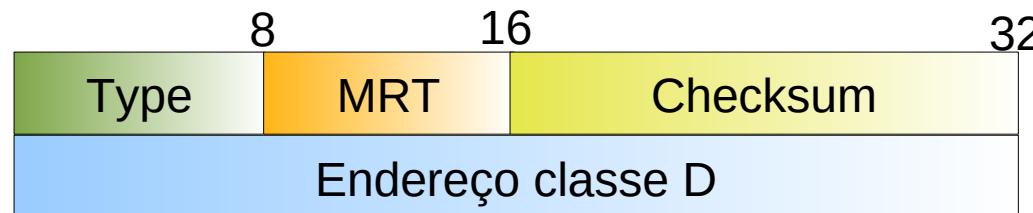


IPv6 Multicast address conversion to IEEE 802 address



Internet Group Membership Protocol (IGMP)

- IGMP version 2, RFC 2236
 - ◆ Operates between the station and the directly connected routers.
 - ◆ Is used for the station to announce to the router that it wants to participate in the multicast session (identified by a class D address).



MRT - Maximum Response Time

- IGMP runs through IP protocol (protocol type = 0x02).
- The packets are sent to the destination address 224.0.0.1 ("All Hosts") with TTL= 1.



IGMP Messages

- GMQ - General Membership Query
 - ◆ Sent by the routers to ask the stations if they participate in a multicast session.
- SMQ - Specific Membership Query
 - ◆ Sent by the routers to ask if there is any station that participates in a specific multicast session.
- MR - Membership Report
 - ◆ Sent by the stations to signal that they participate in a multicast session .
- LGR - Leave Group Report (Optional)
 - ◆ Sent by the stations to signal that they will leave a multicast session
 - ◆ Update can be done through the Membership Report.
- In each network, the Querier Router is the router with lower IP address among all interfaces connected to the network, and is the one that maintains the IGMP messages with the terminals.

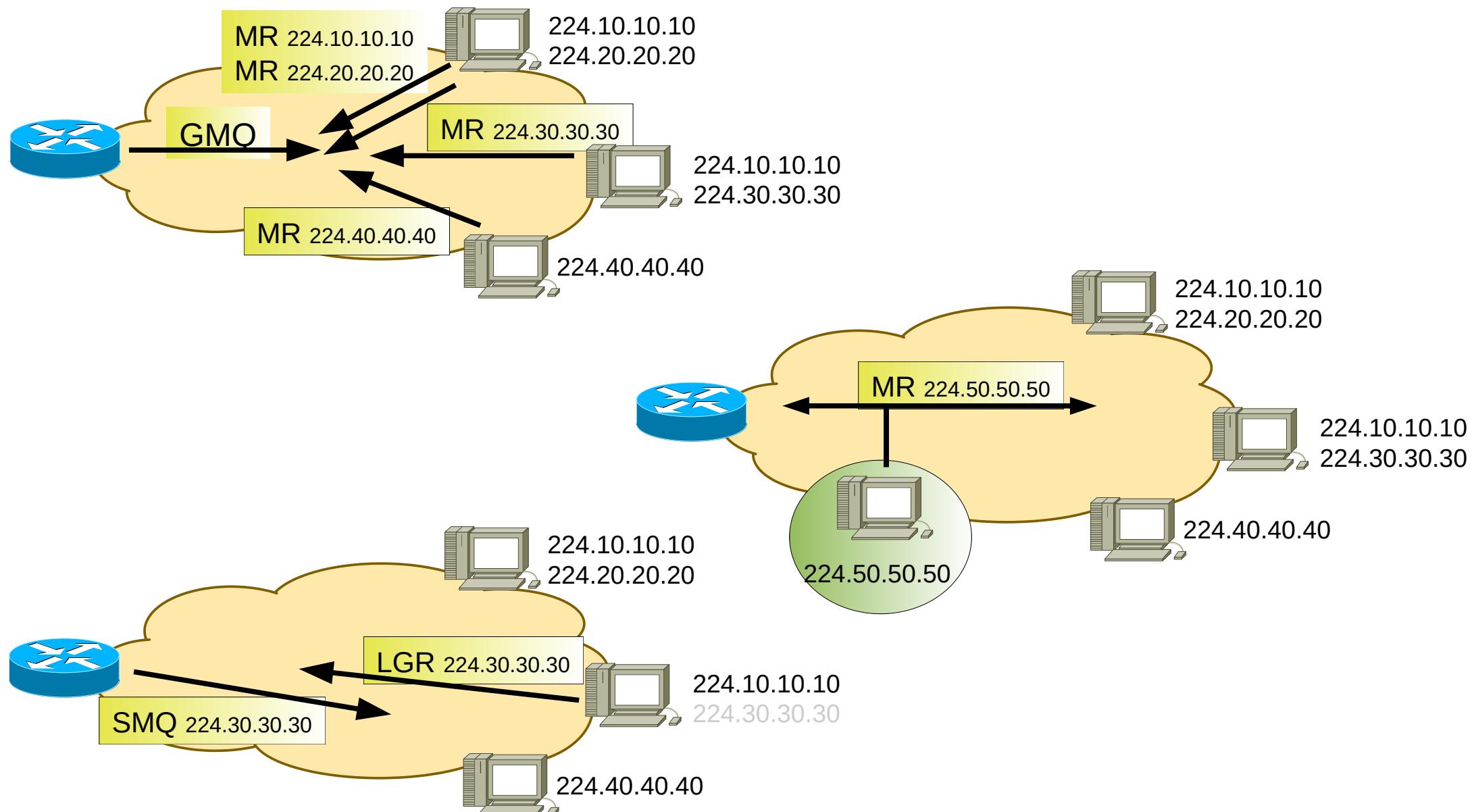


IGMP Protocol

- Routers send periodically a GMQ specifying a *Maximum Response Time* (MRT).
 - Each station waits a random time between 0 and MRT to answer to a MR specifying a *multicast address*.
 - If in the meanwhile the station ‘sees’ a MR for the same session, it aborts the sending of the MR.
-
- Each station sends a MR when it wants to belong to a multicast session .
 - Optionally, a station sends a LGR when it does not belong anymore to a multicast session.
 - When a router receives a LGR, it sends a SMQ to verify if there are still any stations belonging to that session.



IGMP Protocol



IGMPv1/v2 – Final Conclusions

- Any station can join a multicast session receiving and sending information.
- The formation of multicast sessions is initiated by the receivers.
 - ◆ Senders do not specify nor control the stations that can receive information.
- The network does not provide filtering, ordering or privacy to multicast packets.
- The multicast IP service model follows the same philosophy of the unicast:
 - ◆ Simple and reliable protocol layer in which additional functionalities are provided by the upper layers.



IGMPv3

- IGMPv3 adds support to "source filtering".
 - ◆ Allows a terminal to report interest in a specific multicast session/group, from
 - ONLY a specific source.
 - INCLUDE Mode.
 - ALL sources EXCEPT specific sources.
 - Exclude Mode.
 - A blank list means interest in all sources.
 - ◆ Allows simultaneous requests to multiple multicast sessions.
- Has a new “Report” message format.
 - Version 3 Membership Report.
- Allows IGMPv1 and IGMPv2 interoperability.
 - ◆ Supports Version 1 Membership Report, Version 2 Membership Report and Version 2 Leave Group.



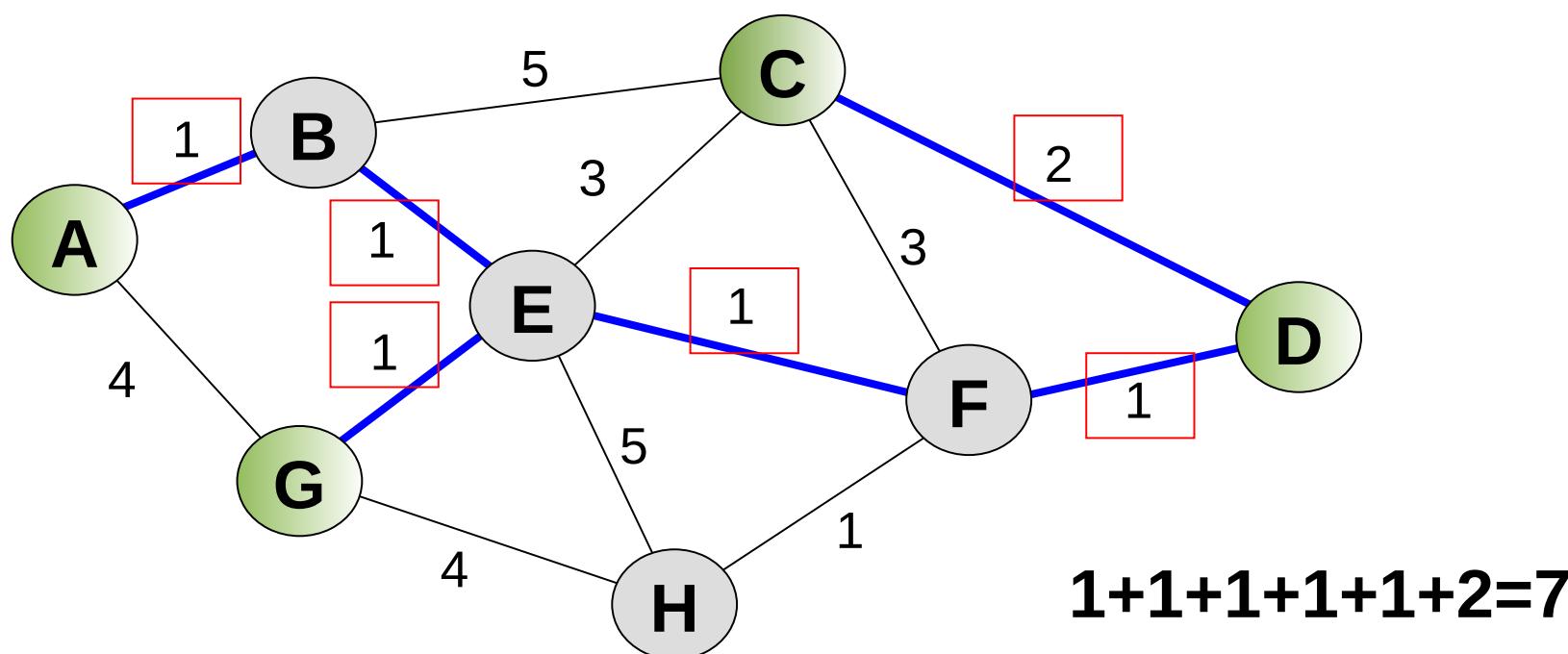
Multicast Routing

- Group-shared tree
 - ◆ It is based on determining a routing tree per multicast session that connects all routers with stations belonging to the session.
 - ◆ Minimal spanning (Steiner tree).
 - ◆ It is not used in practice:
 - High computational complexity.
 - Requires information about the overall network.
 - Monolithic: executed every time a router needs to join/leave a session.
 - ◆ Minimum cost tree to a central node (“rendezvous point”).
- Source-based tree
 - ◆ It is based on determining a routing tree, per multicast session, and per sender.
 - ◆ One router is identified as the “central point”.



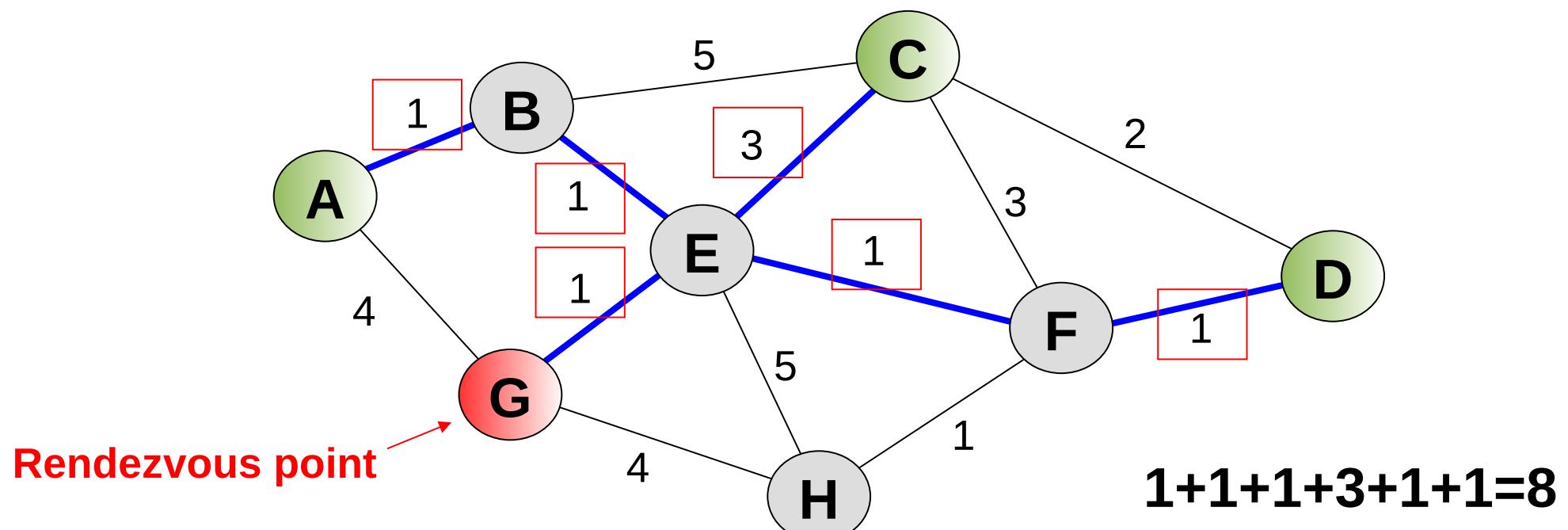
Group-shared tree

- Steiner Tree: determines the minimum cost tree that interconnects all nodes with stations of a session.
 - ◆ Requires a link-state protocol.
 - ◆ Algorithm with exponential complexity.



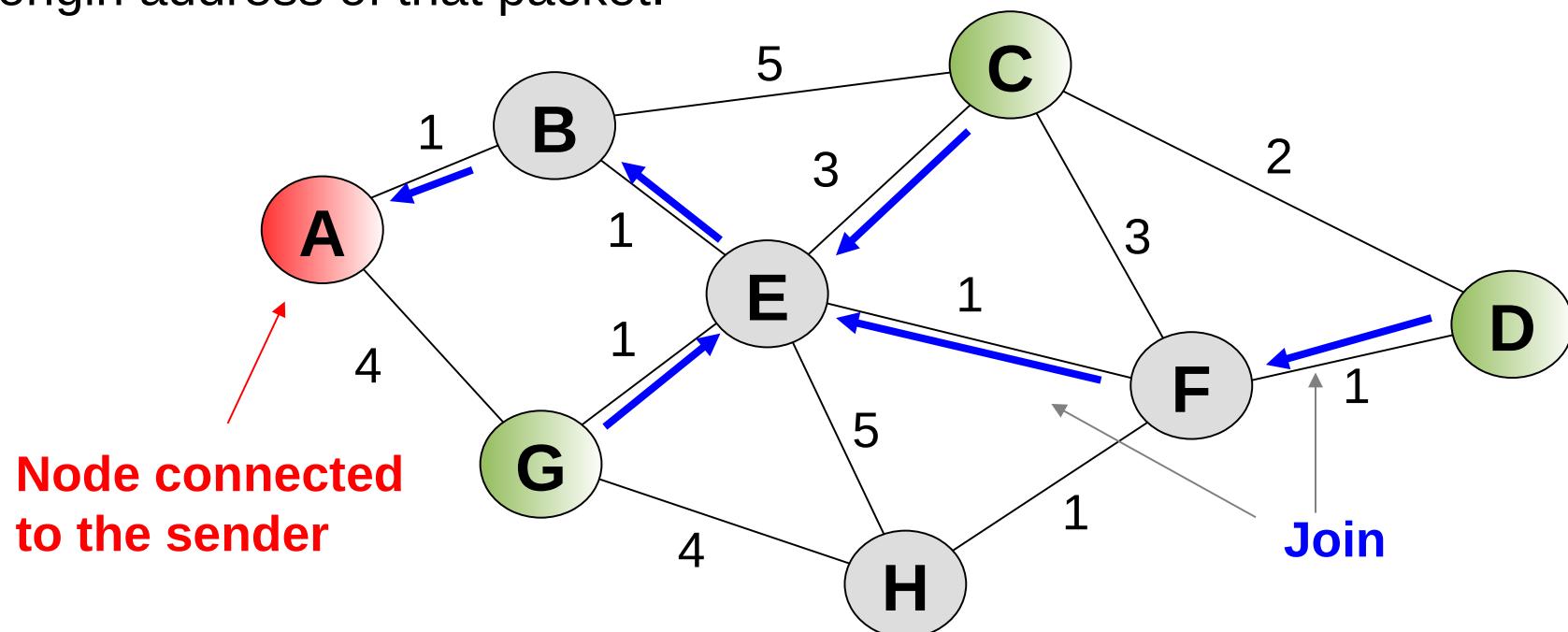
Group-shared tree

- Minimum cost tree of the paths to a central node (“rendezvous point”).
 - ◆ Central node is previously chosen (belonging to the tree even if it does not contain stations that belong to the session) and known by other nodes.
 - ◆ To join the tree, the nodes with stations belonging to the session send a “join” message by the path with minimum cost between them and the central node (rendezvous point).



Source-based tree (when the sender is known)

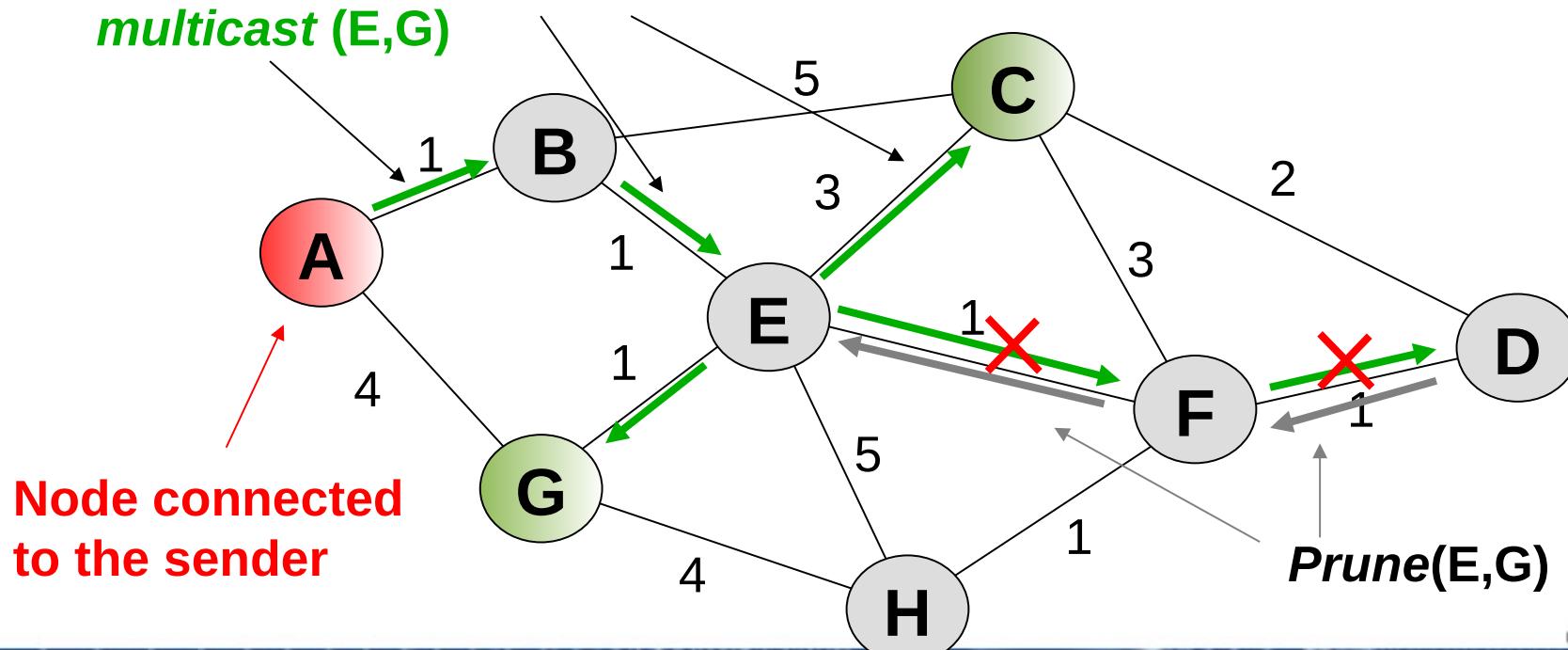
- Each node with receivers interested in a multicast session G of a sender with address E sends a $\text{join}(E, G)$ message towards the address of E by the minimum cost path (unicast path).
- In the path of the join message, each node receives the message in interface F0, re-sends it through interface F1 and builds a routing table (E, G) entry stating that multicast packets that enter in F1 can be routed through F0.
- Multicast routing is based not only on the destination address, but also on the origin address of that packet.



Source-based tree (when the sender is known)

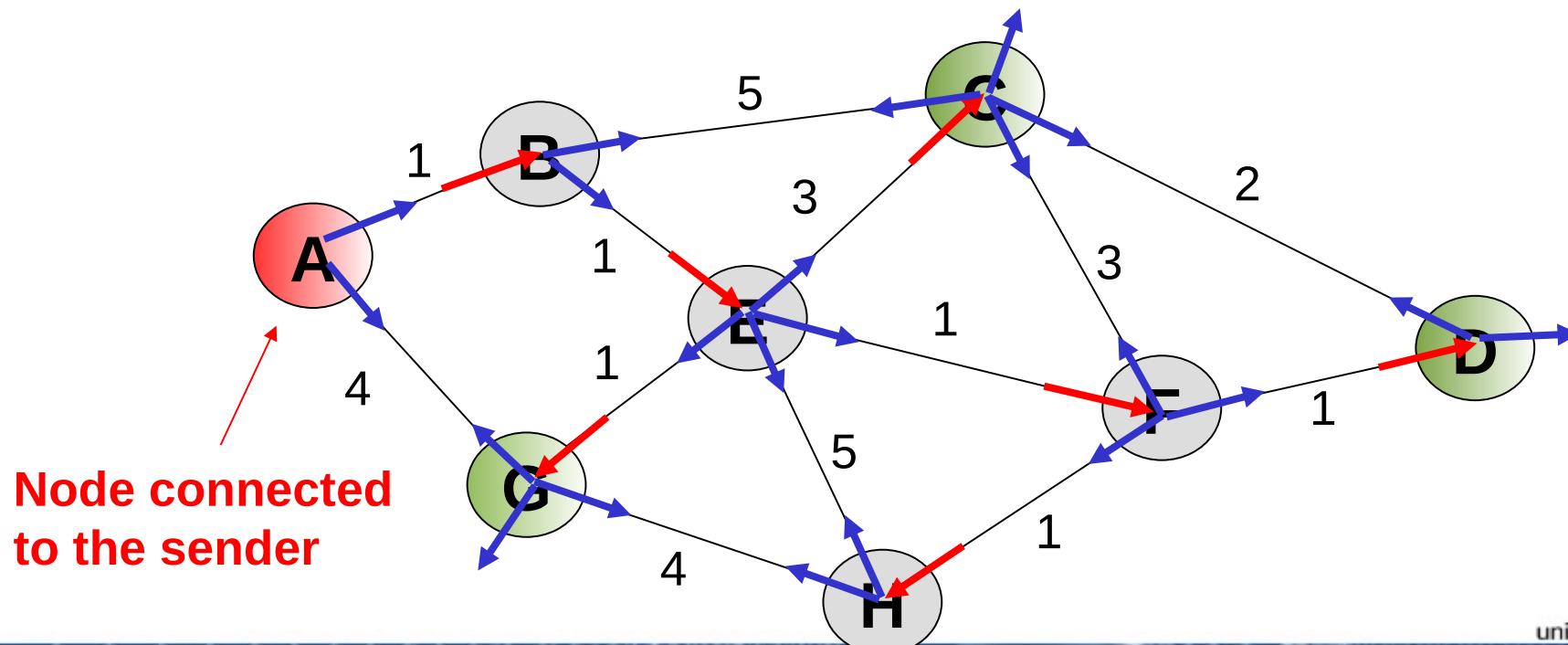
- When there are multiple senders, several multicast routing trees are established, one by each sender.
- When a receiver is not interested in a multicast session, its connected node sends a prune(E,G) message towards the address of sender.
- The prune message is re-sent by the nodes that do not belong anymore to the multicast routing tree.

Árvore de encaminhamento
multicast (E,G)



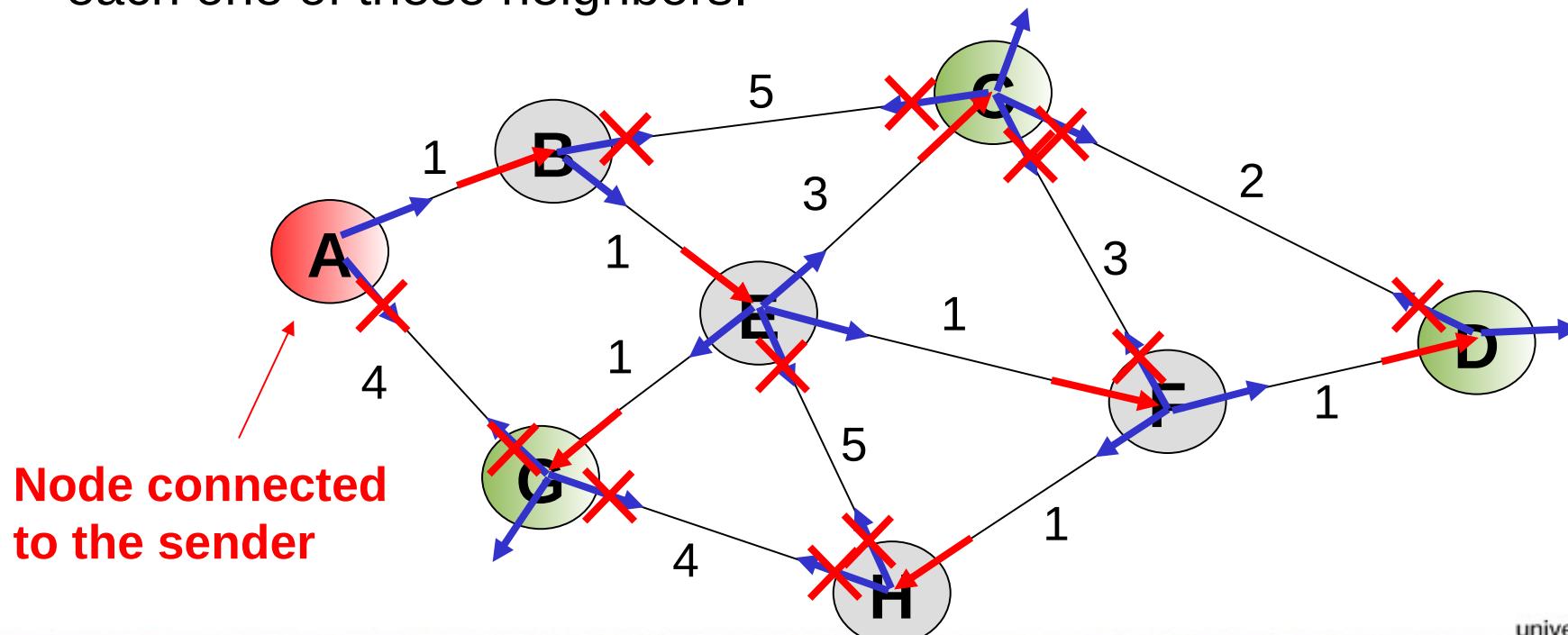
Source-based tree (when the sender is unknown)

- “Reverse Path Forwarding”: creation of a virtual spanning tree
 - ◆ Each node N routes the packets from the source node O, which are received from neighbor V, to all its neighbors only if V is the last node in the minimum cost path from O to N.
 - ◆ Figure: in node E, packets originated from node A are routed to all other ports only if they come from node B because this is the previous node of E in the minimum cost path from A to E.



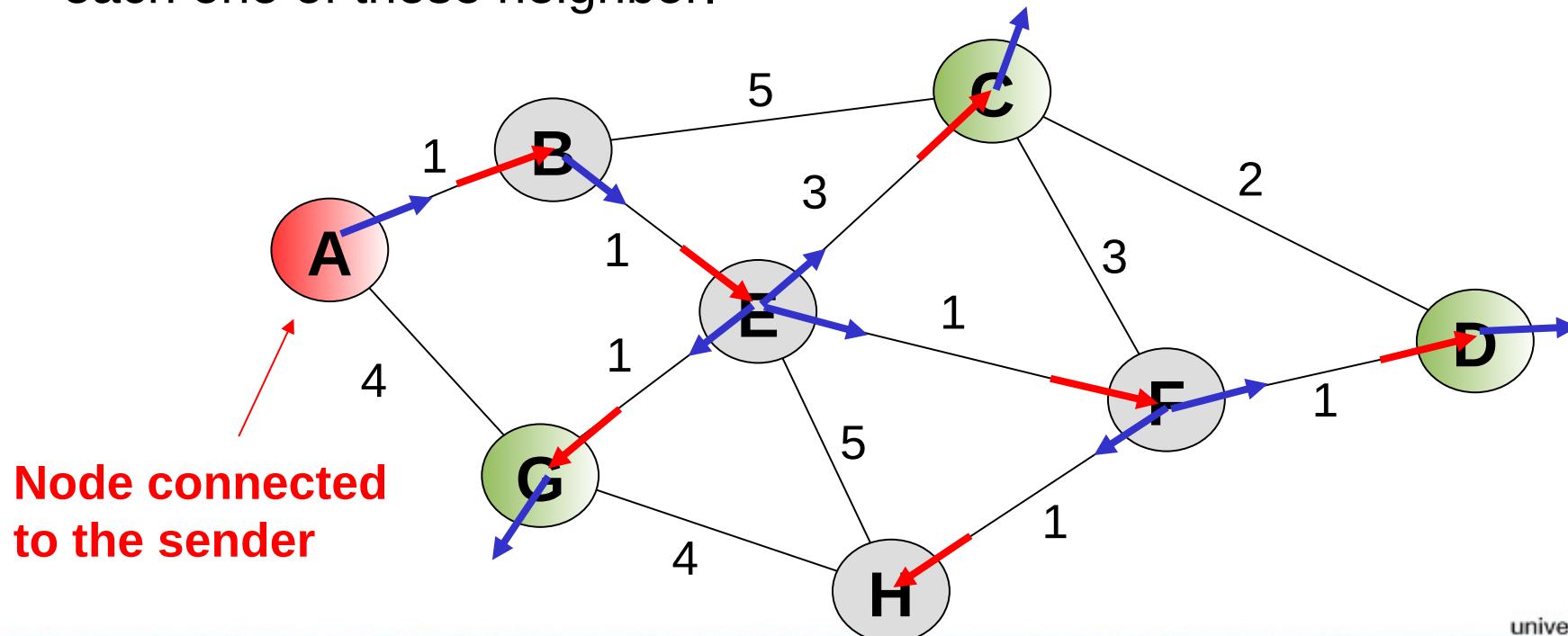
Source-based tree (when the sender is unknown)

- “Reverse Path Forwarding with Pruning” (I)
 - ◆ For each origin node O, node N knows to which neighbors V it is the last node in the minimum cost path from O to V.
 - ◆ Routing is made to these neighbors.
 - ◆ Figure: node E only routes packets from origin node A to neighbors C, F and G because node E is the last node in the minimum cost path from A to each one of these neighbors.



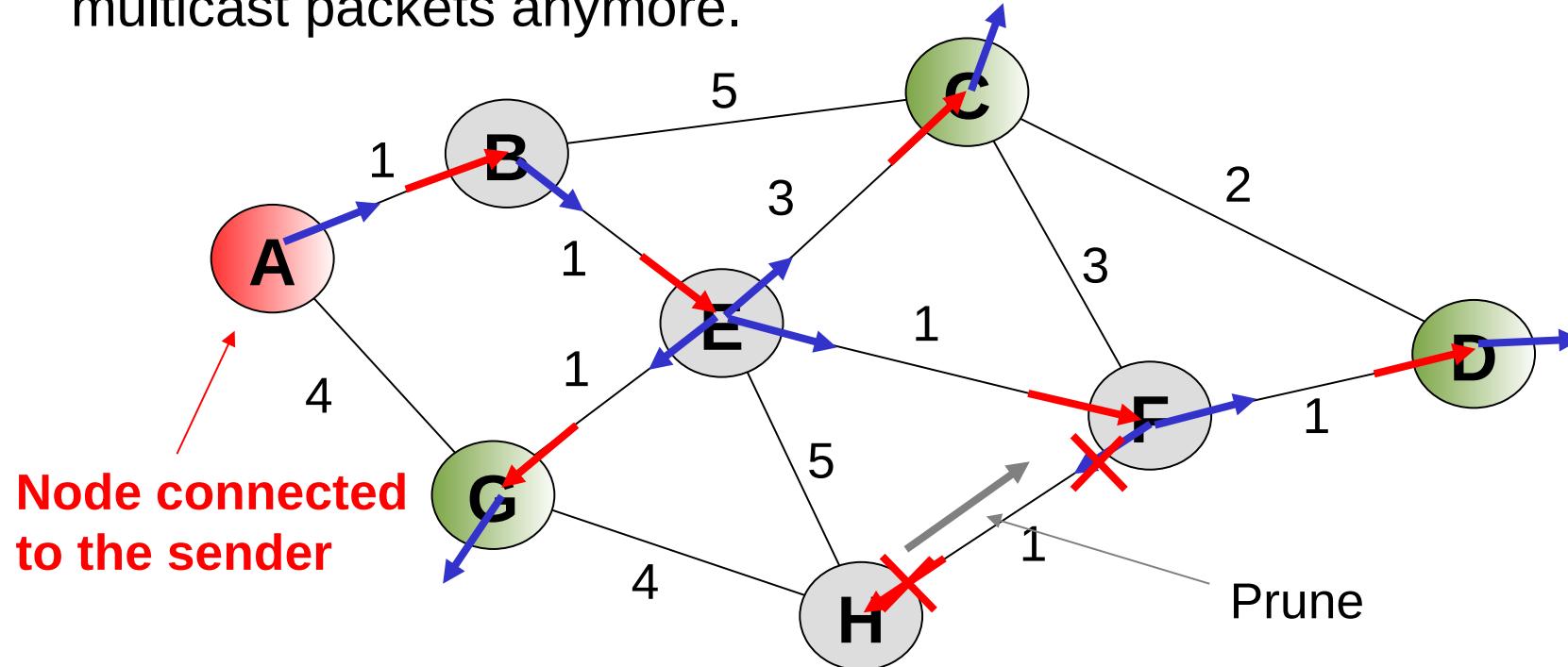
Source-based tree (when the sender is unknown)

- “Reverse Path Forwarding with Pruning” (I)
 - ◆ For each origin node O, node N knows to which neighbors V it is the last node in the minimum cost path from O to V.
 - ◆ Routing is made to these neighbors.
 - ◆ Figure: node E only routes packets from origin node A to neighbors C, F and G because node E is the last node in the minimum cost path from A to each one of these neighbor.



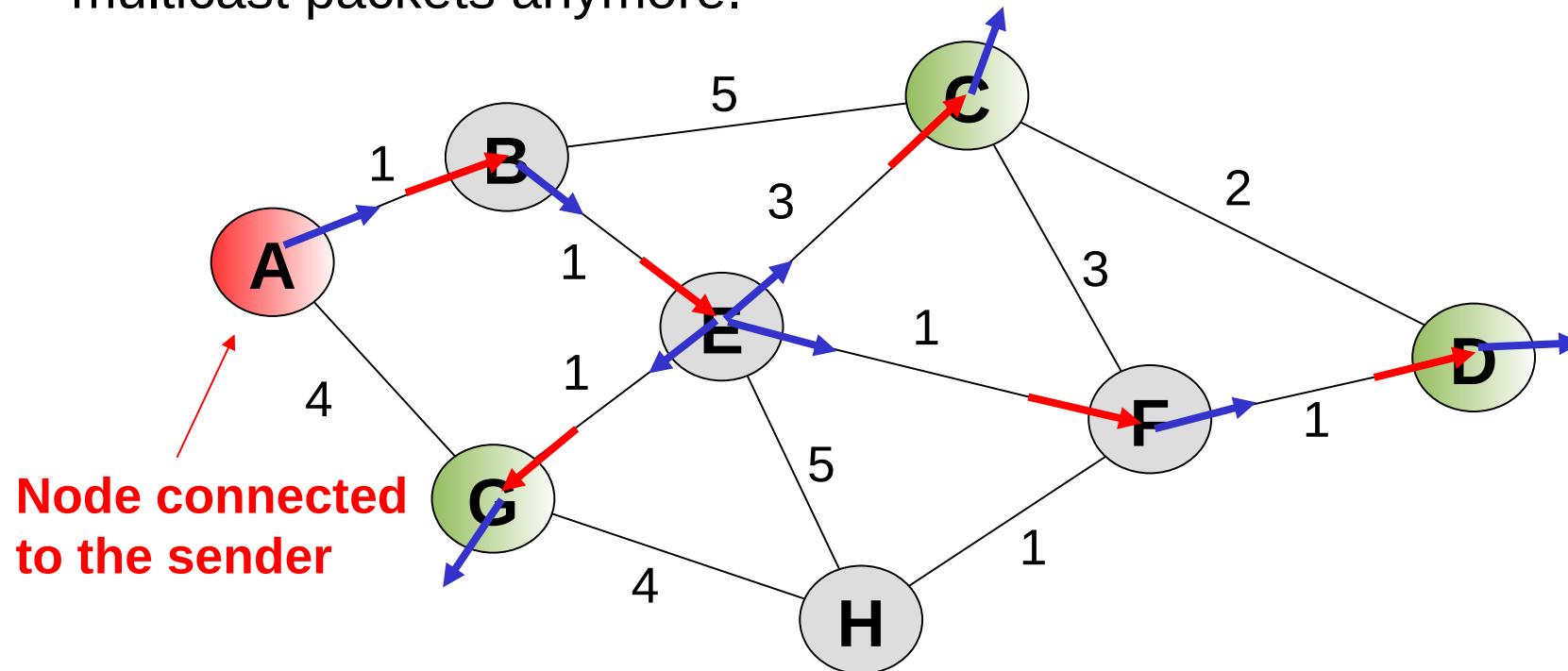
Source-based tree (when the sender is unknown)

- “Reverse Path Forwarding with Pruning” (II)
 - ◆ A node without any terminal stations interested in the multicast session, and without any neighbor nodes to forward the multicast packets to, sends a prune message to the neighbor from which it receives the multicast packets.
 - ◆ Figure: node H sends a prune message to neighbor F and will not receive multicast packets anymore.



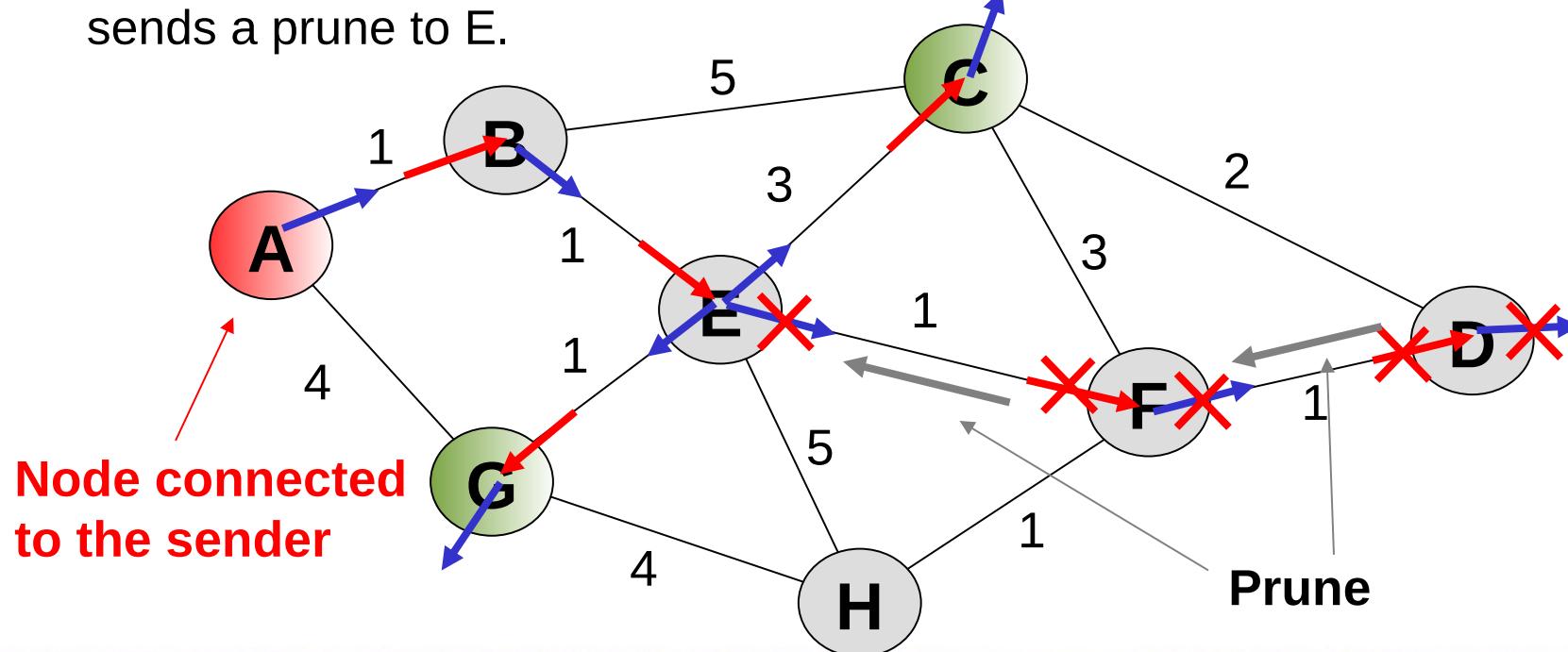
Source-based tree (when the sender is unknown)

- “Reverse Path Forwarding with Pruning” (II)
 - ◆ A node without any terminal stations interested in the multicast session, and without any neighbor nodes to forward the multicast packets to, sends a prune message to the neighbor from which it receives the multicast packets.
 - ◆ Figure: node H sends a prune message to neighbor F and will not receive multicast packets anymore.



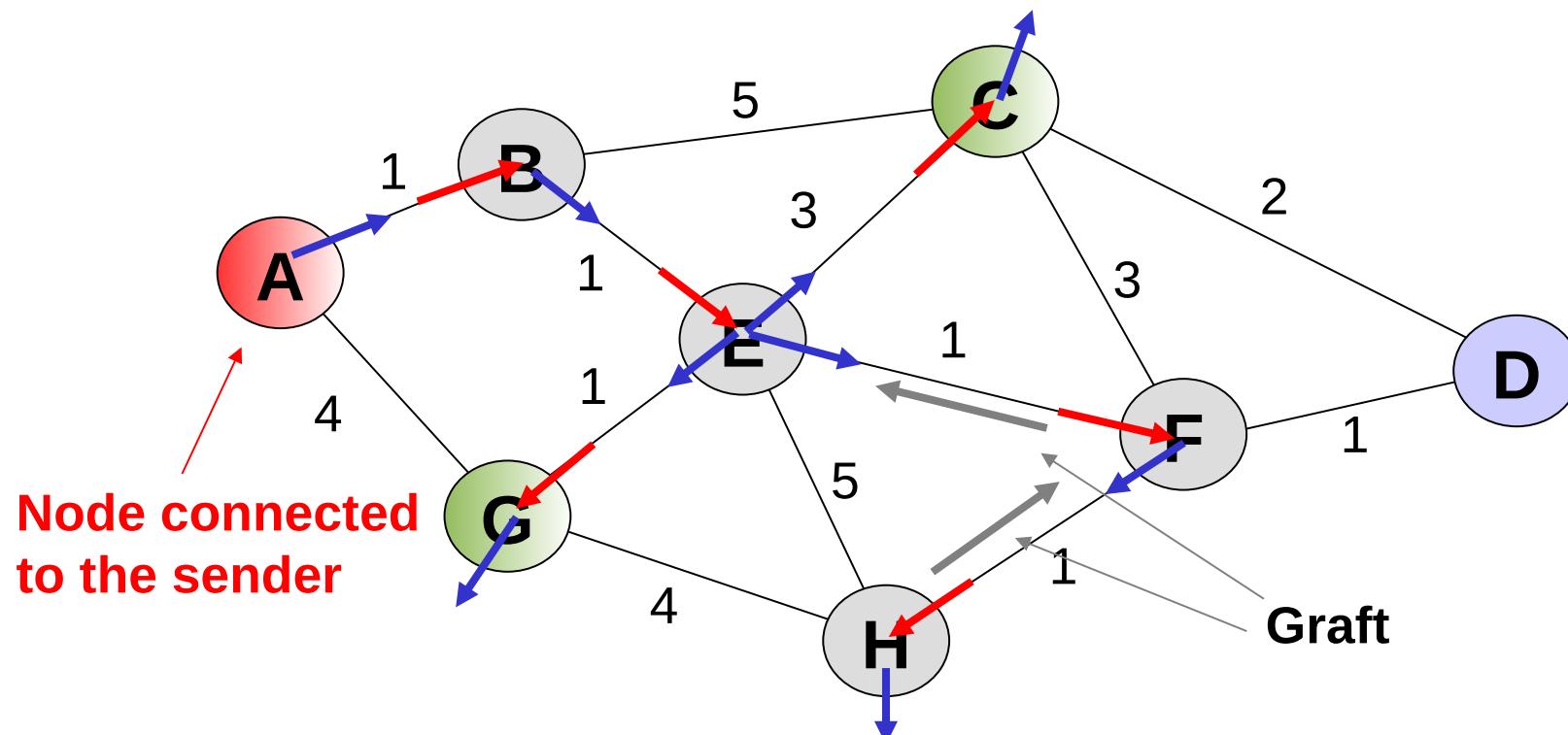
Source-based tree (when the sender is unknown)

- “Reverse Path Forwarding with Pruning” (II)
 - ◆ A node without any terminal stations interested in the session, and without any neighbor nodes to forward the multicast packets to, sends a prune message to the neighbor from which it receives the multicast packets.
 - ◆ Figure: node H sends a prune message to neighbor F and will not receive multicast packets anymore.
 - ◆ Figure: node D has no receivers anymore and sends a prune to F that, in turn, will not have neighbors requiring the forwarding of multicast packets, and also sends a prune to E.



Source-based tree (when the sender is unknown)

- “Reverse Path Forwarding with Pruning” (III): multicast session member nodes appear
 - First strategy: use a graft message to undo the prune.
 - Second strategy: associate a lifetime to the prune; after the lifetime, multicast packets will be forwarded again.



Group-shared tree vs. Source-based tree

- Group-shared tree
 - ◆ Minimizes the state information that needs to be maintained in each router.
 - ◆ Minimizes the number of connections used to support multicast traffic.
 - ◆ Concentrates the traffic congestion problems in a reduced number of nodes.
- Source-based tree
 - ◆ Penalizes the state information in each router, since it needs to know information about each origin.
 - ◆ Distributes the multicast traffic by a larger number of connections.
 - ◆ Less congestion problems.



Distance Vector Multicast Routing Protocol (DVMRP)

- Algorithm of the “source-based tree” type
 - ◆ “Distance Vector” (RFC 1075), similar to RIP.
- Uses a strategy of RPF (reverse path forwarding) with “pruning”
 - ◆ Such as RIP, the distance is given by the number of hops.
 - ◆ Distance vectors represent the distance of each possible origin.
 - ◆ For each possible origin, each router announces to its neighbors when they are the last hop in the path from the origin.
- “prune” messages
 - ◆ Sent with na associated lifetime.
- “graft” messages
 - ◆ To eliminate/recover from a “prune” message.



Distance Vector Multicast Routing Protocol (DVMRP)

- As its name suggests, DVMRP uses a distance-vector routing algorithm.
- Such algorithms require that each router periodically informs its neighbors about its routing table.
- DVMRP routers advertise routes by sending DVMRP report messages.
- For each network path, the receiving router picks the neighbor advertising the lowest cost and adds that entry to its routing table for future advertisement.
- All interfaces are configured with a cost metric and a threshold TTL that limits the scope of the multicast transmission
 - ◆ It is possible to change the metric associated to an interface to promote or demote the preference for some routes.
 - ◆ A multicast router forwards a multicast datagram through an interface if the TTL in its header is larger than the interface threshold TTL.
- Allows the use of tunnels between multicast routers
 - ◆ Tunnels are administratively configured.
 - ◆ Border routers act as neighbors.



Multicast Open Shortest Path First (MOSPF)

- Algorithm of the “source-based tree” type
 - ◆ It is an extension of OSPF (RFC 1584).
 - ◆ Implements a “RPF with pruning” strategy.
- It makes use of the topology knowledge of each router
 - ◆ In this way, each router can locally process the minimum cost tree for each multicast session.
- It does not support tunnels
 - ◆ In the OSPF messages there is a flag that, when set toNull, indicates that the router does not support multicast.
 - ◆ These routers will not belong to the minimum cost tree.
 - ◆ This protocol requires that any 2 multicast routers should have at least one path where all intermediate routers support multicast.
- Each MOSPF router has a local database of multicast groups containing a list of the directly connected group members
 - ◆ The local router forwards the multicast datagrams to the group members based on this information.
 - ◆ The local router will send a group-membership LSA to all other routers in the domain.



Protocol-Independent Multicast (PIM)

- PIM (RFC 2362) addresses two extreme use cases
- PIM, dense mode
 - ◆ When most of the networks contain stations that want to use multicast.
 - ◆ Consequently, the majority of the network routers need to route multicast packets.
- PIM, sparse mode
 - ◆ Stations that want to use multicast are concentrated on a reduced number of networks.
 - ◆ Consequently, the number of routers that need to route multicast packets is small when compared to the total number of routers.

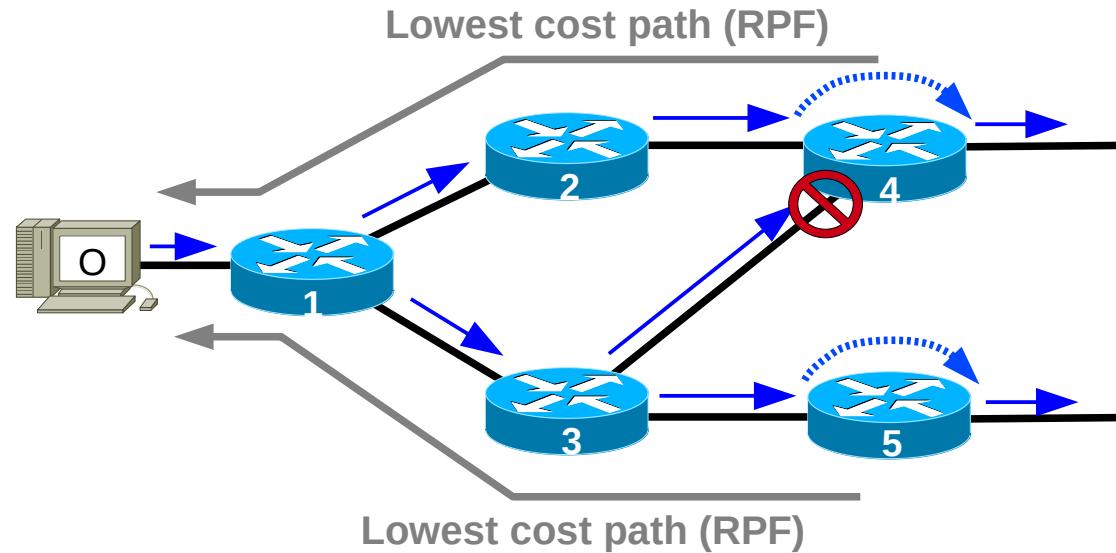


PIM Dense Mode

- Implements a “RPF with pruning” strategy.
- Requires all routers to have the protocol active.
- It is simpler because:
 - ◆ Does not calculate routing tables.
 - ◆ Instead, it uses routing tables built by any other unicast protocol.
 - It is then independent from the routing protocol that is in use.
 - ◆ Assumes that all point-to-point routes are symmetric.
- In case of multiple minimum cost paths, it only accepts the ones corresponding to the highest IP address interface.
- Unicast routing tables do not allow to determine to which neighbors the packets should be forwarded to. So:
 - ◆ By default, routers forward packets to all neighbors that did not send prune messages.
 - ◆ Use prune messages to signal their neighbors that they should not forward packets to them.



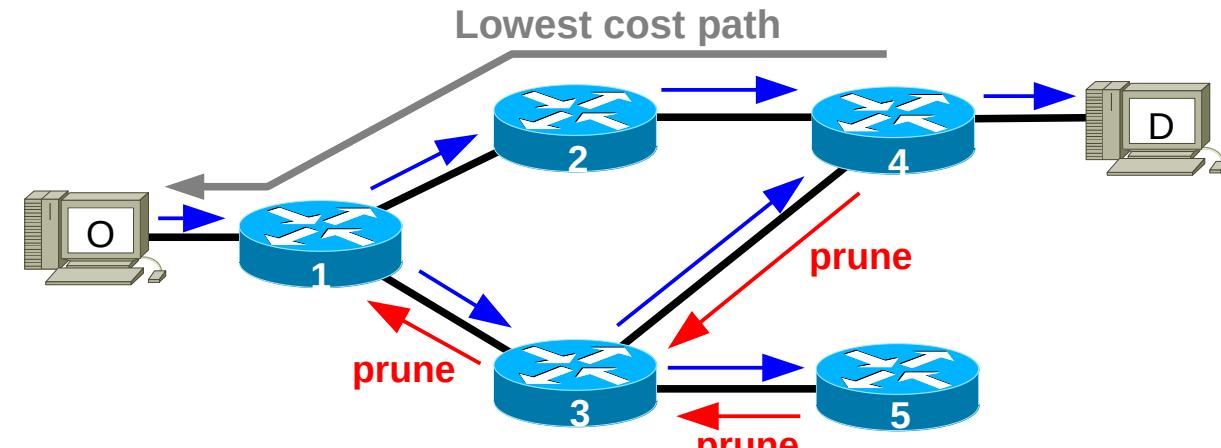
PIM Dense Mode – Initial Flooding



- When a router receives multicast traffic in the interface that provides the lowest cost path to the source (RPF interface), it forwards the traffic to all other interfaces.
- When a router receives multicast traffic in the interface that does not provide the lowest cost path to the source (not the RPF interface), it discards the packets.

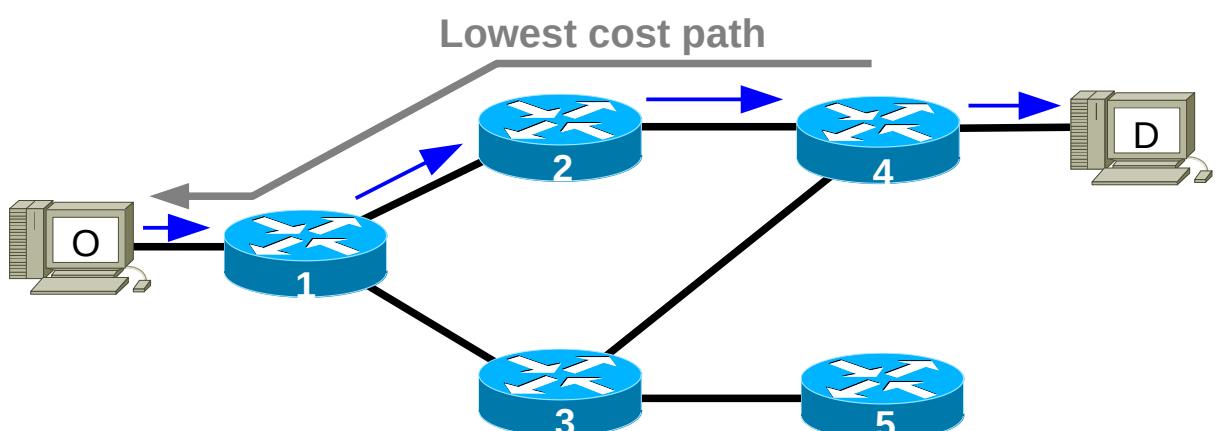


PIM DM – Prune Message

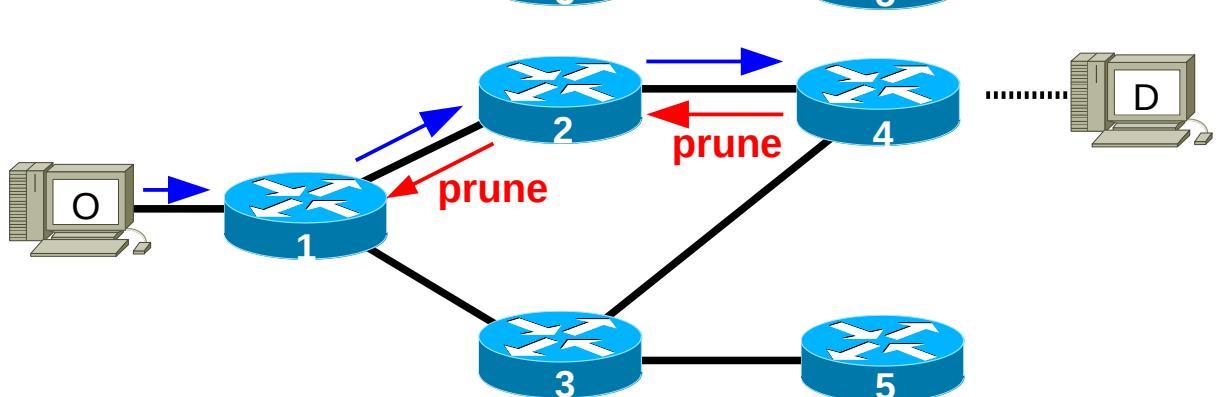


- Routers forward multicast traffic received in one interface to all other interfaces in which they did not receive a *Prune* message.

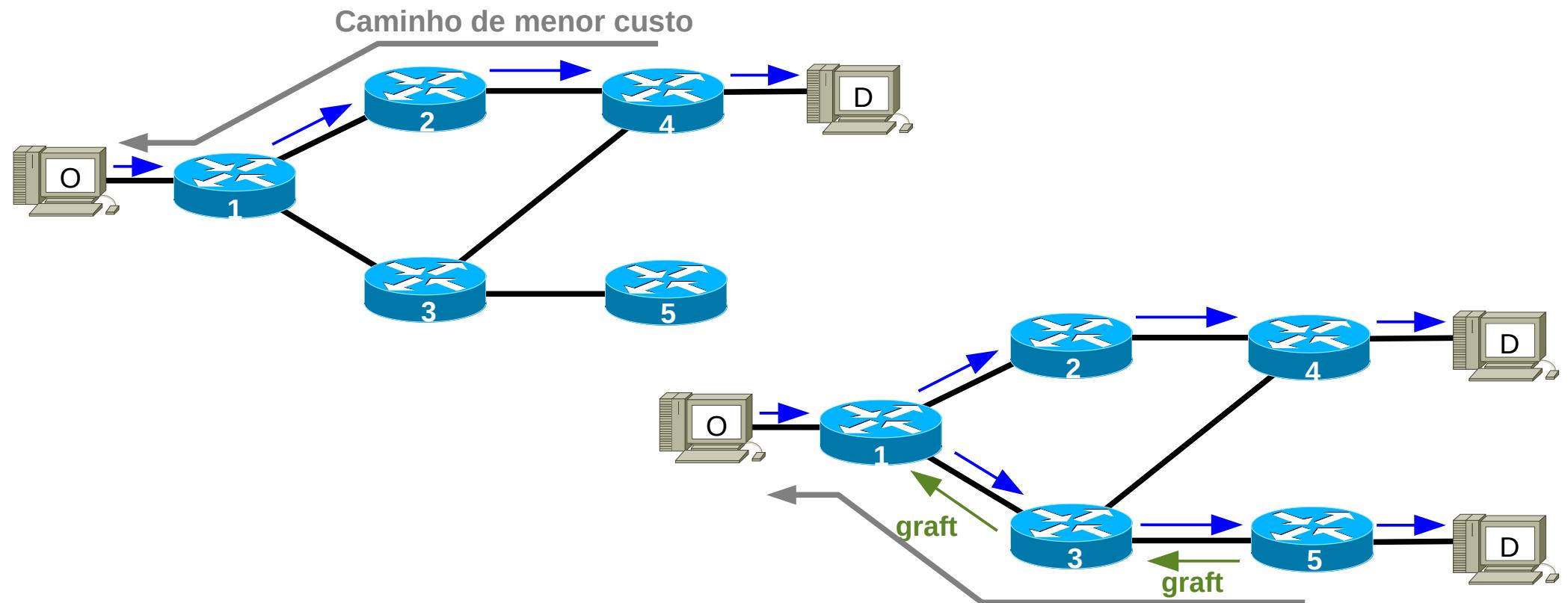
- A *Prune* Message is sent by:
 - ◆ Routers with no clients interested in a specific multicast session (e.g., Router 5).
 - ◆ Routers that received the same multicast traffic in more than one interface (e.g., Router 4)
 - Send the *Prune* message via all interfaces in which the traffic was received, and do not provide the lowest path cost to source.
 - ◆ Routers that received *Prune* messages in all interfaces to which the multicast traffic was forwarded (e.g., Router 3).



- When a router does not have any clients interested in a specific multicast session it sends a *Prune* message via the interface that receives the multicast traffic.



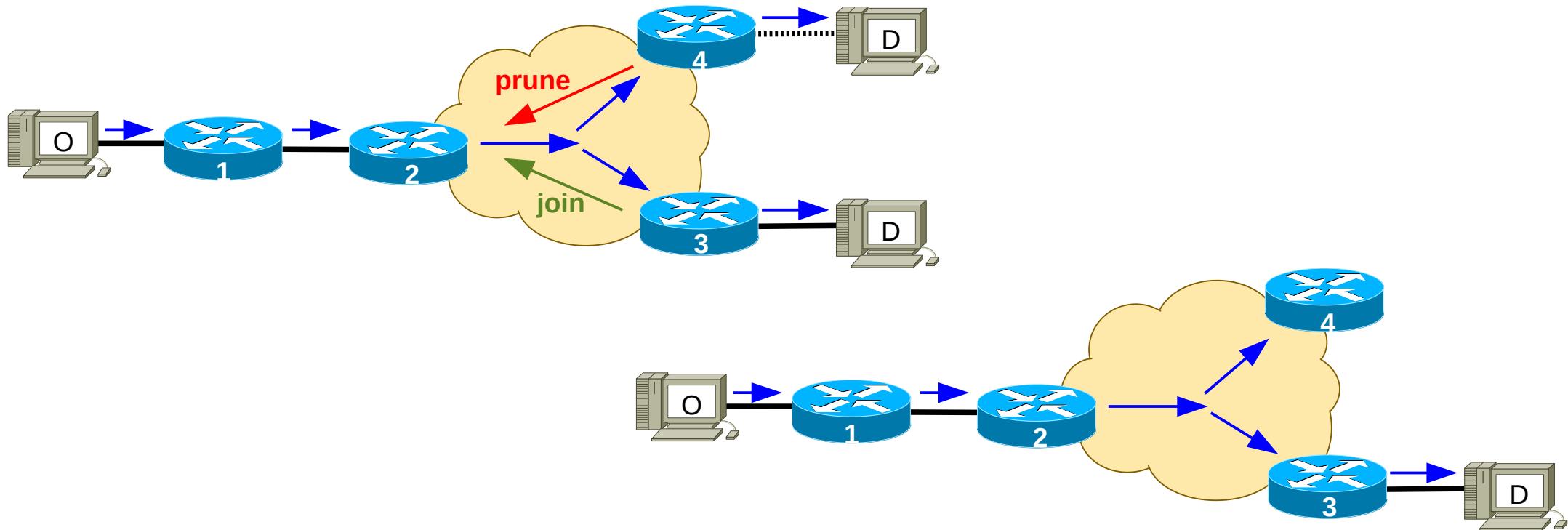
PIM DM – Graft Message



- A router may restart to receive multicast traffic by sending a *Graft* message to cancel a previously sent *Prune* message.



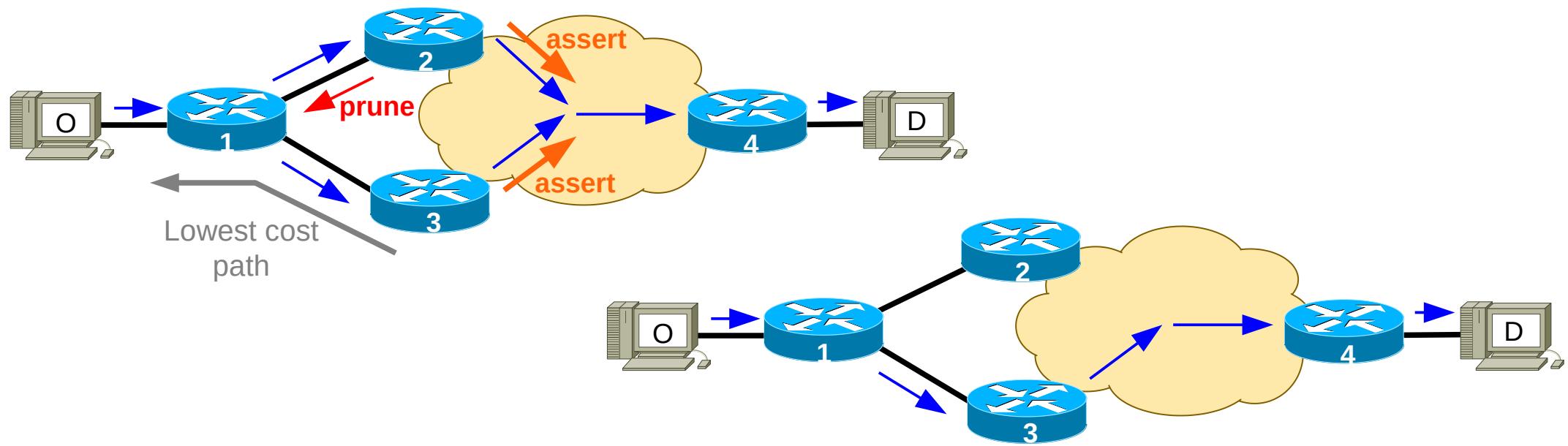
PIM DM – Join Message



- When a router has no more clients for a specific multicast session it send a *Prune* message.
- When the *Prune* message is sent through a shared medium (e.g., LAN), and when other routers have clients for the same multicast sessions, these should send a *Join* message to nullify the sent *Prune* message.



PIM DM – Assert Message



- When there are more than one router sending traffic from a specific multicast session to a shared medium (e.g., LAN), these must decide which one will be responsible for the multicast traffic.
- All routers send an *Assert* message.
 - Message contains the path cost to the multicast source.
- The chosen router is
 - The one that provides the lowest cost path to the source,
 - In case of tie it is chosen the router with the largest IP address (in interfaces connected to the shared medium).



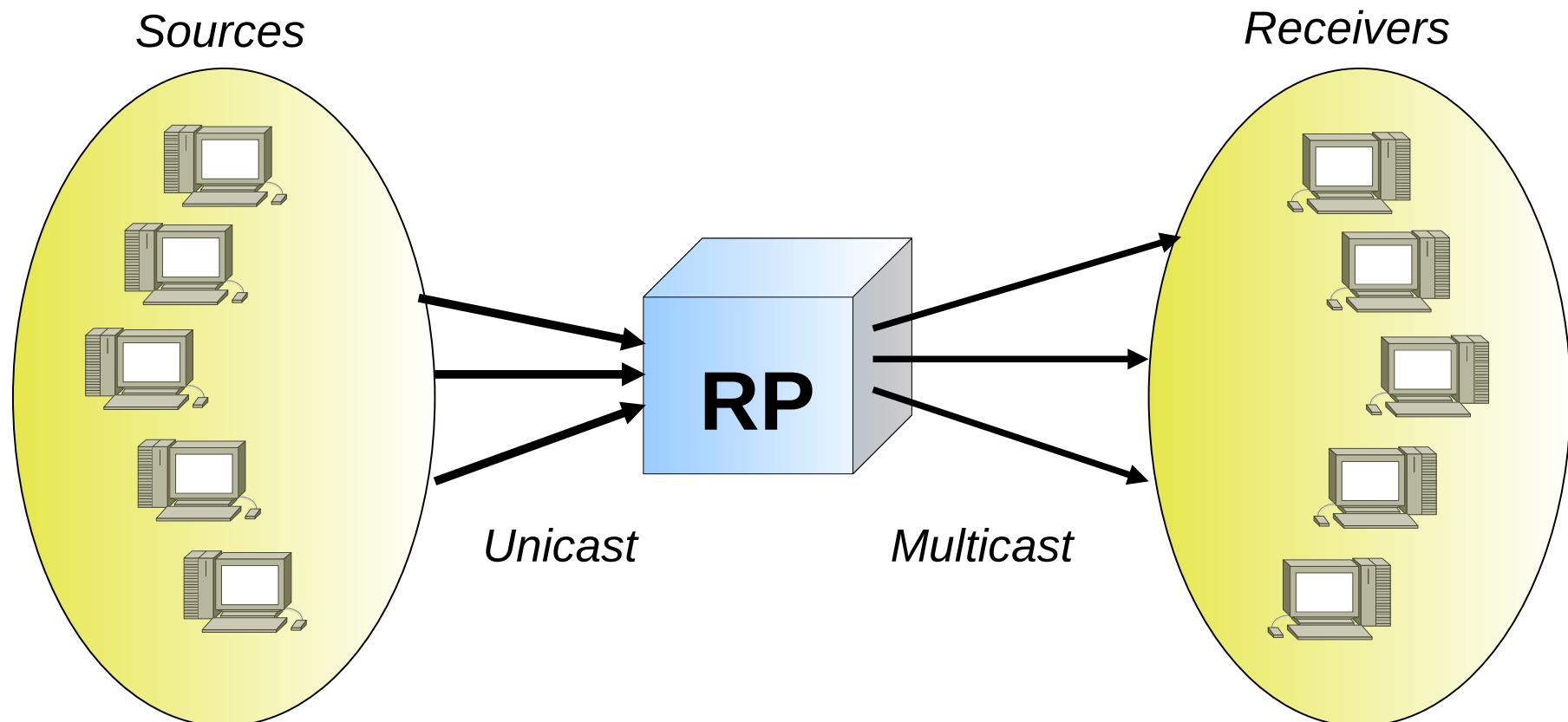
PIM Sparse Mode

- PIM Dense Mode is a data-driven protocol.
 - ◆ Requires that routers that do not have multicast clients must periodically send prune messages to avoid receiving multicast packets.
- PIM Sparse Mode is a receiver-driven protocol.
 - ◆ Each router announces explicitly (with join messages) that it wants to join specific multicast sessions.
- PIM Sparse Mode initially uses a “Group-Shared tree” strategy based on a Rendezvous Point (RP) router.
 - ◆ RP can be administratively configured.
 - ◆ There are also automatic mechanisms, such as CISCO RP Discovery Protocol, that uses the multicast address 224.0.1.40.
 - ◆ There may be different RPs for different multicast sessions.



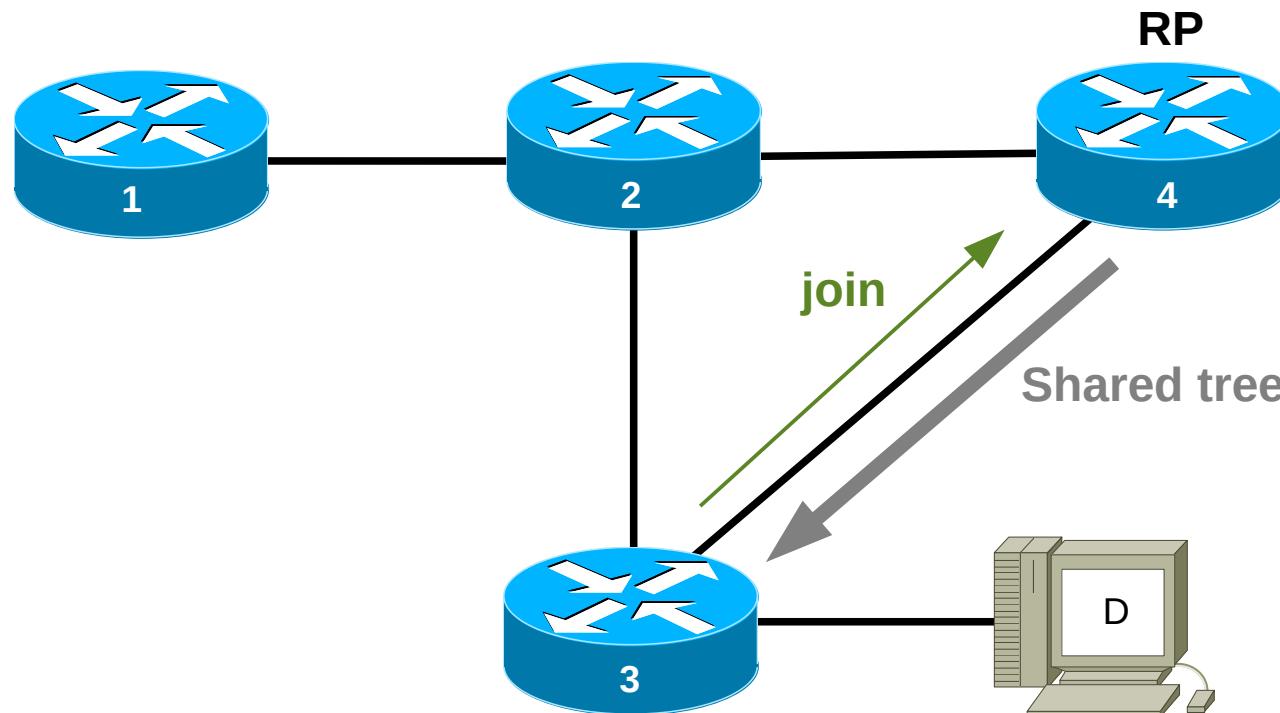
PIM Sparse Mode

- At start, multicast packets from a specific session are sent to the RP using unicast (Multicast over PIM tunnel).
- At the RP, the multicast packets are resent to all interested clients.

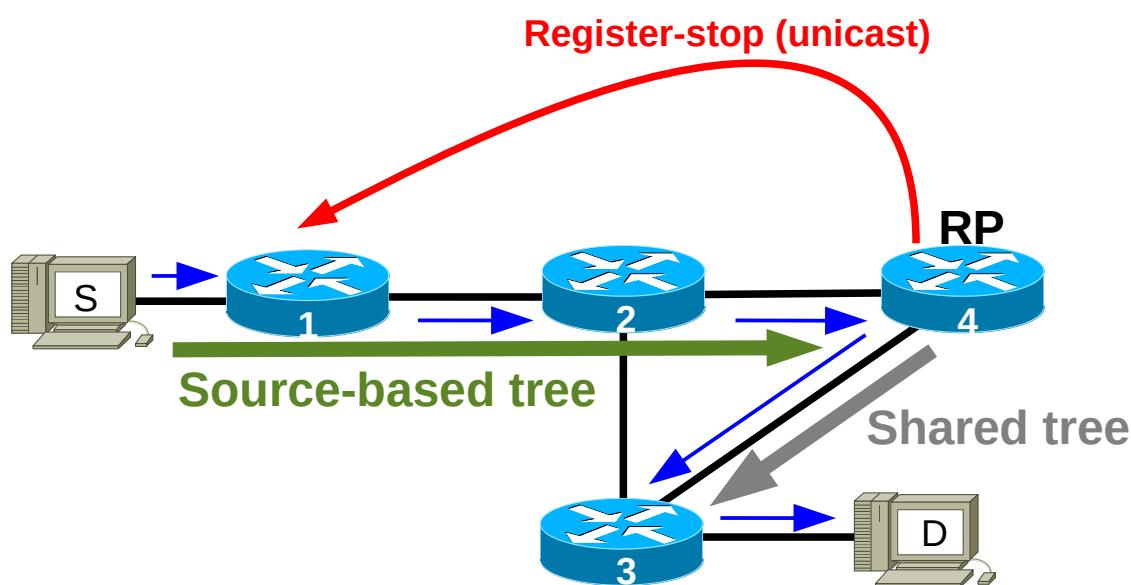
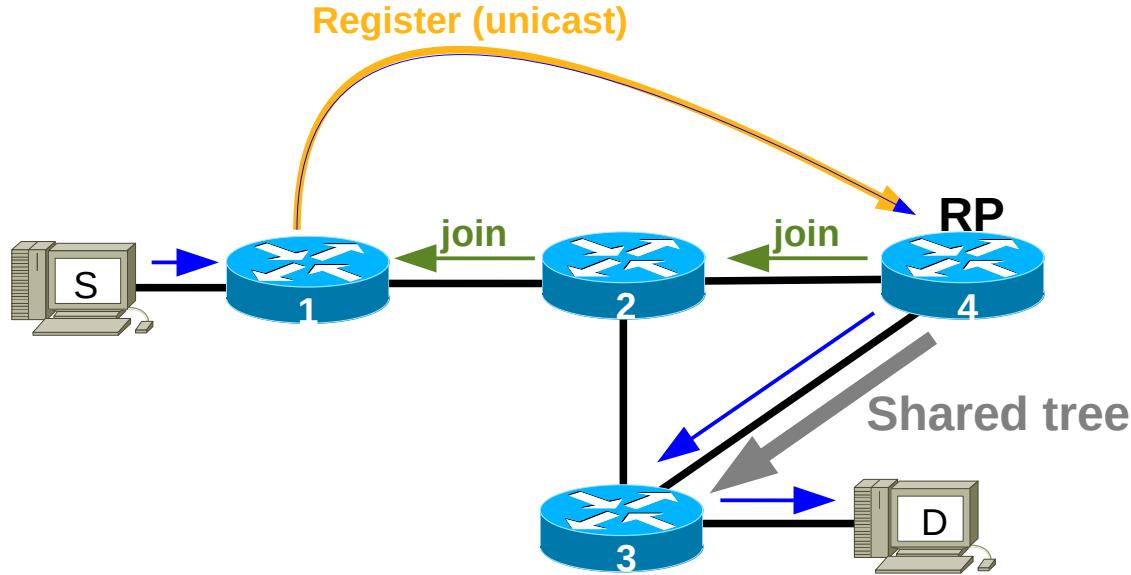


PIM SM – Join to *Group-Shared Tree*

- The routers with clients interested in a specific multicast session, send a *Join* message to the RP in order to join the group-shared tree (which root is the RP).



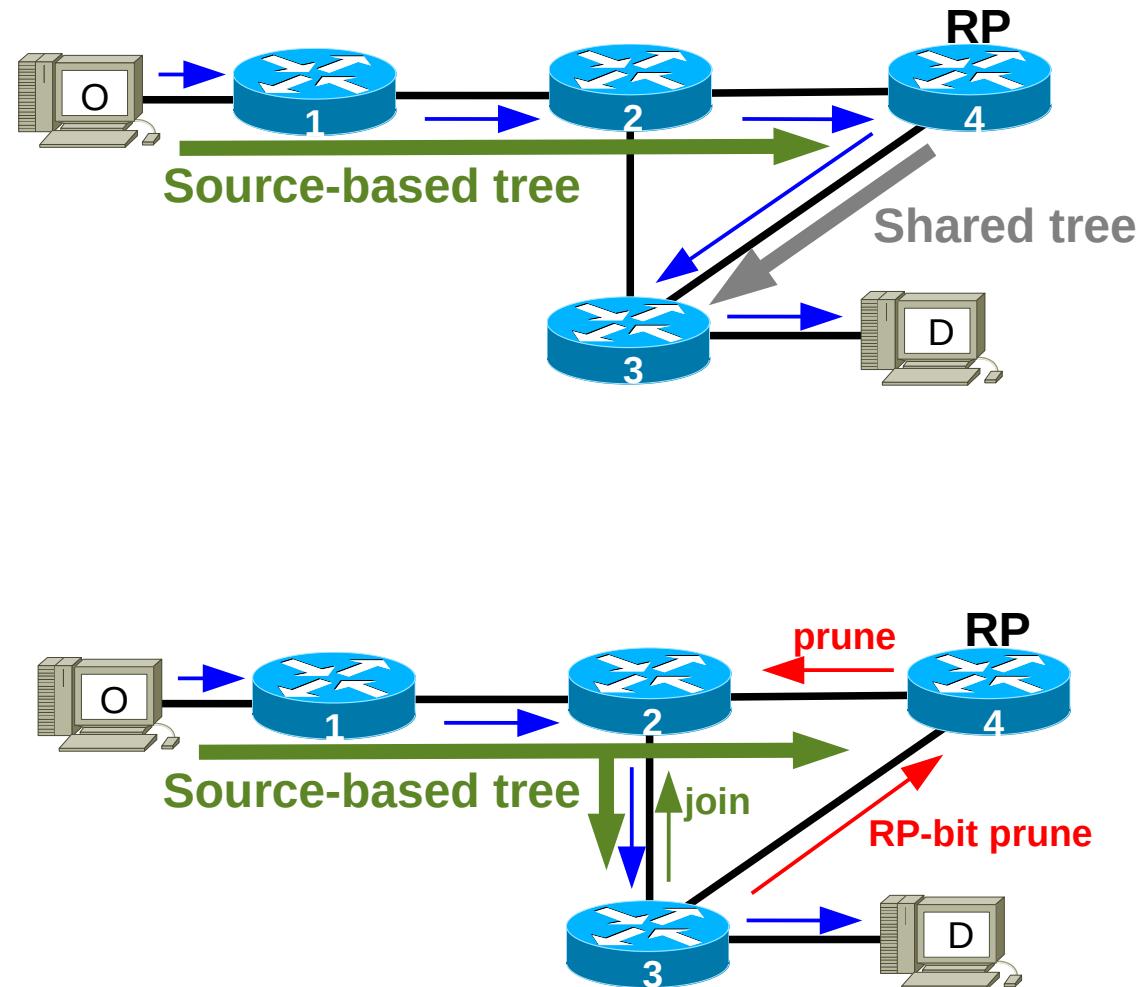
PIM SM – New Source



- When a new multicast source appears, the router that provides connection to the source, re-sends the multicast packets encapsulated in PIM packets (*PIM Register* message), and sends them in unicast to the RP.
- The RP desencapsulates the *PIM Register* messages, receives the multicast packets, and simultaneously:
 - Re-sends the multicast packets to the clients, using the Group-shared tree.
 - Just after receiving the first *PIM Register*, sends a *Join* message to the source router to create a source-based tree.
- As soon the source router starts to route the multicast packets over the source-based tree, and the RP receives them, the RP send a *PIM Register-Stop* message to notify the source router to stop sending the encapsulated multicast packets.



PIM SM – Commuting to Source-based Tree



- When the aggregated bit rate (of all sources) exceeds a predefined threshold, a router may choose to join the Source-based tree, leaving the Group-shared tree.
 - By sending a *Join* message towards the multicast source, until it finds a router that belongs to that source's Source-based tree.
- When a multicast packet arrives through the Source-based tree, the router send a *RP-bit Prune* message towards the RP, to leave the Group-shared tree.
 - Later any router may re-join the Group-shared tree.



Multicast em IPv6

- IGMP is replaced by Multicast Listener Discovery (MLD) protocol.
 - ◆ MLD is equivalent to IPv4 IGMP,
 - ◆ MLD was adapted to IPv6 semantics:
 - ◆ MLDv1 <=> IGMPv2,
 - ◆ MLDv2 <=> IGMPv3.
 - ◆ MLD messages are transported over ICMPv6.
 - ◆ Multicast Listener Query, Multicast Listener Report and Multicast Listener Done.
 - ◆ MLD uses link local addresses as sources.
- Multicast routing tree may be *Sparse or Source-Specific*.
 - ◆ Dense-Mode is not available in IPv6!
 - ◆ Source-Specific requires MLDv2.
 - ◆ PIM protocols can also be used with IPv6 addresses.
 - ◆ PIM-SM and PIM-SSM, only.



MLD Messages

- MLD messages are used to determine group membership on a network segment, also known as a link or subnet.
- Multicast Listener Query
 - ◆ Sent by a multicast router to poll a network segment for group members. Queries can be general, requesting group membership for all groups, or can request group membership for a specific group.
- Multicast Listener Report
 - ◆ Sent by a host when it joins a multicast group, or in response to an MLD Multicast Listener Query sent by a router.
- Multicast Listener Done
 - ◆ Sent by a host when it leaves a host group and is the last member of that group on the network segment.



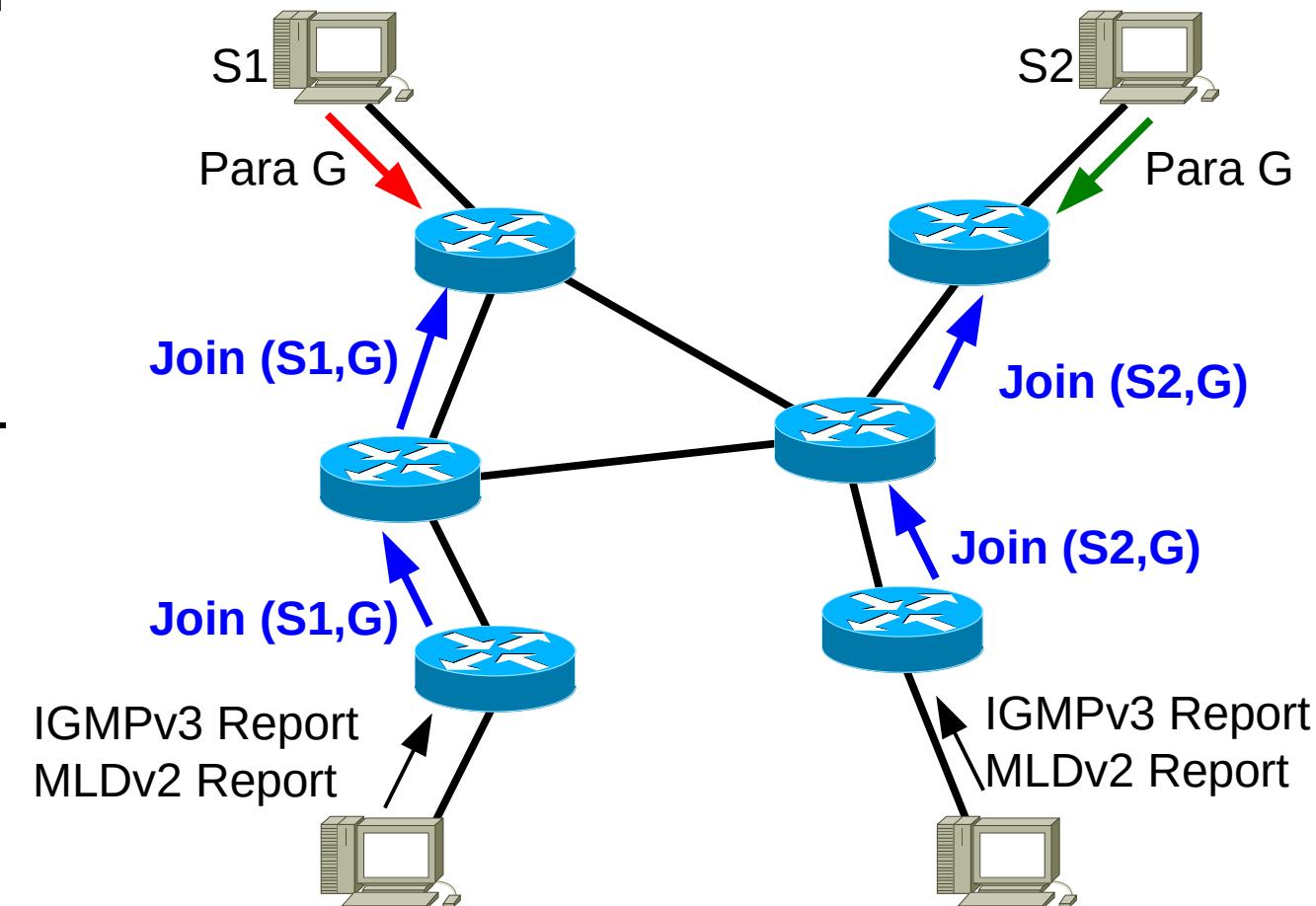
Source-Specific Multicast (SSM)

- SSM is an extension of the Multicast routing model.
 - ◆ Routing is uniquely made using Source-based trees.
 - ◆ The deployment of SSM with PIM (PIM-SSM) uses a sub-set of the PIM-SM mechanisms.
- Receivers specify not only the multicast session/group, but also a specific source.
 - ◆ The receiver router establishes a Source-based tree by sending a *Join* message to the specific source.
- Reserved SSM addresses:
 - ◆ IPv4 range - 232.0.0.0/8.
 - ◆ IPv6 range - FF3x::/32.
- Requires IGMPv3 or MLDv2.
 - ◆ Because client must specify the desired source.
 - ✚ IGMPv3 new INCLUDE mode functionality.



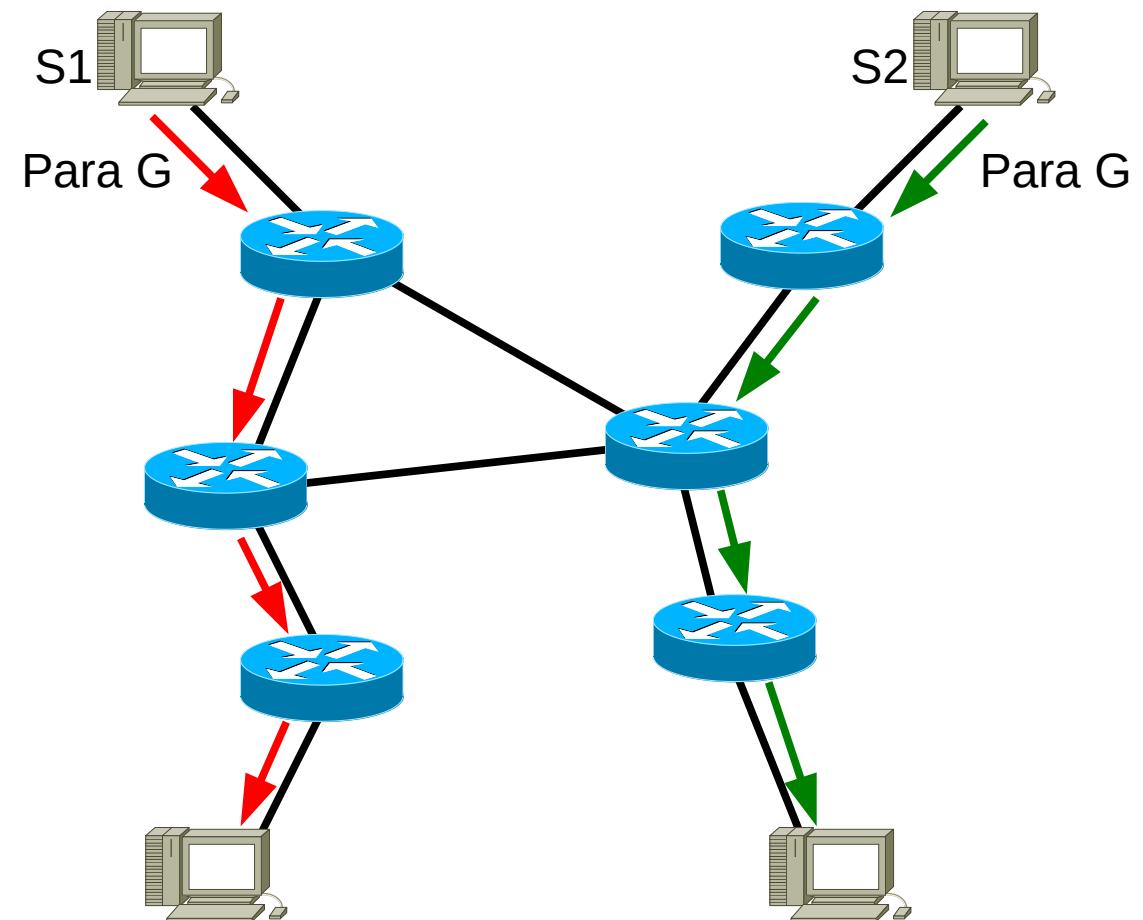
PIM-SSM Example (1)

- Reports are sent to join multicast group G from sources S1 or S2.
- Routers send PIM *Join* messages to construct the two distinct Source-based trees to group G from S1, and group G from S2.



PIM-SSM Example (2)

- After the Source-based trees creations, routers forward the multicast traffic to terminals using the respective source tree.



Traffic Engineering (TE) & Multiprotocol Label Switching (MPLS)

Redes de Comunicações II

**Licenciatura em
Engenharia de Computadores e Informática
DETI-UA**



universidade de aveiro

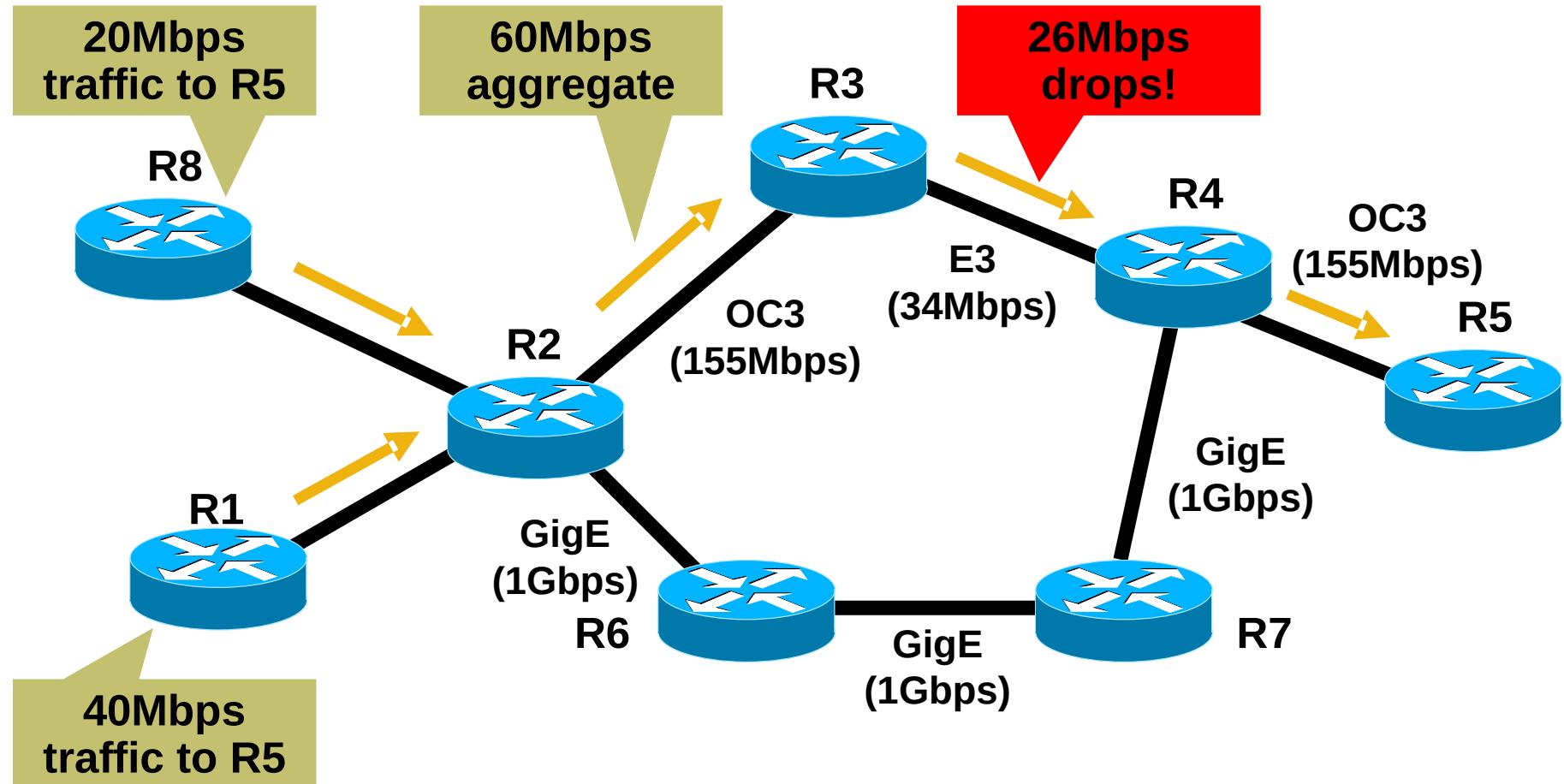
deti.ua.pt

Traffic Engineering (TE)

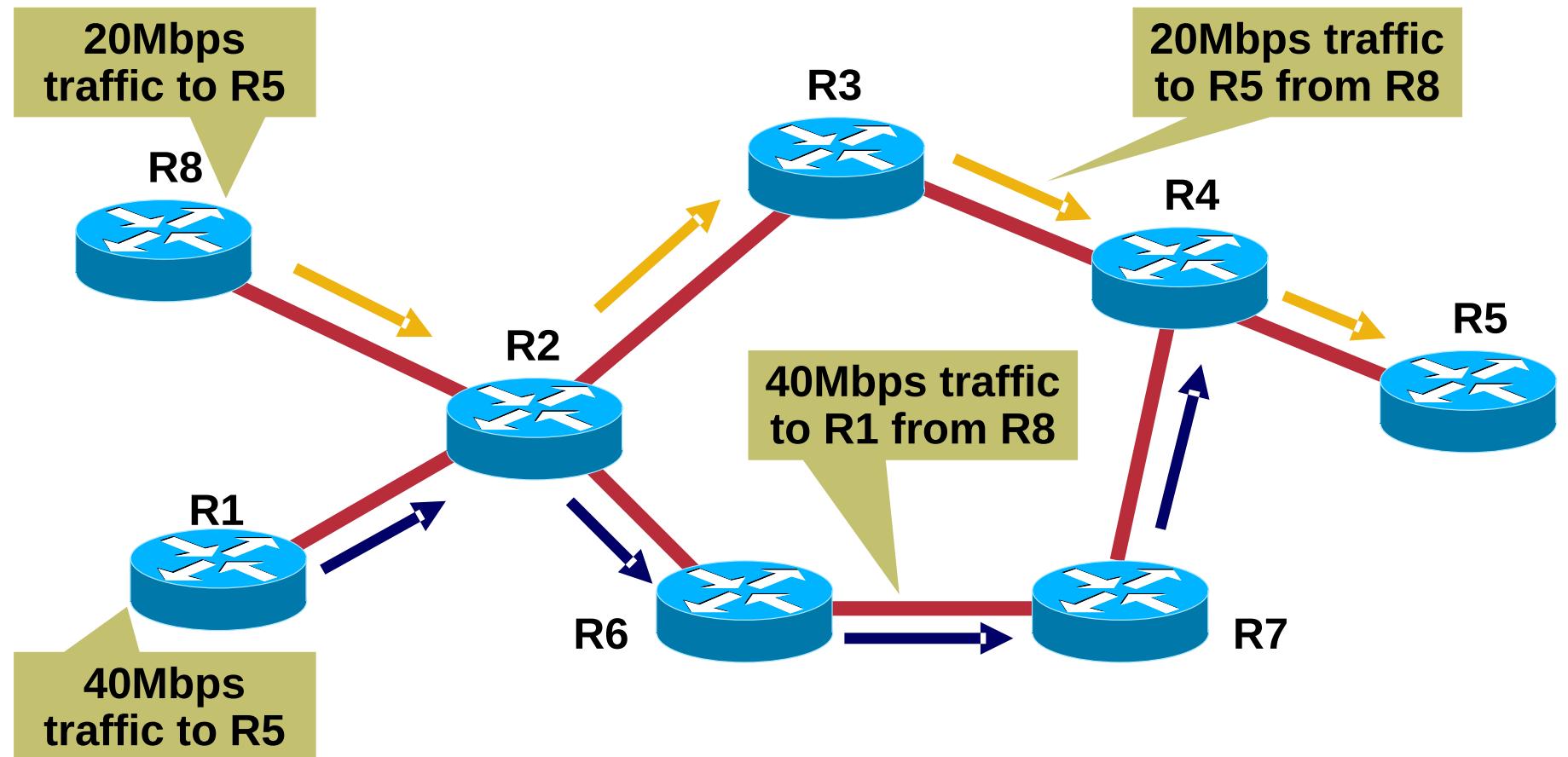
- Network Engineering
 - ◆ Build your network to carry your predicted traffic!
 - ◆ Traffic patterns are impossible to predict!
 - ◆ Routing is based on the destination and does not allow to take the maximum possible advantage of the network resources.
 - ◆ IP source routing (using options field of IP header) is not usable in practice due to security reasons.
- Traffic Engineering
 - ◆ Manipulate your traffic path to fit your network!
 - Can be done with routing protocol costs (difficult deployment), or MPLS.
 - With RIP or OSPF or ANY OTHER IGP it is not possible to condition multiple traffic flows.
 - ◆ Increase efficiency of bandwidth resources.
 - Prevent over-utilized (congested) links whilst other links are under-utilized.
 - ◆ Ensure the most desirable/appropriate path for some/all traffic.
 - Override the shortest path selected by the routing protocols.



Shortest Path and Congestion



A TE Solution



Tunnels are **UNI-DIRECTIONAL**

Normal path: R8 > R2 > R3 > R4 > R5

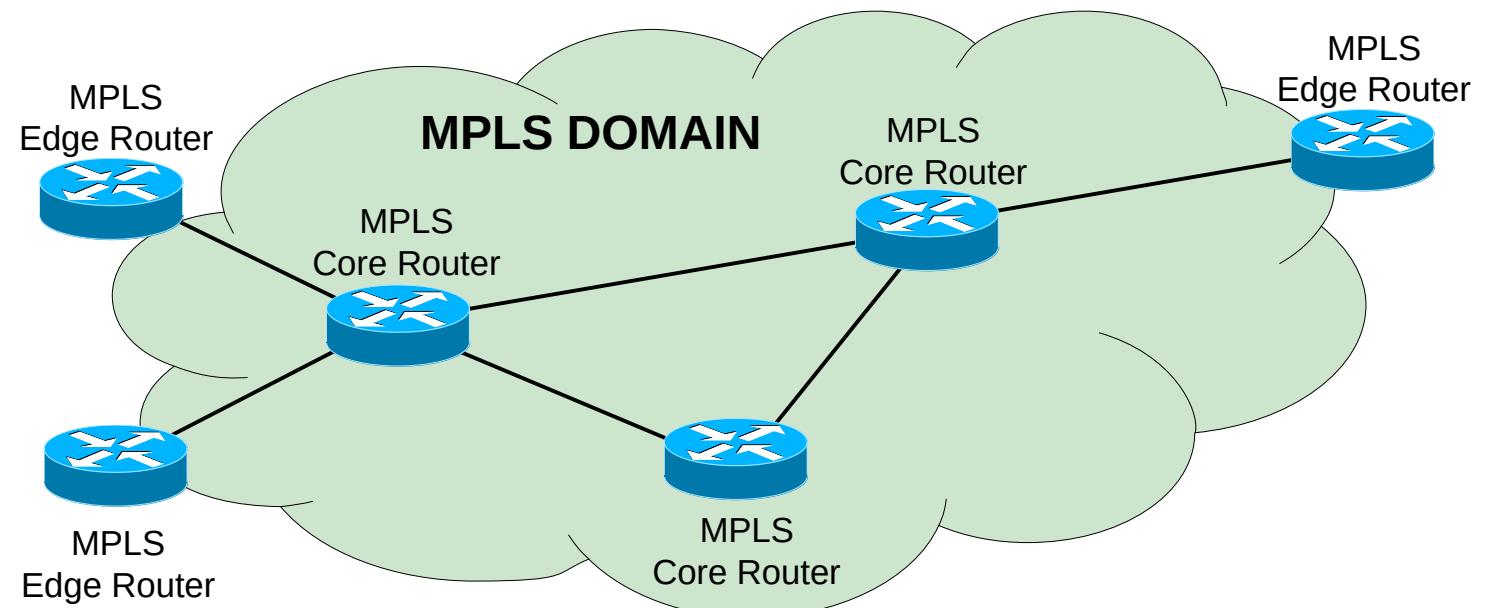
Tunnel path: R1 > R2 > R6 > R7 > R4



Multiprotocol Label Switching (MPLS)

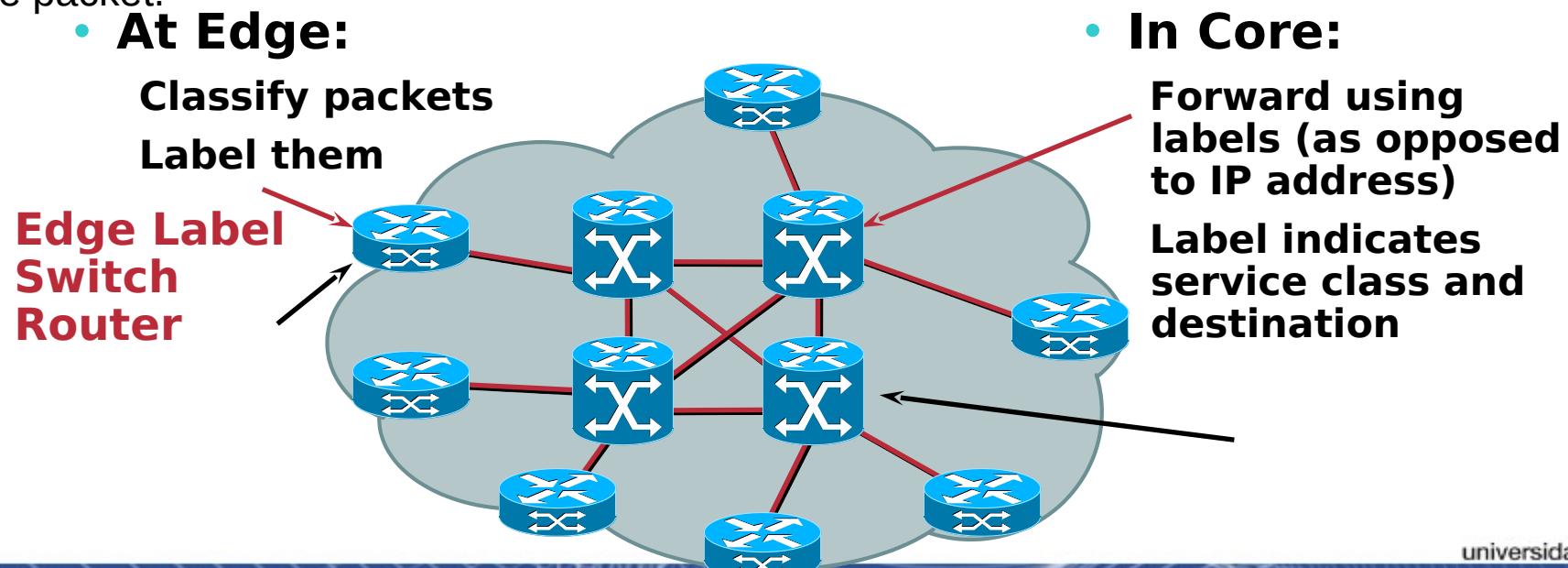
- Packets are labeled at the source with the label of the first hop.
- As a packet travels from one router to the next, each router makes an independent forwarding decision for that packet based on a label.
- Advantages

- ◆ Simplification of the packet routing process on routers.
- ◆ Traffic engineering capability.
- ◆ Simplification of the network management (a single protocol layer).

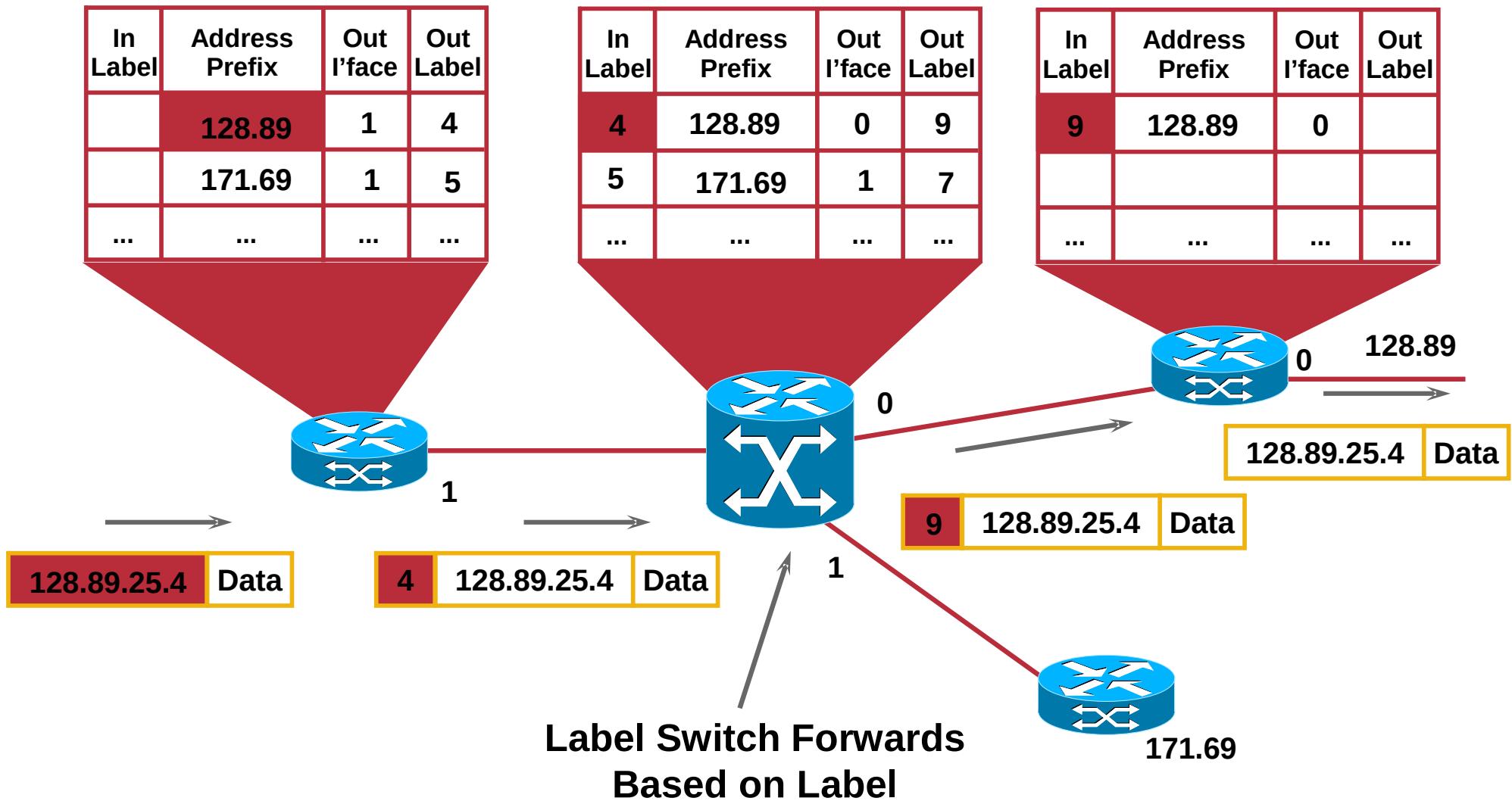


MPLS Fundamentals

- Based on the label-swapping and forwarding paradigm.
- As a packet enters an MPLS network, it is assigned a label based on its **Forwarding Equivalence Class (FEC)** as determined at the edge of the MPLS network.
- FECs are groups of packets forwarded over the same **Label Switched Path (LSP)** by **Label Switching Routers (LSR)**.
- Need a mechanism that will create and distribute labels to establish LSP paths.
- Separated into two planes:
 - ◆ Control Plane - Responsible for maintaining correct label tables among Label Switching Routers.
 - ◆ Forwarding Plane - Uses label carried by packet and label table maintained by LSR to forward the packet.

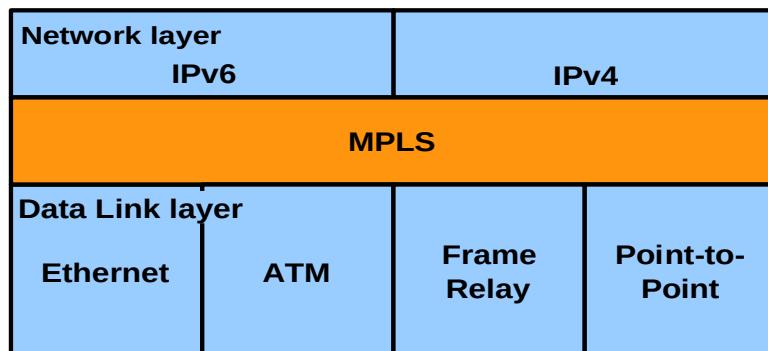


MPLS Switching

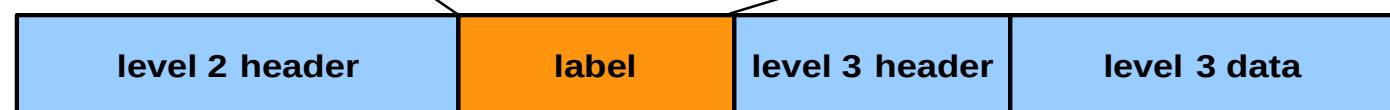


MPLS Labels

- On some Data Link (level 2) technologies, label is given by the appropriate fields of their header.
 - ◆ ATM technology : VPI (Virtual Path ID) and VCI (Virtual Channel ID) fields.
 - ◆ Frame Relay technology: DLCI (Data Link Connection Identifier) field.
- On other Data Link technologies (Point-to-Point, Ethernet), the label is inserted between layer 2 and layer 3 headers.
- Label is a 20-bit field that carries the actual value of the Label.
- TTL field is IP independent – Similar purpose.

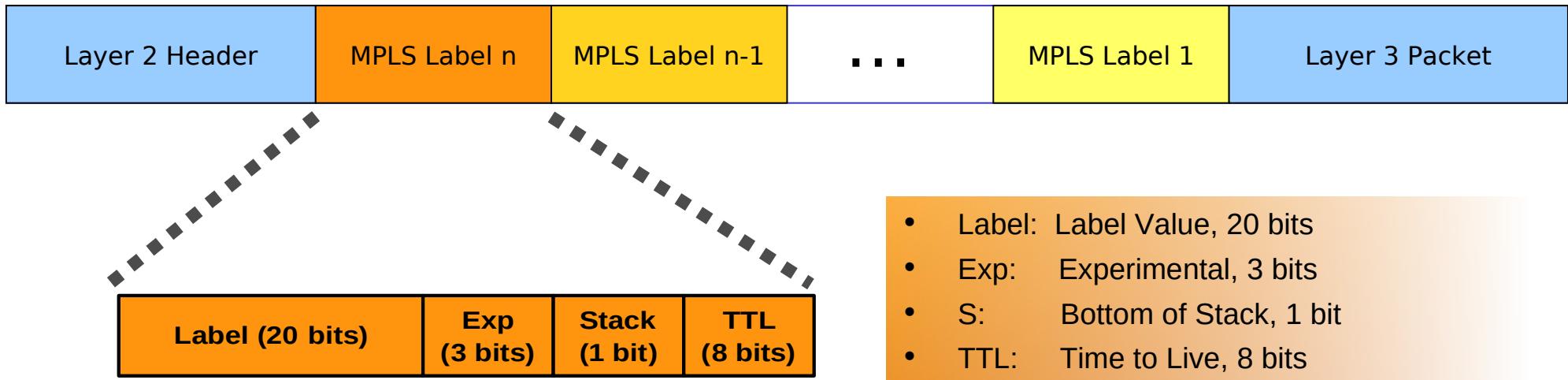


- Label: Label Value, 20 bits
- Exp: Experimental, 3 bits
- S: Bottom of Stack, 1 bit
- TTL: Time to Live, 8 bits



MPLS Label Stacking

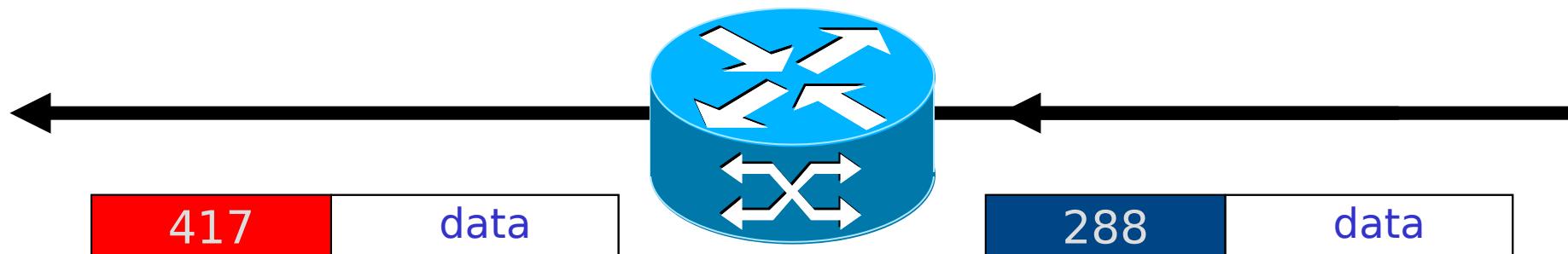
RFC 3032: MPLS Label Stack Encoding



- Labels are arranged in a stack to support multiple services:
 - ◆ Inner labels are used to designate services, FECs, etc.
 - ◆ Outer label is used to switch the packets in MPLS core.
- Bottom of Stack (S) bit is set to one for the last entry in the label stack (i.e., for the bottom of the stack), and zero for all other labels.



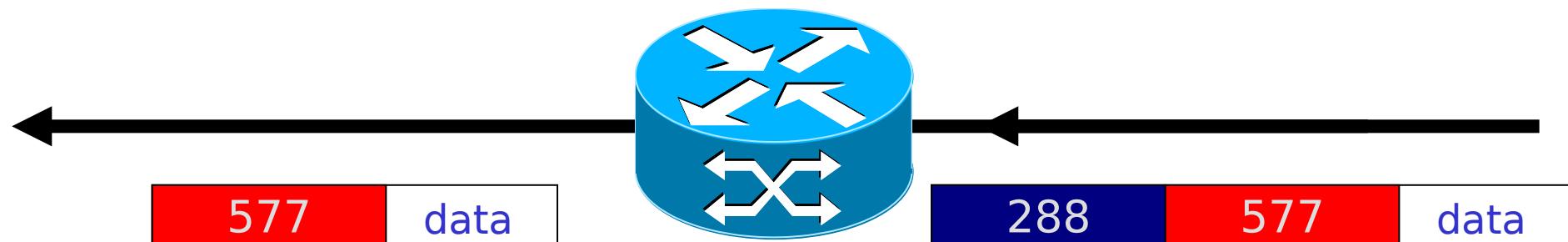
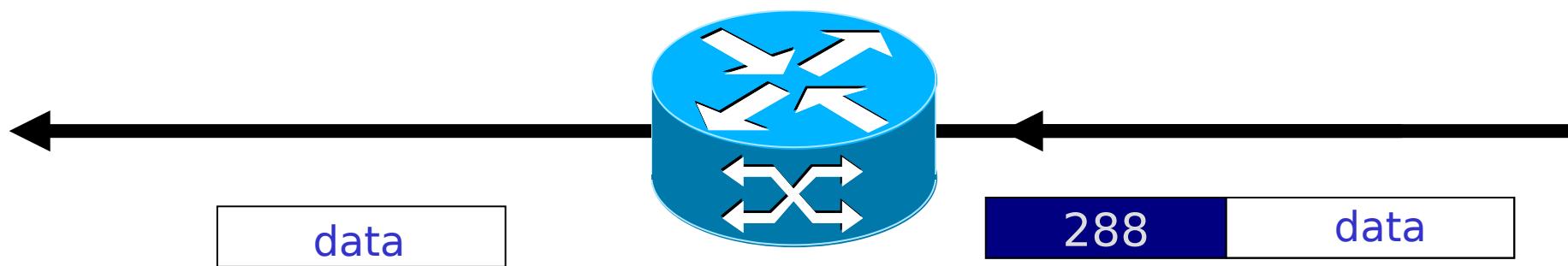
Forwarding via Label Swapping



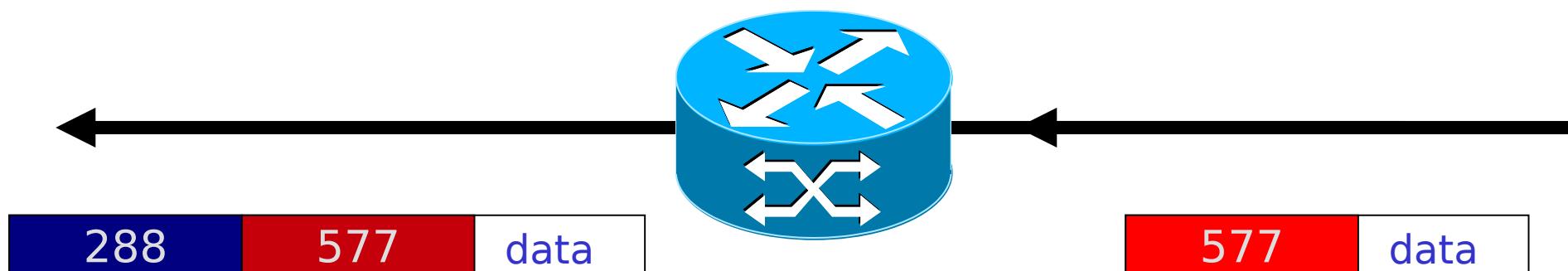
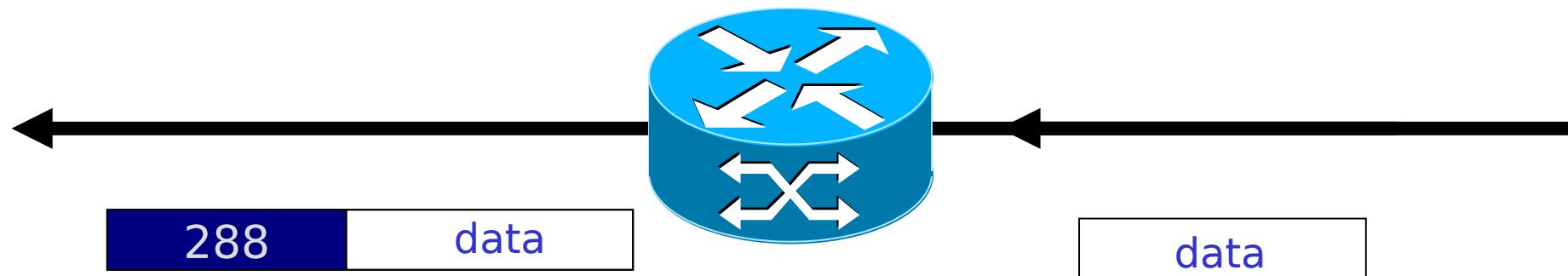
Labels are short, fixed-length values.



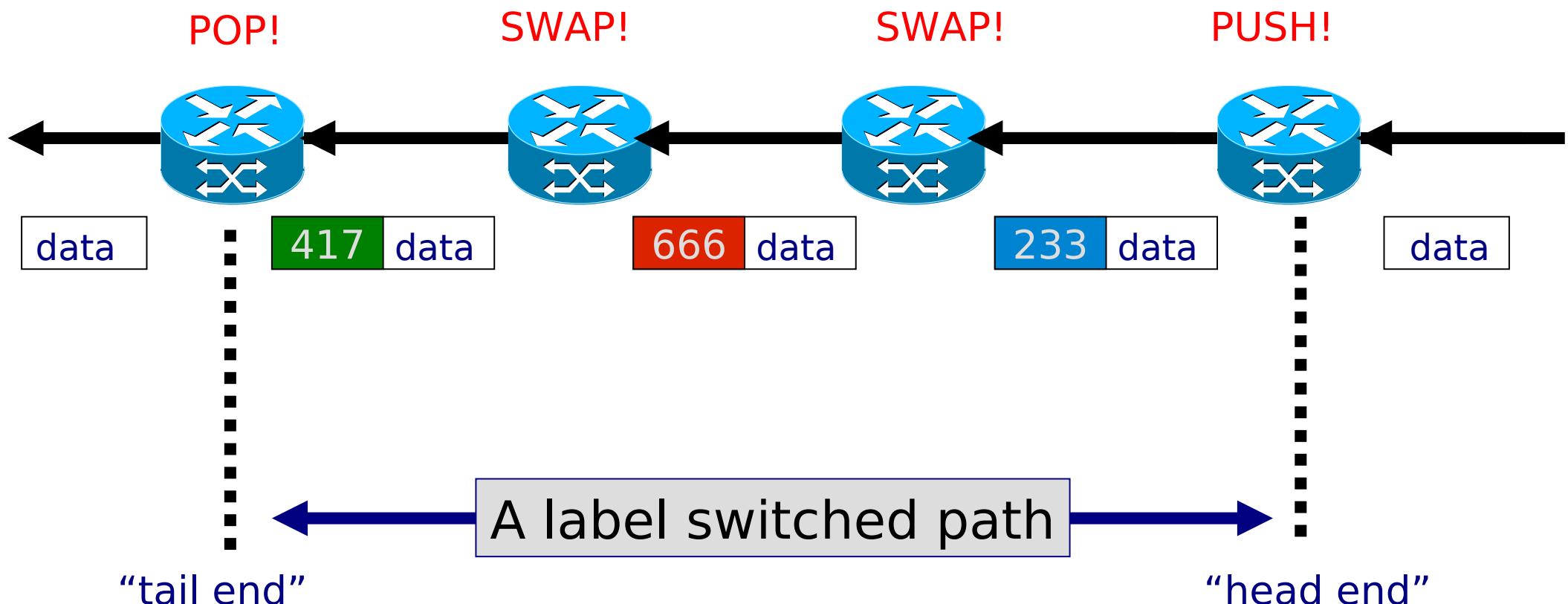
Popping Labels



Pushing Labels



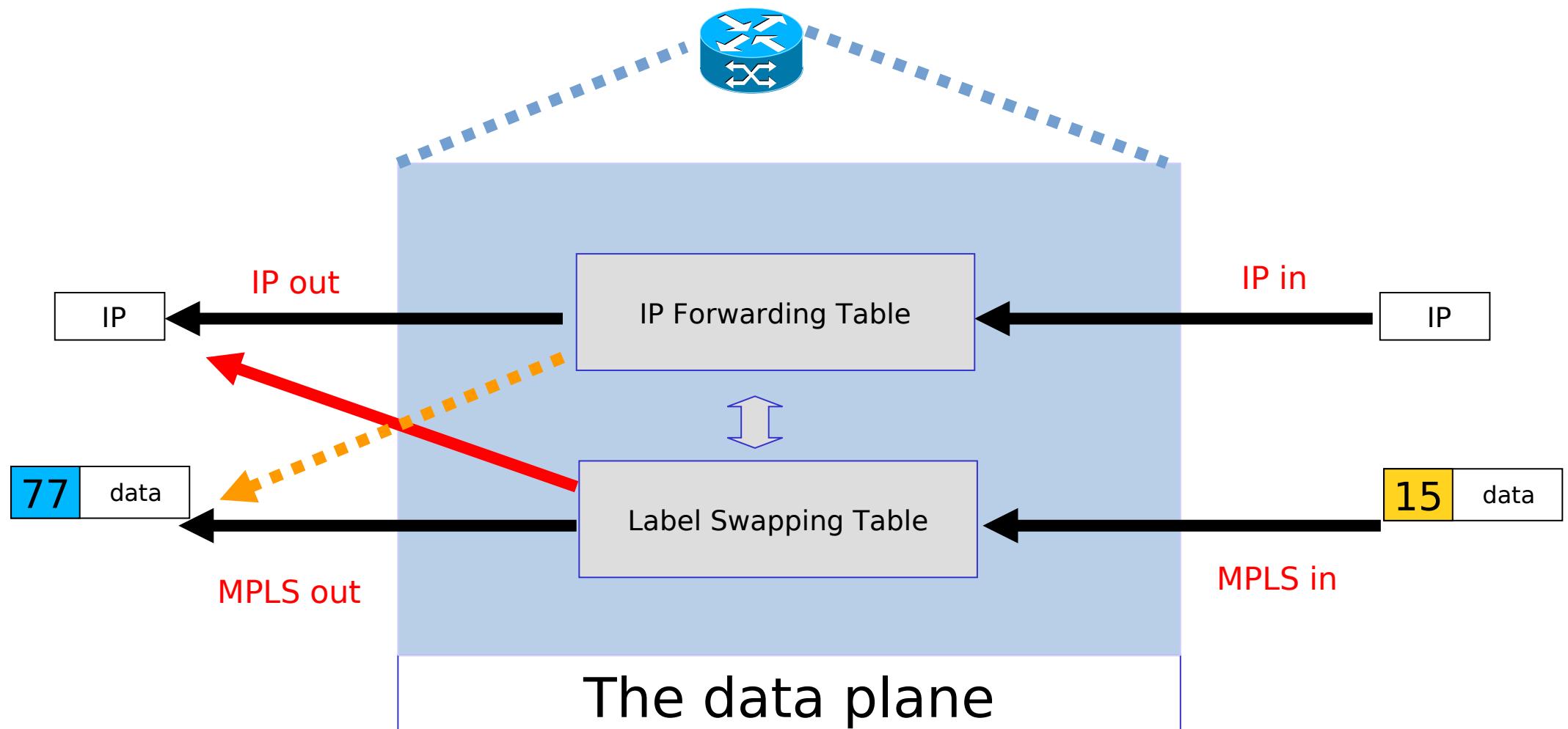
A Label Switched Path (LSP)



Often called an MPLS tunnel: payload headers are not Inspected inside of an LSP. Payload could be MPLS ...



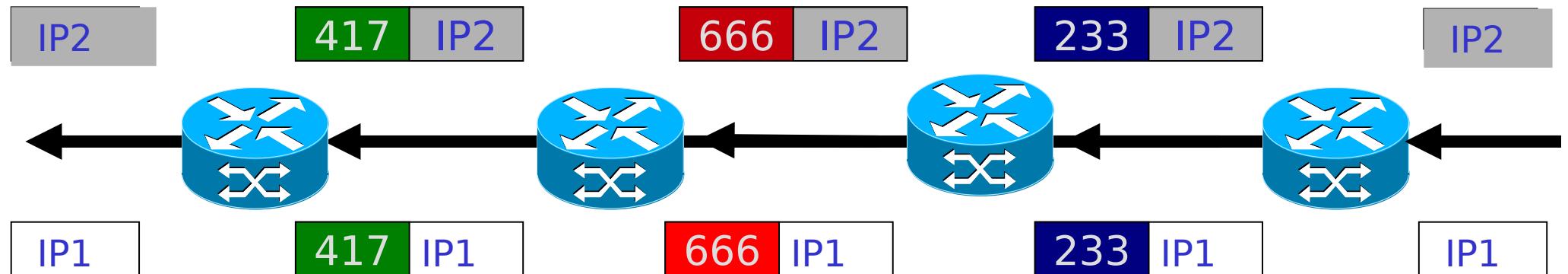
Label Switched Router



IP Lookup + Label PUSH
Label POP + IP lookup



Forwarding Equivalence Class (FEC)

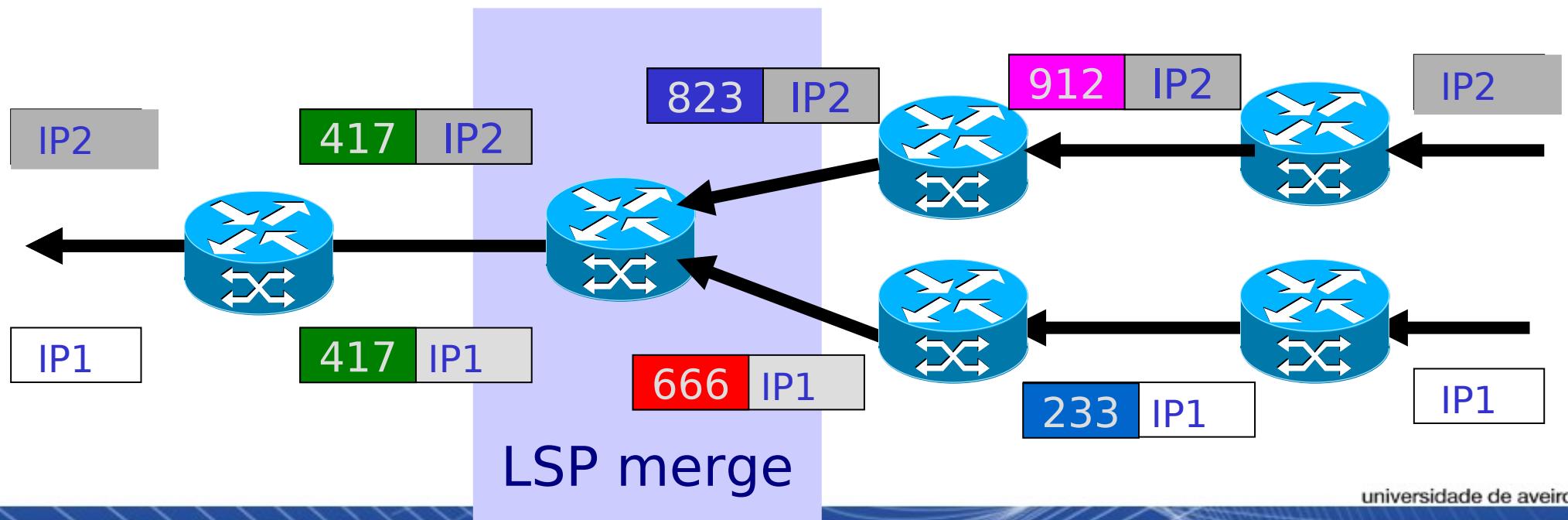
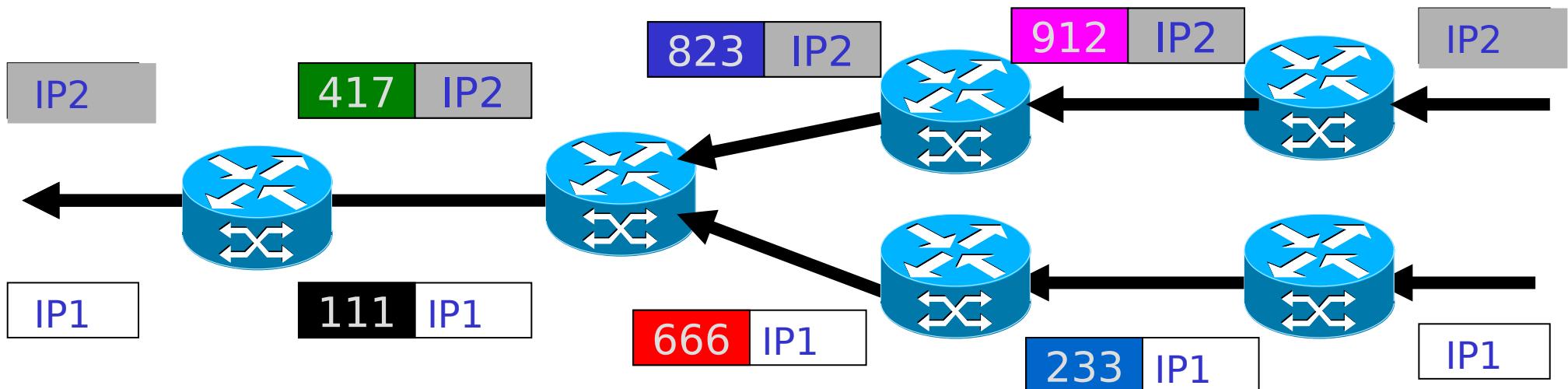


Packets IP1 and IP2 are forwarded in the same way --- they are in the same FEC.

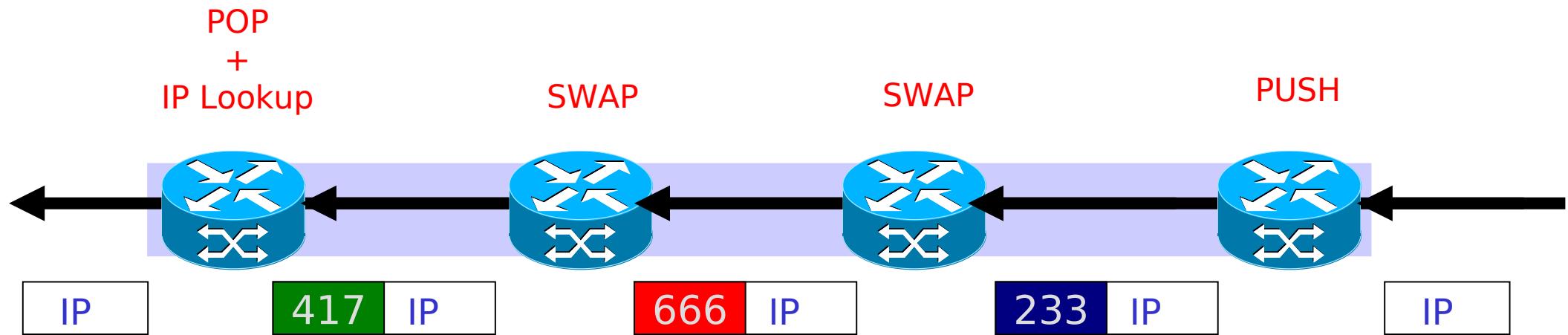
Network layer headers are not inspected inside an MPLS LSP. This means that inside of the tunnel the LSRs do not need full IP forwarding table.



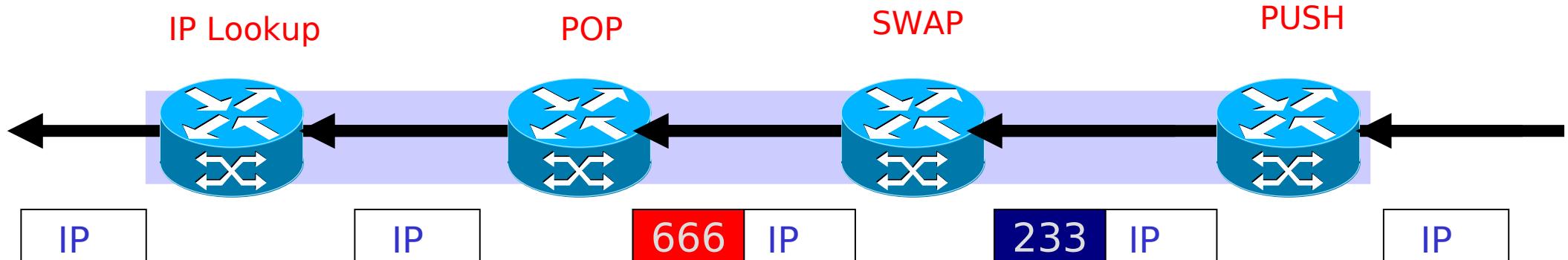
LSP Merge



Penultimate Hop Popping (PHP)



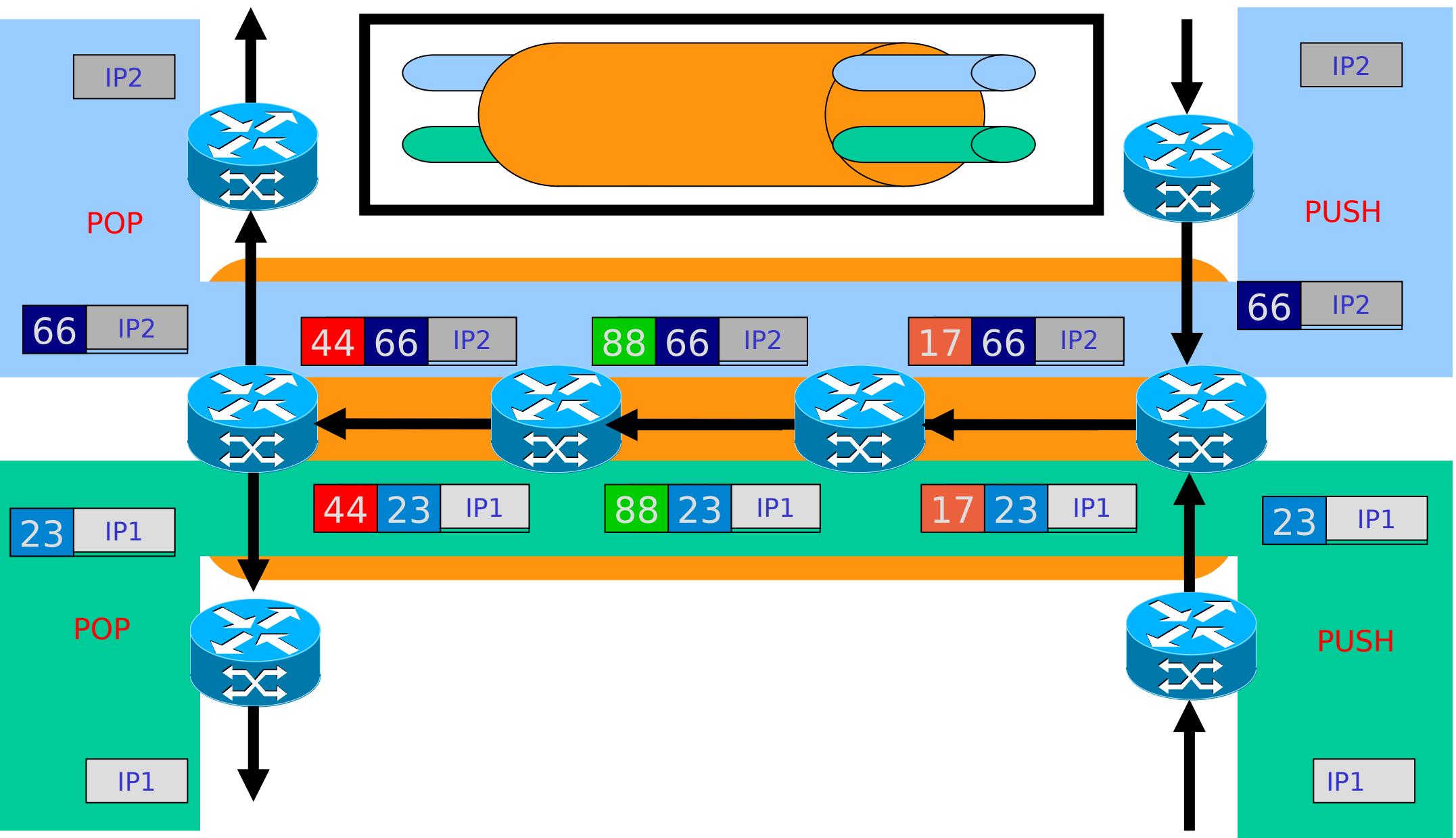
Without PHP



With PHP - Reduces Label Edge Router load



LSP Hierarchy via Label Stacking



Label Distribution Protocols

- Unconstrained routing
 - ◆ Label Distribution Protocol (LDP).
 - ◆ Path is chosen based on IGP shortest path.
- Constrained routing
 - ◆ Constrained by explicit path definition and/or performance requirements (e.g., available bandwidth).
 - ◆ Resource Reservation Protocol with Traffic Engineering (RSVP-TE).
 - ◆ Evolution of RSVP to support traffic engineering and label distribution.
 - ◆ Constrained based Routing LDP (CR-LDP).
 - ◆ Evolution of LDP to support constrained routing.
 - ◆ Deprecated!
- MPLS VPN scope
 - ◆ MP-BGP using address family VPN IPv4 and family specific MP_REACH_NLRI attribute.



Label Distribution Protocol (LDP)

RFC 5036: LDP Specification. (10/2007)

- Dynamic distribution of label binding information.
- LSR discovery.
- Reliable transport with TCP.
- Incremental maintenance of label swapping tables (only deltas are exchanged).
- Designed to be extensible with Type-Length-Value (TLV) coding of messages.
- Modes of behavior that are negotiated during session initialization
 - ◆ Label distribution control (ordered or independent).
 - ◆ Label retention (liberal or conservative).
 - ◆ Label advertisement (unsolicited or on-demand).



LDP Messages

- Discovery messages
 - ◆ Announce and maintain the presence of an LSR in a network.
 - ◆ **Hello Messages** (UDP) sent to “all-routers” multicast address.
 - ◆ Once neighbor is discovered, a LDP session is established over TCP.
- Session messages
 - ◆ Establish (**Initialization Message**) and maintain (**KeepAlive Message**) sessions between LDP peers.
- Advertisement messages
 - ◆ When a new LDP session is initialized and before sending label information an LSR advertises its interface addresses with one or more **Address Messages**.
 - ◆ An LSR withdraw previously advertised interface addresses with **Address Withdraw Messages**.
 - ◆ Create, change, and delete label mappings for FECs.
 - ◆ **Label Mapping, Label Request, Label Abort Request, Label Withdraw, and Label Release Messages.**
- Notification messages
 - ◆ Provide advisory information and to signal error information.

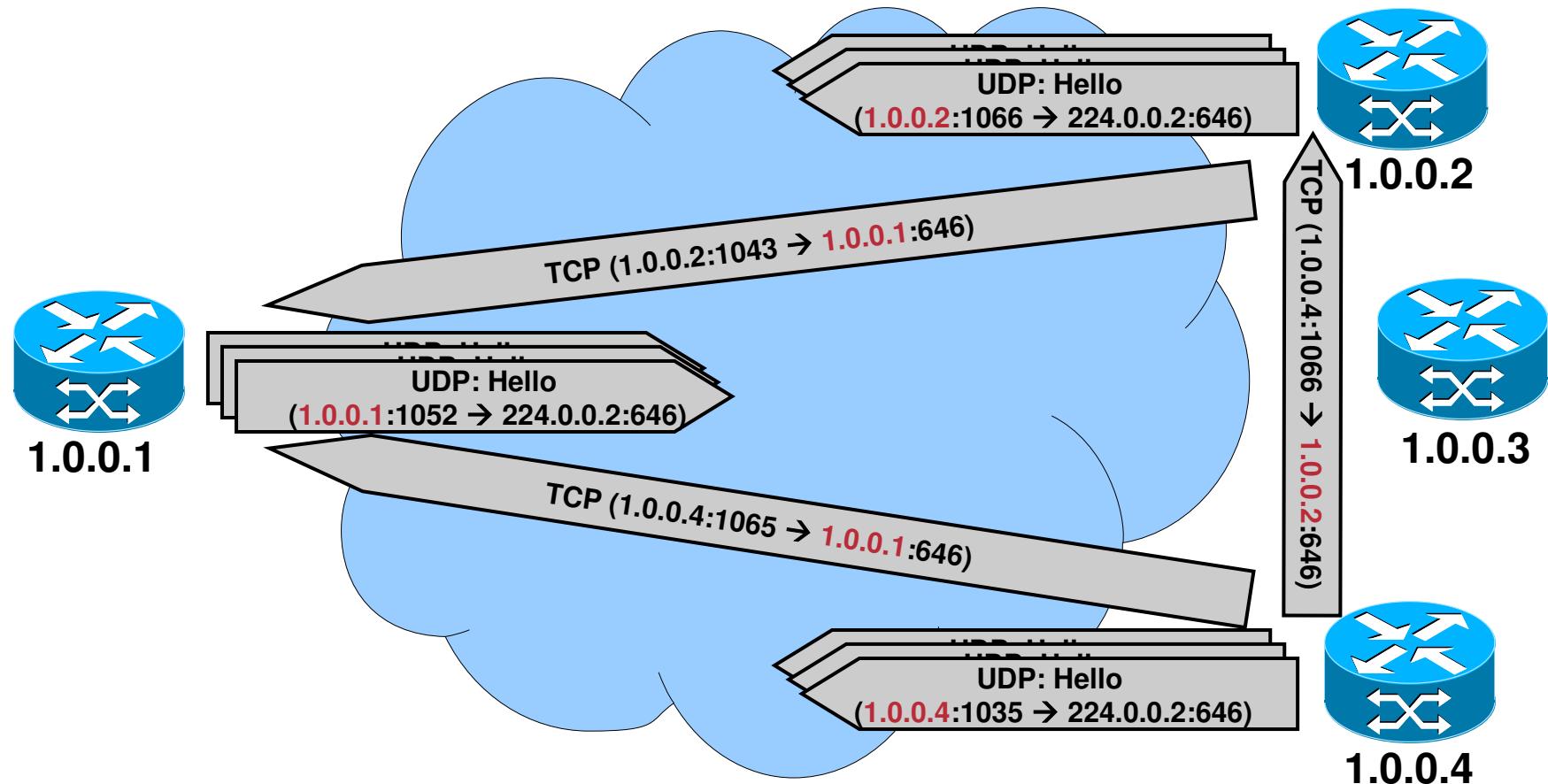


LDP Session Establishment

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

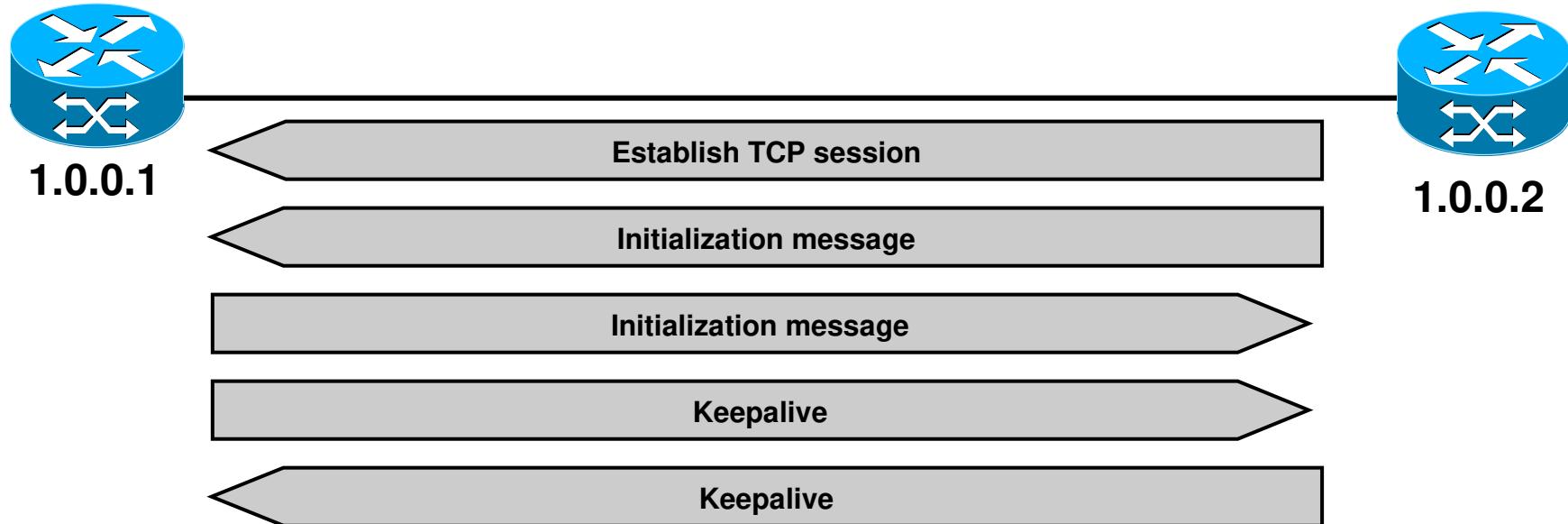


LDP Neighbor Discovery



- LDP Session is started by the router with higher IP address.

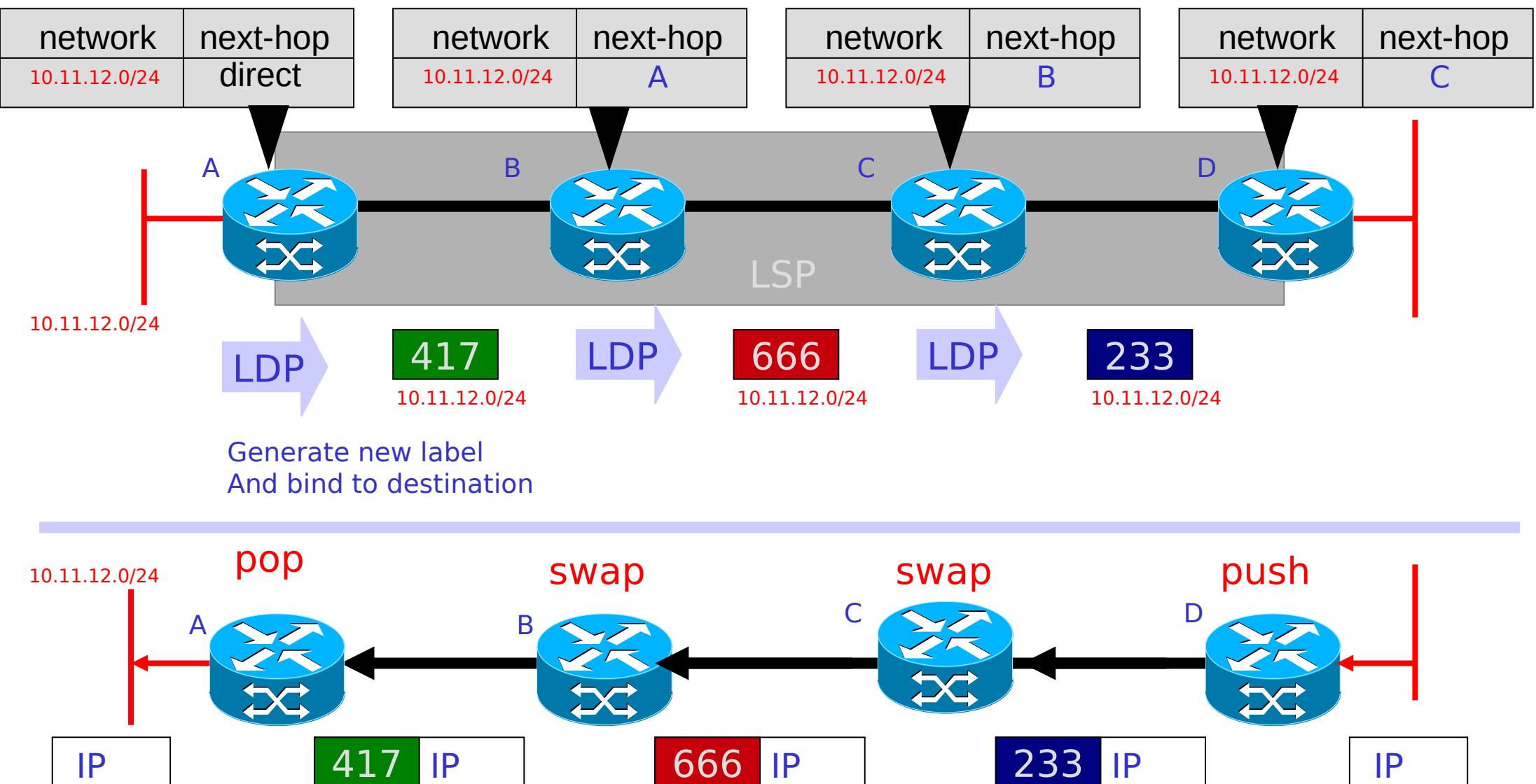
LDP Session Negotiation



- Peers first exchange initialization messages.
- The session is ready to exchange label mappings after receiving the first keepalive.
 - ◆ Keepalives are resent periodically to maintain the LDP/TCP session active.



LDP and Hop-by-Hop routing



Constraint Based Routing

Basic components

1. Specify path constraints
 2. Extend topology database to include resource and constraint information
 3. Find paths that do not violate constraints and optimize some metric
 4. Signal to reserve resources along path
 5. Set up LSP along path (with explicit route)
 6. Map ingress traffic to the appropriate LSPs
- Extend Link State Protocols (IS-IS, OSPF)
- Extend RSVP or LDP or both!

Note: (3) could be offline, or online (perhaps an extension to OSPF)

Problem here: OSPF areas hide information for scalability. So these extensions work best only within an area...

Problem here: what is the “correct” resource model for IP services?

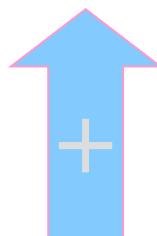


Resource Reservation + Label Distribution

Two competing approaches:

Add label distribution and explicit routes to a resource reservation protocol

RSVP-TE



RSVP

RSVP-TE:
RFC 3209: RSVP-TE: Extensions to RSVP for LSP Tunnels

CR-LDP



LDP

CR-LDP
RFC 3212: Constraint-Based LSP Setup using LDP

As of February 2003, the IETF MPLS working group deprecated CR-LDP and decided to focus purely on RSVP-TE.

RFC 3468: The Multiprotocol Label Switching (MPLS) Working Group decision on MPLS signaling protocols



Resource Reservation Protocol with Traffic Engineering (RSVP-TE)

RFC 3209: RSVP-TE: Extensions to RSVP for LSP Tunnels. (12/2001)

RFC 5151: Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions. (2/2008)

- Evolution of RSVP.
- To map traffic flows onto the physical network topology through label switched paths, requires resource and constraint network information.
 - ◆ Provided by Extend Link State Protocols (IS-IS or OSPF with TE extensions).
 - RFC 3630: Traffic Engineering (TE) Extensions to OSPF Version 2. (9/2003)
 - RFC 5305: IS-IS Extensions for Traffic Engineering. (10/2008)

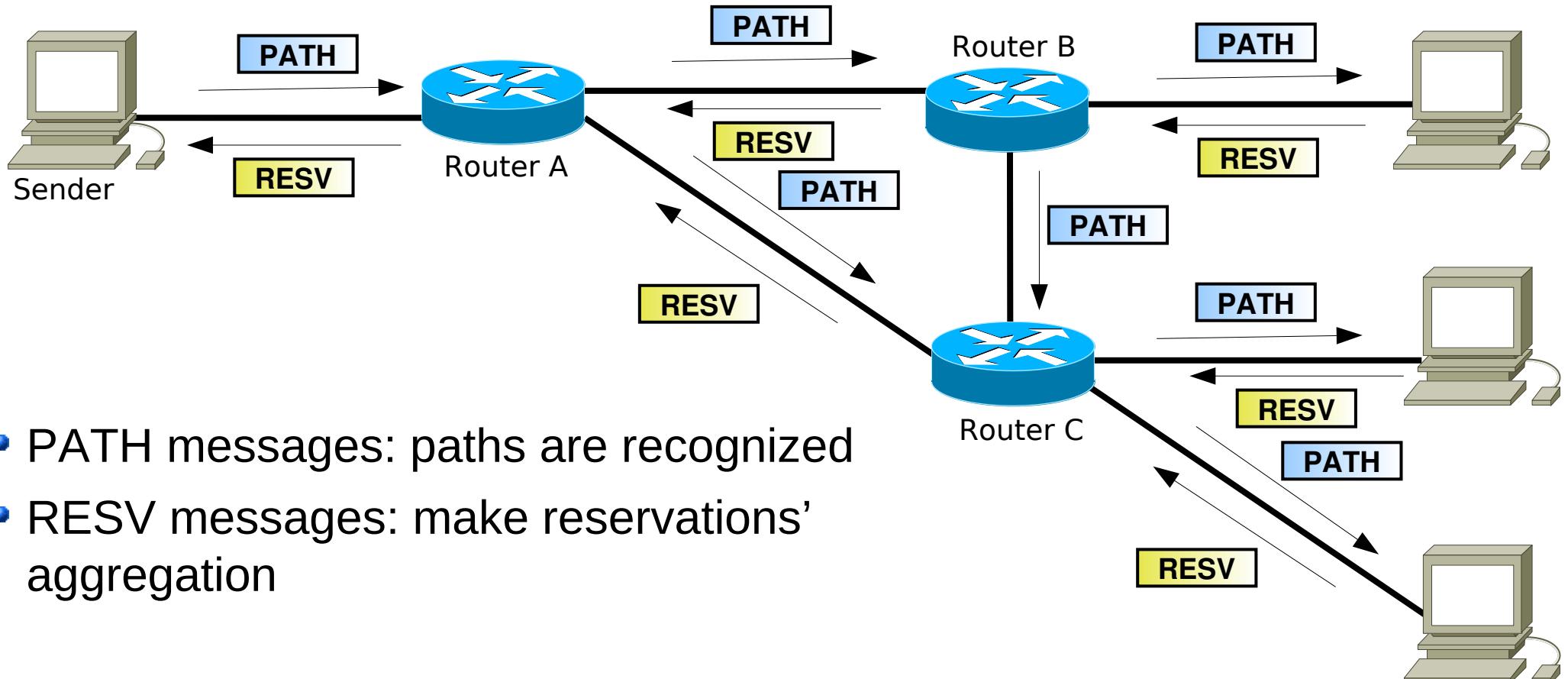


ReSerVation Protocol (RSVP)

- The resource ReSerVation Protocol (RSVP) was developed to communicate resource needs between hosts and network devices (RFC 2205-2215)
- RSVP allows:
 - ◆ The source do describe the characteristics of the IP packets flow.
 - ◆ Destinations to describe the reservation they want.
 - ◆ Routers to know how to process the packets flow in order to fulfill the requested reservation.
- Encapsulated on IP; protocol type = 46 (0x2E)
- Signaling is based on the exchange of PATH and RESV messages.
 - ◆ PATH announces the traffic characteristics at the sender.
 - ◆ RESV achieves reservations that were initiated by the receivers.
 - ◆ If the reservation is not possible, a RESV ERR message is sent.
- The routers reservation states have to be periodically refreshed (soft states).
- RSVP defines a "Session" to be a data flow with a particular destination and transport-layer protocol.
 - ◆ RSVP treats each session independently.



RSVP Signaling



- PATH messages: paths are recognized
- RESV messages: make reservations' aggregation

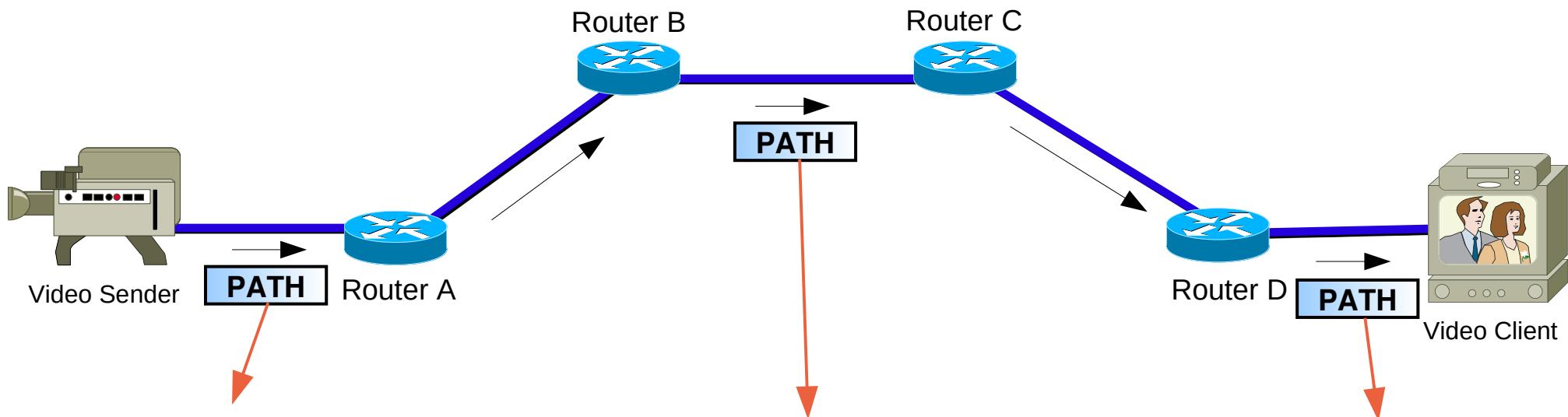


RSVP messages

- PATH (*Type* = 0x01)
 - ◆ Tspec (“flow traffic specification”): contains the parameters that describe the traffic source based on the “Token Bucket” model
- RESV (*Type* = 0x02)
 - ◆ Tspec: the same that was received on the PATH message
 - ◆ FilterSpec (“*filter specification*”): contains the flow descriptor that enables routers to identify packets belonging to this reservation (source address, destination address, protocol type, source port number, destination port number, any combination of these parameters)
 - ◆ Rspec (“*flow reservation specification*”): contains the parameters describing the reservation that the receiver wants to become supported
 - ✚ Rspec is specified if the receiver wants a service of the “*guaranteed service*” type; when it is not specified, it means that the receiver wants a service of the “*controlled load*” type



RSVP PATH (Example)



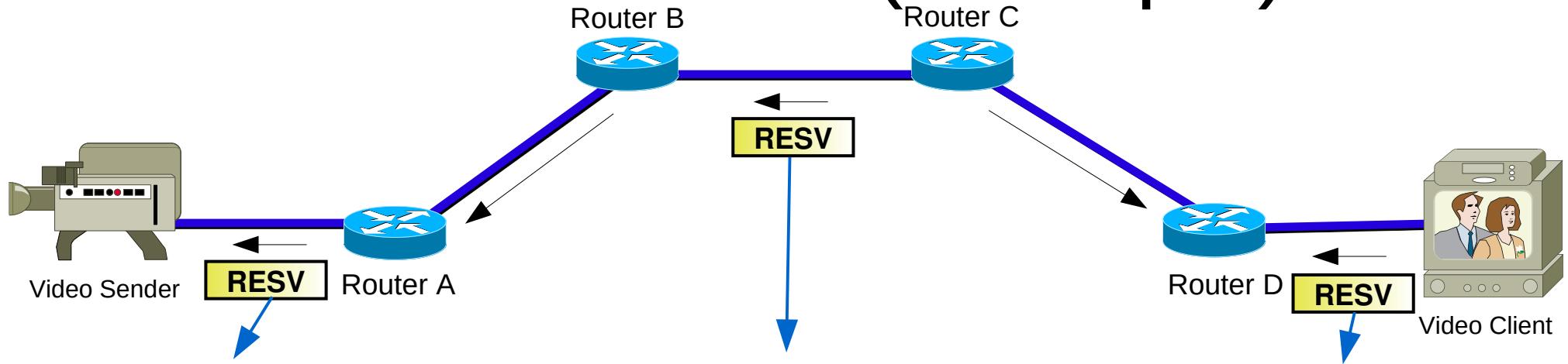
Vs.: 4	iHL: 5	Service	Total Length: 60			
Identification		Flg	Fragment Offset			
Time to Live	Protocol: 46	Header Checksum				
Source Address: Video Server						
Destination Address: Video Client						
1	0	Type: 1	Checksum			
Send_TTL	0	Message Length: 40				
SESSION Length.: 12		Class Nº: 1	Class Type: 1			
Destination Address: Video Client						
Protocol ID	Flags	Destination port				
RSVP_HOP Length. : 12		Class Nº: 3	Class Type: 1			
Last Hop Address: Video Server						
Logical Interface Handle of the last node (LIH)						
TIME_VALUES Length: 8	Class Nº: 5	Class Type: 1				
Update Period (ms)						

Vs.: 4	iHL: 5	Service	Total Length: 60			
Identification		Flg	Fragment Offset			
Time to Live	Protocol: 46	Header Checksum				
Source Address: Video Server						
Destination Address: Video Client						
1	0	Type: 1	Checksum			
Send_TTL	0	Message Length: 40				
SESSION Length: 12		Class Nº: 1	Class Type: 1			
Destination Address: Video Client						
Protocol ID	Flags	Destination Port				
RSVP_HOP Length: 12		Class Nº: 3	Class Type: 1			
Last Hop Address: Router B						
Logical Interface Handle of the last node (LIH)						
TIME_VALUES Length: 8	Class Nº: 5	Class Type: 1				
Update Period (ms)						

Vs.: 4	iHL: 5	Service	Total Length: 60			
Identification		Flg	Fragment Offset			
Time to Live	Protocol: 46	Header Checksum				
Source Address: Video Server						
Destination Address: Video Client						
1	0	Type: 1	Checksum			
Send_TTL	0	Message Length: 40				
SESSION Length: 12		Class Nº: 1	Class Type: 1			
Destination Address: Video Client						
Protocol ID	Flags	Destination Port				
RSVP_HOP Length: 12		Class Nº: 3	Class Type: 1			
Last Hop Address: Router D						
Logical Interface Handle of the last node (LIH)						
TIME_VALUES Length: 8	Class Nº: 5	Class Type: 1				
Update Period (ms)						



RSVP RESV (Example)



Vs.: 4		iHL: 5	Service		Total Length										
			Identification	Flg	Fragment Offset										
Time to Live		Protocol: 46	Header Checksum												
Source Address:		Router A													
Destination Address:		Video Server													
1	0	Type: 2	Checksum												
Send_TTL	0		Message Length												
SESSION Length: 12		Class Nº: 1	Class Type: 1												
Destination Address:		Video Client													
Protocol Id	Flags	Destination protocol port													
RSVP_HOP Length: 12		Class Nº: 3	Class Type: 1												
Address of the last node:		Router A													
Logical Interface Handle of the last node (LIH)															
TIME_VALUES Length: 8		Class Nº: 5	Class Type: 1												
Update period (ms)															
STYLE Object Length : 8		Class Nº: 8	Class Type: 1												
Flags	Style Option Vector: 0x00000A (FF)														
FLOWSPEC Length		Class Nº: 9	Class Type												
FLOWSPEC object contents															
FILTER_SPEC Length: 12		Class Nº: 10	Class Type: 1												
Source Address: Video Server															
Reserved	Reserved		Source protocol port												

Vs.: 4		iHL: 5	Service		Total Length										
			Identification	Flg	Fragment Offset										
Time to Live		Protocol: 46	Header Checksum												
Source Address:		Router C													
Destination Address:		Router B													
1	0	Type: 2	Checksum												
Send_TTL	0		Message Length												
SESSION Length: 12		Class Nº: 1	Class Type: 1												
Destination Address:		Video Client													
Protocol Id	Flags	Destination protocol port													
RSVP_HOP Length: 12		Class Nº: 3	Class Type: 1												
Address of the last node:		Router C													
Logical Interface Handle of the last node (LIH)															
TIME_VALUES Length: 8		Class Nº: 5	Class Type: 1												
Update period (ms)															
STYLE Object Length : 8		Class Nº: 8	Class Type: 1												
Flags	Style Option Vector: 0x00000A (FF)														
FLOWSPEC Length		Class Nº: 9	Class Type												
FLOWSPEC object contents															
FILTER_SPEC Length: 12		Class Nº: 10	Class Type: 1												
Source Address: Video Server															
Reserved	Reserved		Source protocol port												

Vs.: 4		iHL: 5	Service		Total Length										
			Identification	Flg	Fragment Offset										
Time to Live		Protocol: 46	Header Checksum												
Source Address:		Video Client													
Destination Address:		Router D													
1	0	Type: 2	Checksum												
Send_TTL	0		Message Length												
SESSION Length: 12		Class Nº: 1	Class Type: 1												
Destination Address:		Video Client													
Protocol Id	Flags	Destination protocol port													
RSVP_HOP Length: 12		Class Nº: 3	Class Type: 1												
Address of the last node:		Video Client													
Logical Interface Handle of the last node (LIH)															
TIME_VALUES Length: 8		Class Nº: 5	Class Type: 1												
Update period (ms)															
STYLE Object Length : 8		Class Nº: 8	Class Type: 1												
Flags	Style Option Vector: 0x00000A (FF)														
FLOWSPEC Length		Class Nº: 9	Class Type												
FLOWSPEC object contents															
FILTER_SPEC Length: 12		Class Nº: 10	Class Type: 1												
Source Address: Video Server															
Reserved	Reserved		Source protocol port												



Extensions to RSVP for LSP Tunnels

- The SENDER_TEMPLATE (or FILTER_SPEC) object together with the SESSION object uniquely identifies an LSP tunnel (flow).
- LSP Tunnel related new objects
 - ◆ Explicit Route
 - ▶ Carried in PATH and contains a series of variable-length data items called sub-objects.
 - ▶ Possible sub-objects: IPv4 prefix, IPv6 prefix, and autonomous system number.
 - ◆ Label Request
 - ▶ Carried in PATH requesting a label for a specific tunnel/flow.
 - ▶ Request can be without label range, with an ATM label range, or with an Frame Relay label range.
 - ◆ Label
 - ▶ Carried in RESV messages and contain a single label for a specific tunnel/flow.
 - ◆ Record Route
 - ▶ Carried in PATH and RESV, used to collect detailed path information and useful for loop detection and diagnostics.
 - ◆ Session Attribute
 - ▶ Carried in PATH, used to define the type and name of the session/tunnel/flow, also used to define priority values.
- LSP Tunnel related new object types
 - ◆ Session object new types
 - ▶ LSP_TUNNEL_IPv4 and LSP_TUNNEL_IPv6
 - ◆ Sender Template object new types
 - ▶ LSP_TUNNEL_IPv4 and LSP_TUNNEL_IPv6
 - ◆ Filter Specification object new types
 - ▶ LSP_TUNNEL_IPv4 and LSP_TUNNEL_IPv6



RSVP-TE PATH and RESV (example)

Resource Reservation Protocol (RSVP): PATH Message. SESSION: IPv4-LSP

▷ RSVP Header. PATH Message.

▷ SESSION: IPv4-LSP, Destination 192.2.0.11, Tunnel ID 2, Ext ID c002000a.

▷ HOP: IPv4, 200.10.2.10

▷ TIME VALUES: 30000 ms

▷ EXPLICIT ROUTE: IPv4 200.10.2.2, IPv4 200.2.11.2, IPv4 200.2.11.11,

▷ LABEL REQUEST: Basic: L3PID: IP (0x0800)

▷ SESSION ATTRIBUTE: SetupPrio 7, HoldPrio 7, SE Style, [RA_t2]

▷ SENDER TEMPLATE: IPv4-LSP, Tunnel Source: 192.2.0.10, LSP ID: 8.

▷ SENDER TSPEC: IntServ, Token Bucket, 18750 bytes/sec.

▷ ADSPEC

▷ Resource Reservation Protocol (RSVP): RESV Message. SESSION: IPv4-LSP

▷ RSVP Header. RESV Message.

▷ SESSION: IPv4-LSP, Destination 192.2.0.11, Tunnel ID 2, Ext ID c002000a.

▷ HOP: IPv4, 200.10.2.2

▷ TIME VALUES: 30000 ms

▷ STYLE: Shared-Explicit (18)

▷ FLOWSPEC: Controlled Load: Token Bucket, 18750 bytes/sec.

▷ FILTERSPEC: IPv4-LSP, Tunnel Source: 192.2.0.10, LSP ID: 8.

▷ LABEL: 19



Traffic Engineering Extensions to OSPF

- RFC 3630: Traffic Engineering (TE) Extensions to OSPF Version 2. (9/2003)
- OSPF Traffic Engineering (TE) extensions are used to advertise TE Link State Advertisements (TE-LSAs) containing information about TE-enabled links.
 - ◆ Traffic Engineering LSA is a type 10 Opaque LSAs, which have an area flooding scope.
- TE-LSA contains one of two possible top-level Type Length Values (TLVs)
 - ◆ **Router Address:** specifies a stable IP address of the advertising router that is always reachable if there is any connectivity to it; this is typically implemented as a "loopback address";
 - ◆ **Link:** describes a single link with a set of sub-TLVs (Link type, Link ID, Local interface IP address, Remote interface IP address, Traffic engineering metric, Maximum bandwidth, Maximum reservable bandwidth, Unreserved bandwidth, and Administrative group).
- The information made available by these extensions can be used to build an extended link state database
 - ◆ Can be used to:
 - ◆ Monitoring the extended link attributes;
 - ◆ Local constraint-based source routing;
 - ◆ Global traffic engineering.



OSPF-TE Opaque Area Database

- Router Address TLV

LS age: 250

Options: (No TOS-capability, DC)

LS Type: Opaque Area Link

Link State ID: 1.0.0.0

Opaque Type: 1

Opaque ID: 0

Advertising Router: 192.2.0.2

LS Seq Number: 80000001

Checksum: 0xDACD

Length: 28

Fragment number : 0

MPLS TE router ID : 192.2.0.2

Number of Links : 0

- Link TLV

LS age: 246

Options: (No TOS-capability, DC)

LS Type: Opaque Area Link

Link State ID: 1.0.0.2

Opaque Type: 1

Opaque ID: 2

Advertising Router: 192.2.0.2

LS Seq Number: 80000001

Checksum: 0x2FBB

Length: 124

Fragment number : 2

Link connected to Broadcast network

Link ID : 200.1.2.2

Interface Address : 200.1.2.2

Admin Metric : 1

Maximum bandwidth : 12500000

Maximum reservable bandwidth : 64000

Number of Priority : 8

Priority 0 : 64000 Priority 1 : 64000

Priority 2 : 64000 Priority 3 : 64000

Priority 4 : 64000 Priority 5 : 64000

Priority 6 : 64000 Priority 7 : 64000

Affinity Bit : 0x0

IGP Metric : 1

Number of Links : 1

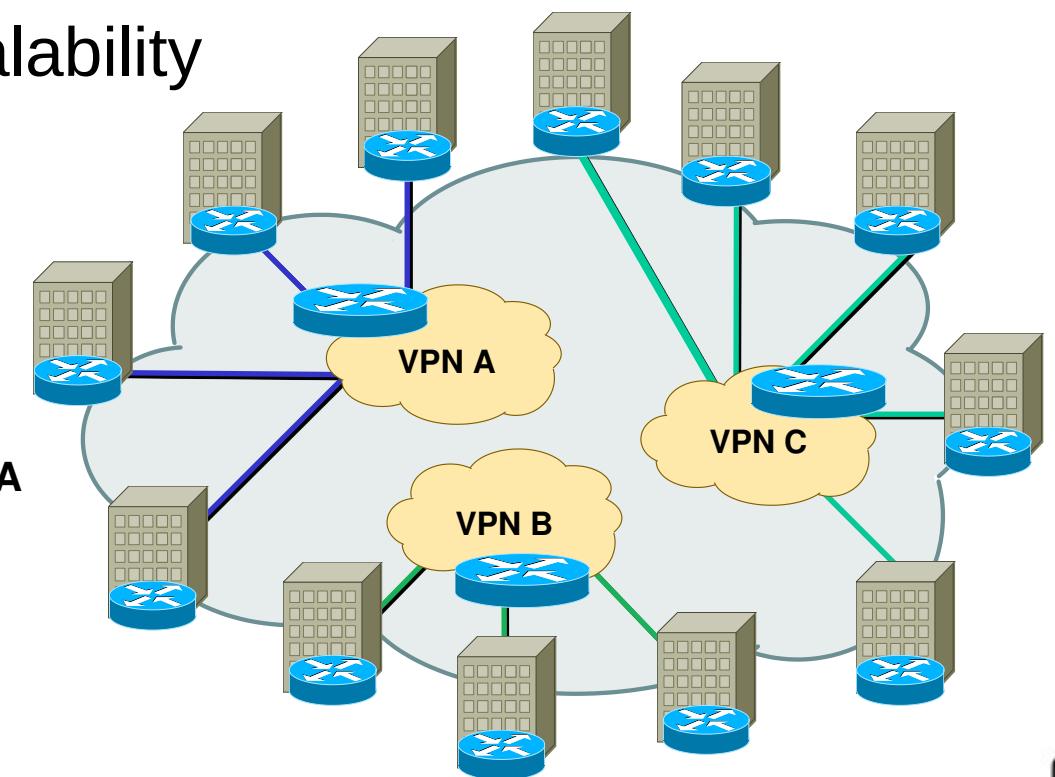
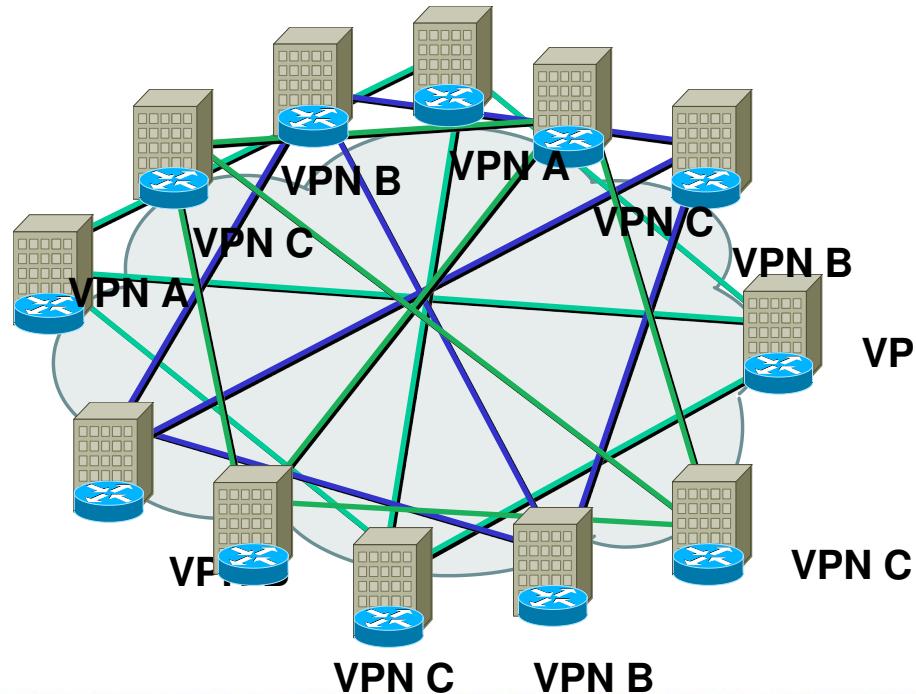


MPLS VPN

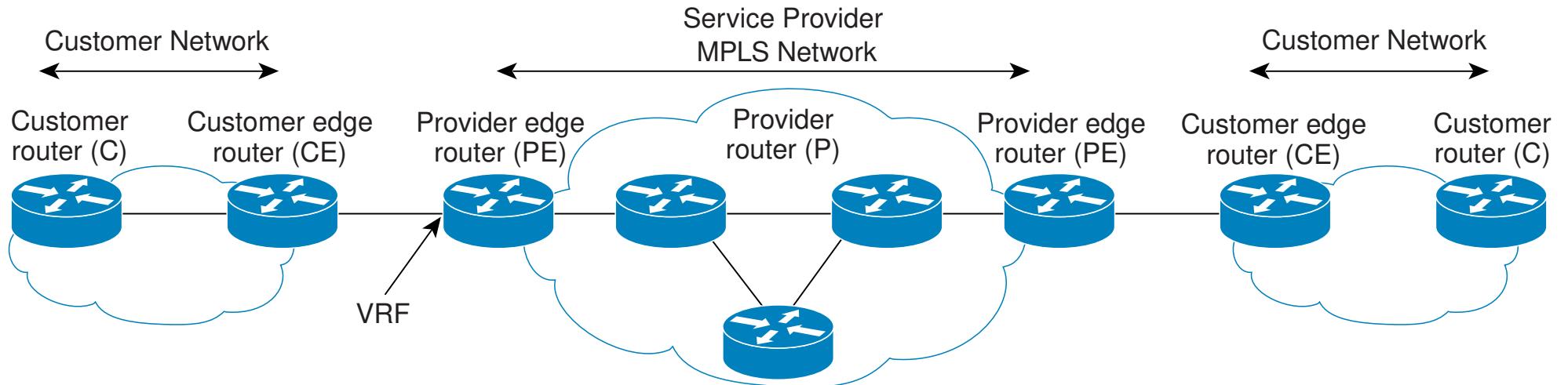


MPLS L3 VPNs using BGP (RFC2547)

- End user perspective
 - ◆ Virtual Private IP service.
 - ◆ Simple routing – just point default to provider.
 - ◆ Full site-site connectivity without the usual drawbacks (routing complexity, scaling, configuration, cost).
- Major benefit for provider – scalability



MPLS VPN Terminology

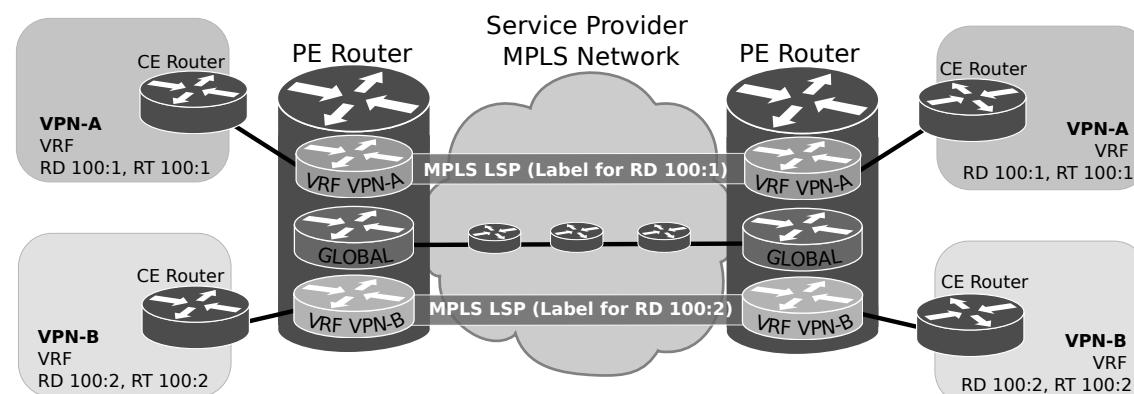


- Customer router (C) is connected only to other customer devices.
- Customer Edge (CE) router peers at Layer 3 to the Provider Edge (PE).
 - ◆ The PE-CE Interface runs either a dynamic routing protocol (eBGP, RIPv2, EIGRP, or OSPF) or has static routing (Static, Connected).
- Provider (P) router, resides in the core of the provider network.
 - ◆ Participates in the control plane for customer prefixes. The P router is also referred to as a Label Switch Router (LSR), in reference to its primary role in the core of the network, performing label switching/swapping of MPLS traffic.
- Provider Edge (PE) router, sits at the edge of the MPLS SP network.
 - ◆ In an MPLS VPN context, separate VRF routing tables are allocated for each user group.
 - ◆ Contains a global routing table for routes in the core SP infrastructure.
 - ◆ The PE is sometimes referred to as a Label Edge Router (LER) or Edge Label Switch Router (ELSR) in reference to its role at the edge of the MPLS cloud, performing label imposition and disposition.

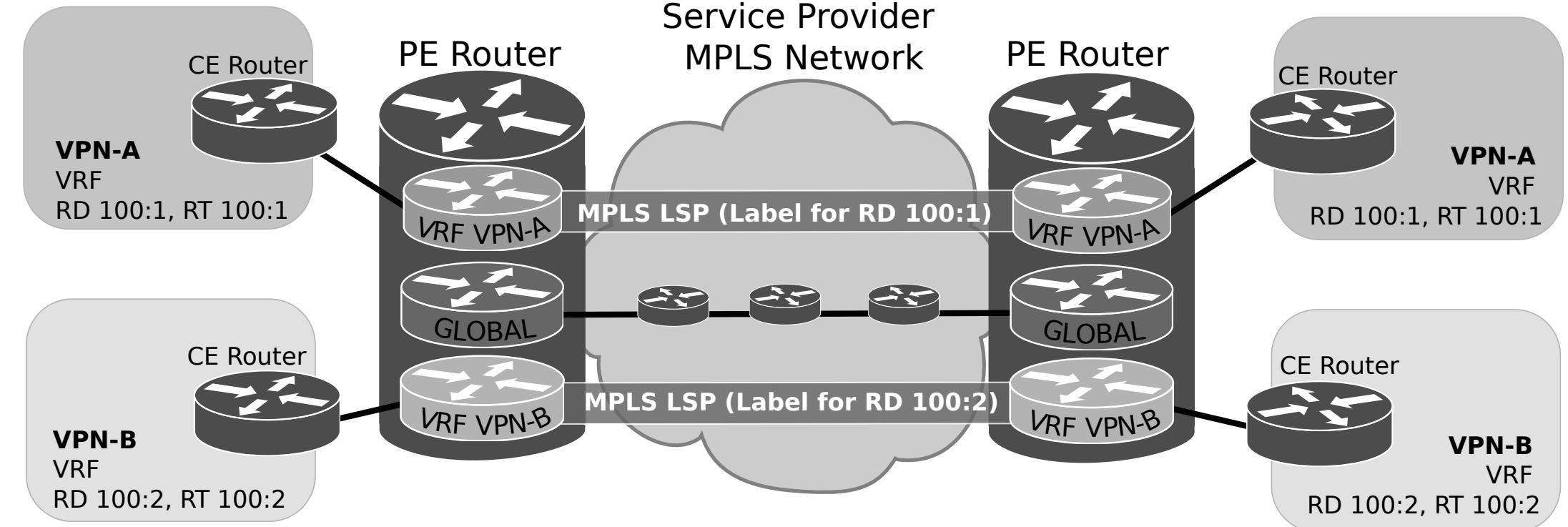


Virtual Routing and Forwarding (VRF)

- Virtual Routing and Forwarding (VRF) instance, is separate from the global routing table that exists on PE routers.
- PE routers maintain separate routing tables:
 - ◆ Global routing table
 - ➡ Contains all PE and P routes (perhaps BGP).
 - ➡ Populated by the VPN backbone IGP .
 - ◆ VRF table
 - ➡ Routing and forwarding table associated with one or more directly connected sites (CE routers).
 - ➡ VRF is associated with any type of interface, whether logical or physical (e.g. sub/virtual/tunnel) .
 - ➡ Interfaces may share the same VRF if the connected sites share the same routing information.
 - ➡ Routes are injected into the VRF from the CE-PE routing protocols for that VRF and any MP-BGP announcements that match the defined VRF.



MPLS-VPN & VRF



Carrying VPN Routes in BGP

- Need some way to get the VRF routing information off the PE and to other PEs.
- This is done with MP-BGP.
- Additions to MP-BGP to carry MPLS-VPN info:
 - ◆ Route Target (RT) sent in EXTENDED_COMMUNITY attribute.
 - ◆ MP_REACH_NLRI attribute for Labeled VPN IPv4 (VPNv4) address family,
 - ◆ VPN IPv4 network.
 - ◆ Route Distinguisher (RD).
 - ◆ MPLS Label.

Border Gateway Protocol - UPDATE Message

Marker: ffffffffffffffffffffff

Length: 91

Type: UPDATE Message (2)

Withdrawn Routes Length: 0

Total Path Attribute Length: 68

Path attributes

▷ Path Attribut - ORIGIN: INCOMPLETE

▷ Path Attribut - AS_PATH: empty

▷ Path Attribut - MULTI_EXIT_DISC: 0

▷ Path Attribut - LOCAL_PREF: 100

Path Attribut - EXTENDED_COMMUNITIES

▷ Flags: 0xc0: Optional, Transitive, Complete

Type Code: EXTENDED_COMMUNITIES (16)

Length: 8

▷ Carried extended communities: (1 community)

▷ Community Transitive Two-Octet AS Route Target: 200:1

Path Attribut - MP_REACH_NLRI

▷ Flags: 0x80: Optional, Non-transitive, Complete

Type Code: MP_REACH_NLRI (14)

Length: 33

Address family: IPv4 (1)

Subsequent address family identifier: Labeled VPN Unicast (128)

▷ Next hop network address (12 bytes)

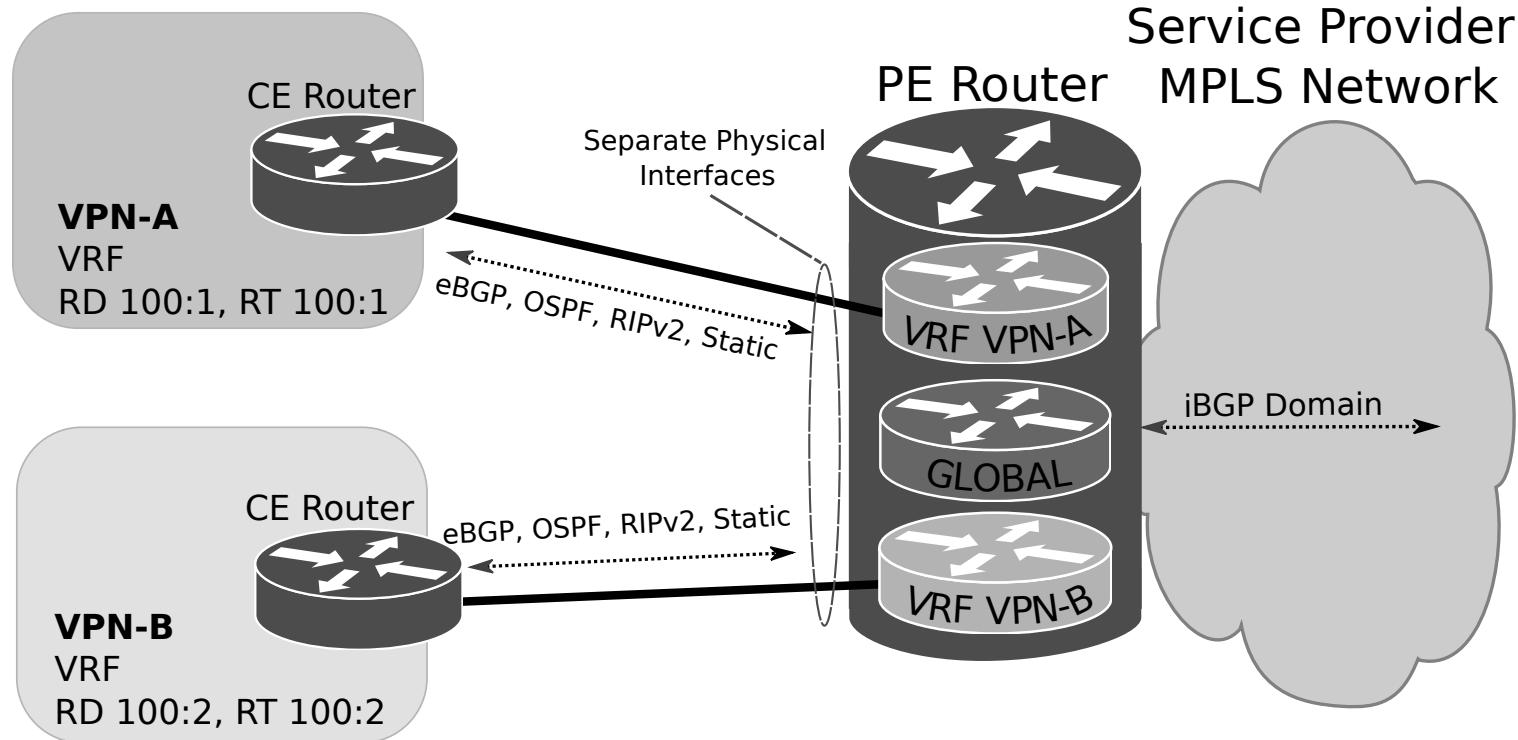
Subnetwork points of attachment: 0

▷ Network layer reachability information (16 bytes)

▷ Label Stack=24 (bottom) RD=200:1, IPv4=192.1.1.0/25



VRF Route Population



- VRF is populated locally through PE and CE routing protocol exchange.
 - ◆ EBGP, OSPF, RIPv2, and Static routing.
 - ◆ “Connected” is also supported.
- Separate routing context for each VRF.
 - ◆ Routing protocol context (e.g., MP-BGP).
 - ◆ Separate process (e.g., OSPF).

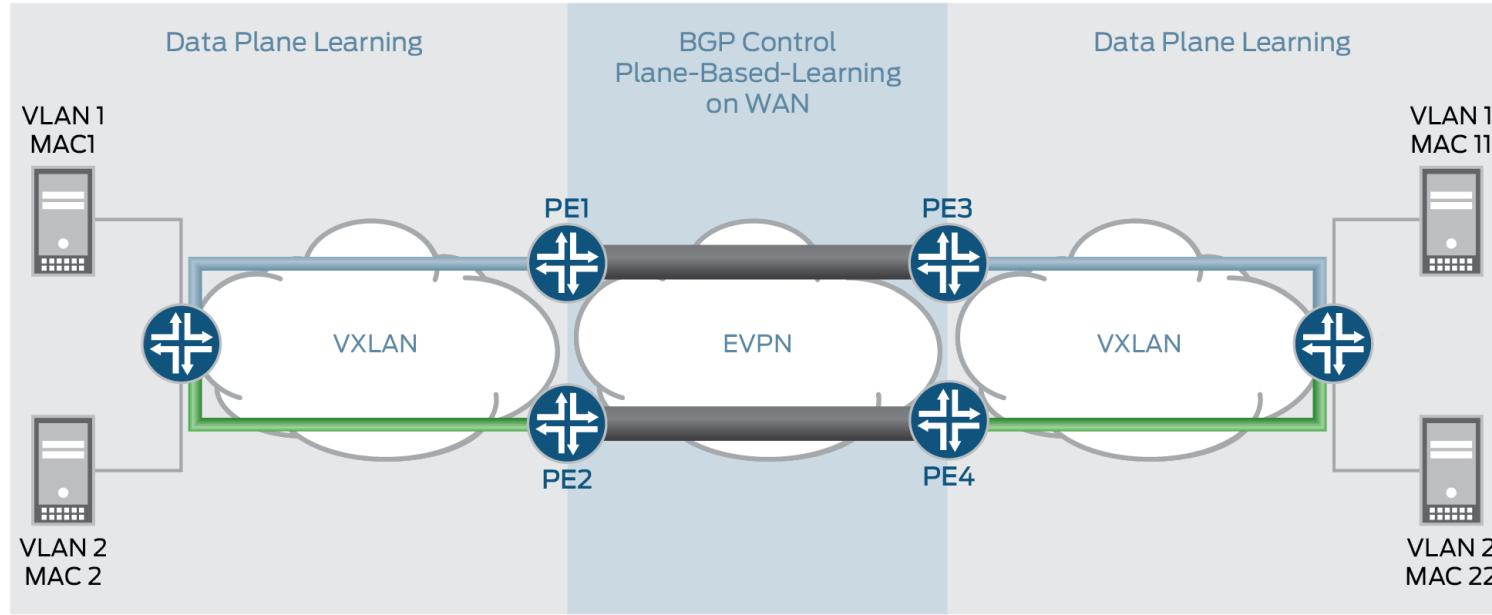


MPLS-VPN Packet Forwarding

- Between PE and CE, regular IP packets (currently)
- Within the provider network—label stack
 - ◆ Outer label: “get this packet to the egress PE”
 - ◆ Inner label: “get this packet to the egress CE”
- MPLS nodes forward packets based on TOP label!!!
 - ◆ any subsequent labels are ignored
- Penultimate Hop Popping procedures used one hop prior to egress PE router (shown in example)



VXLAN-EVPN



- Stands for Virtual Extensible LAN (VXLAN) Ethernet Virtual Private Network (EVPN).
- Multi-Site architecture for seamless Layer 2 and Layer 3 extension.
 - ◆ Commonly used in datacenters.
- Constructed based on MP-BGP Layer2 VPN EVPN family.
 - ◆ Is similar to MP-BGP MPLS IP VPN.
- Defines a new type of BGP network layer reachability information (NLRI).
 - ◆ EVPN NLRI
 - ◆ Defines new BGP EVPN routes to implement MAC address learning and advertisement between Layer 2 networks at different sites.



Access and Core Networks



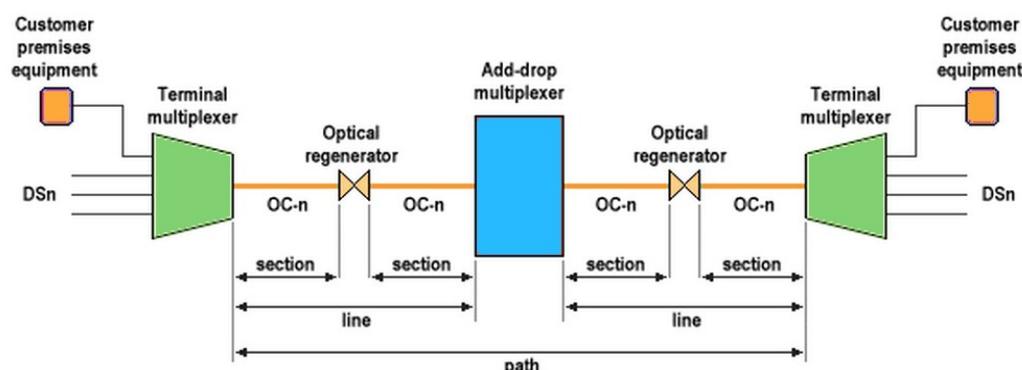
SONET/SDH

- Synchronous Optical NETwork (SONET) – North America

- ◆ TDM physical layer standard for optical fiber communications.
 - ◆ Compatible with US and Canada PDH - 8000 frames/sec - T frame = 125 µsec.
 - ◆ Point-to-point (linear) or ring Optical Carriers (OC)
- ◆ ITU version = Synchronous Digital Hierarchy (SDH) – Rest of the World
 - ◆ Small differences, but interoperable at higher speeds.
- ◆ Direct mapping of lower levels into higher ones
- ◆ SONET frames: STS. SDH frames: STM.
 - ◆ Transport all PDH types in one universal hierarchy.
 - ◆ Also transports ATM cells and general packet data.
- ◆ SONET Add-Drop Multiplexing
 - ◆ Allows taking individual channels in and out without full demultiplexing.

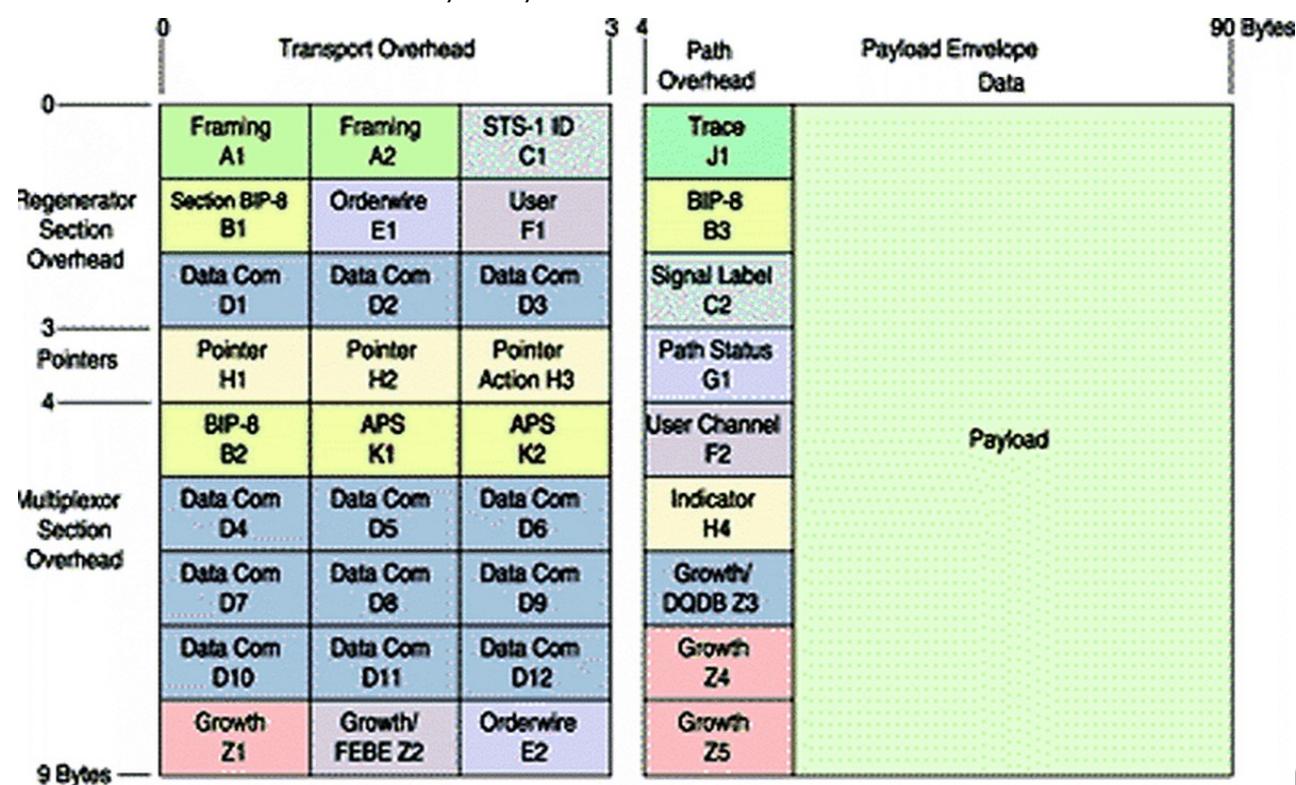
SONET/SDH Designations and bandwidths

SONET Optical Carrier Level	SONET Frame Format	SDH level and Frame Format	Payload bandwidth[nb 3] (Kbit/s)	Line Rate (Kbit/s)
OC-1	STS-1	STM-0	50,112	51,840
OC-3	STS-3	STM-1	150,336	155,520
OC-12	STS-12	STM-4	601,344	622,080
OC-24	STS-24	–	1,202,688	1,244,160
OC-48	STS-48	STM-16	2,405,376	2,488,320
OC-192	STS-192	STM-64	9,621,504	9,953,280
OC-768	STS-768	STM-256	38,486,016	39,813,120
OC-3072	STS-3072	STM-1024	153,944,064	159,252,480



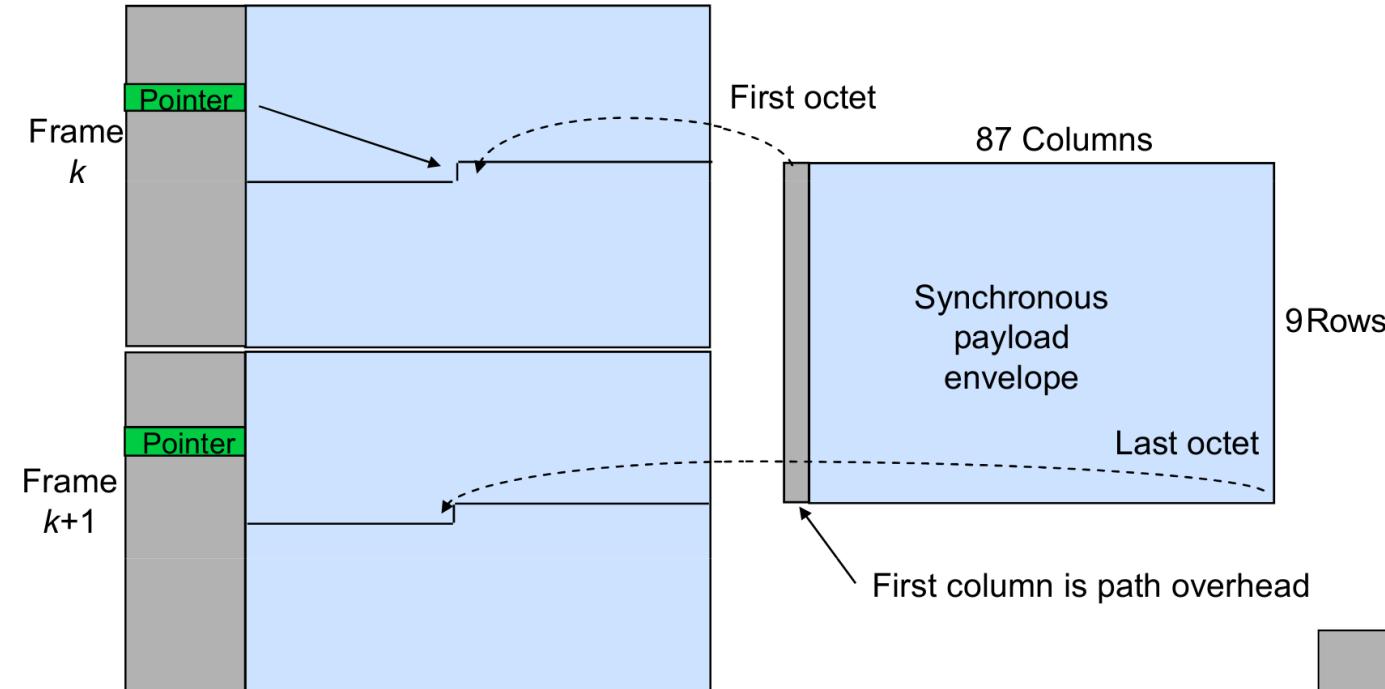
STS Frames

- Transport Overhead (TOH):
 - ◆ Processed at every SONET node.
 - ◆ Occupies a portion of each SONET frame.
 - ◆ Carries management and link integrity information.
- Synchronous Payload Envelope (SPE):
 - ◆ Path Overhead (POH),
 - ◆ Inserted & removed at the ends.
 - ◆ Data/Payload.
- STS-1 Frame: 9 rows x 90 cols.
 - ◆ 810 bytes per frame, 8000 frames/sec → 51.84Mbps
 - ◆ 810x64kbps → 51.84Mbps
- Special OH bytes
 - ◆ H1, H2, H3: Pointer Action

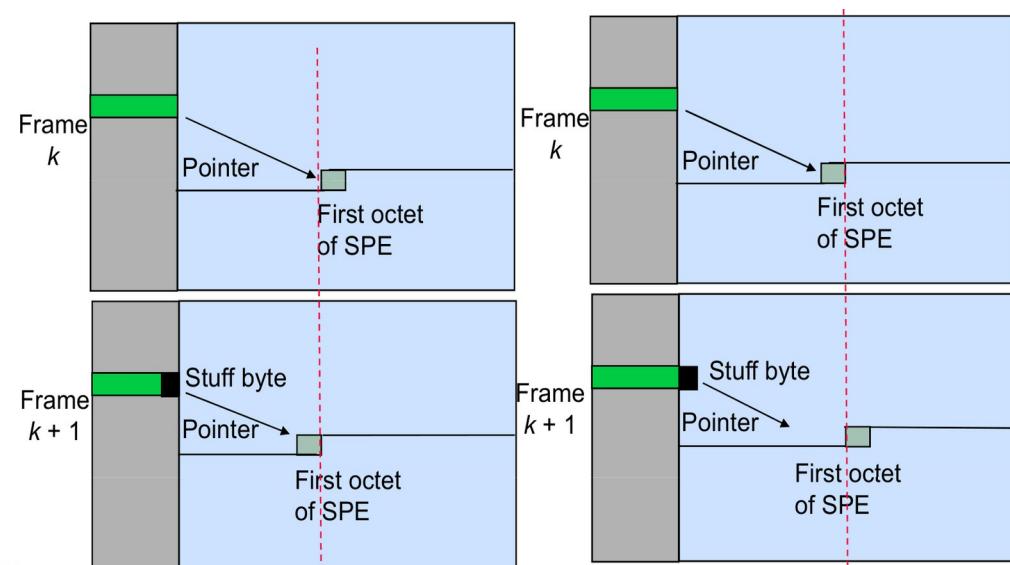


SPE Over Consecutive Frames

- Pointer indicates where SPE begins within a frame

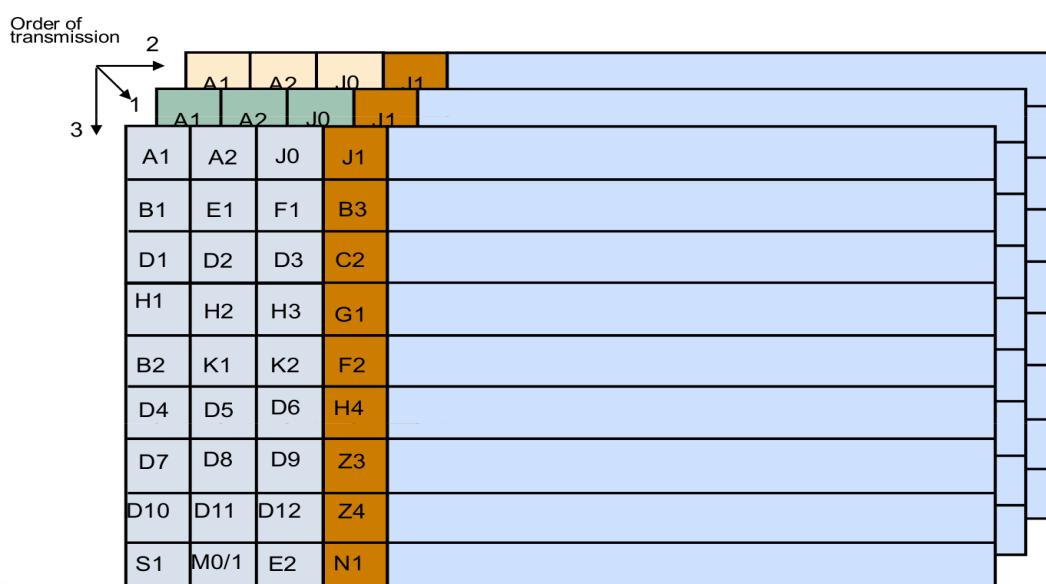
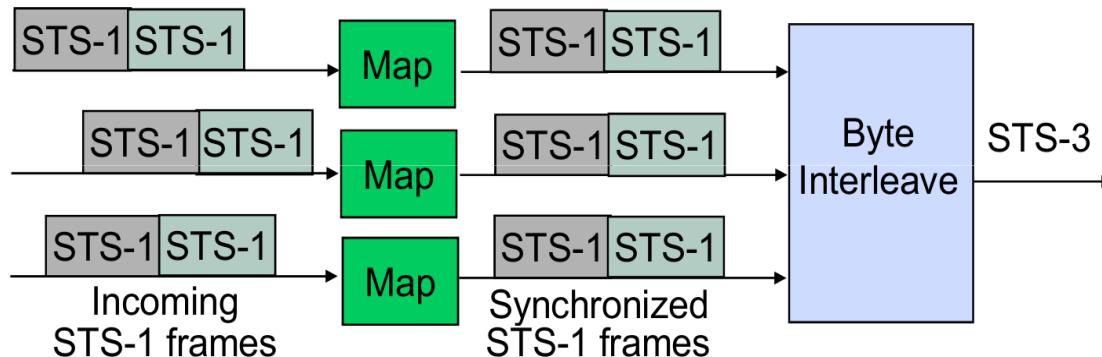


- Pointer enables add/drop capability
 - Positive/negative byte stuffing.



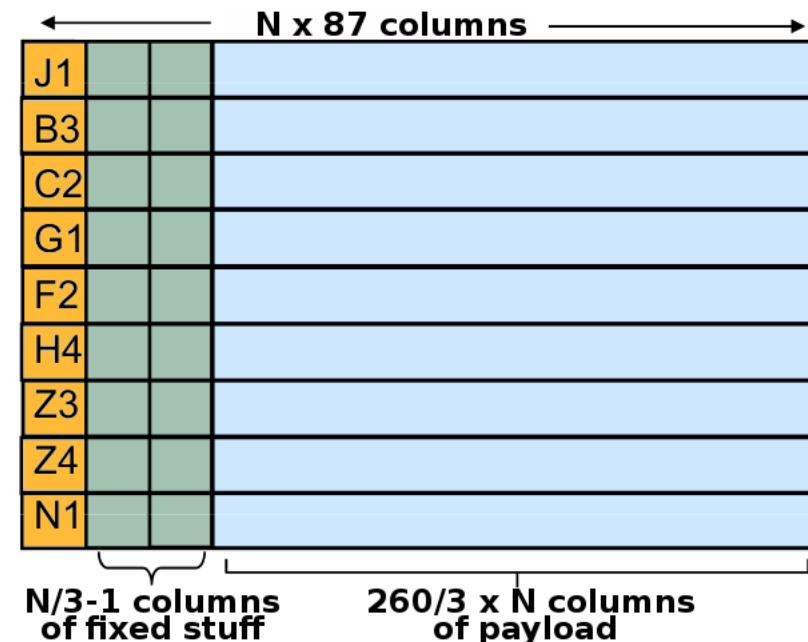
STS Multiplexing

- Synchronous/Channelized
 - ◆ Synchronize each incoming STS-1 to local clock → STS-1s.
 - ◆ All STS-1s are byte interleaved to produce STS-n.



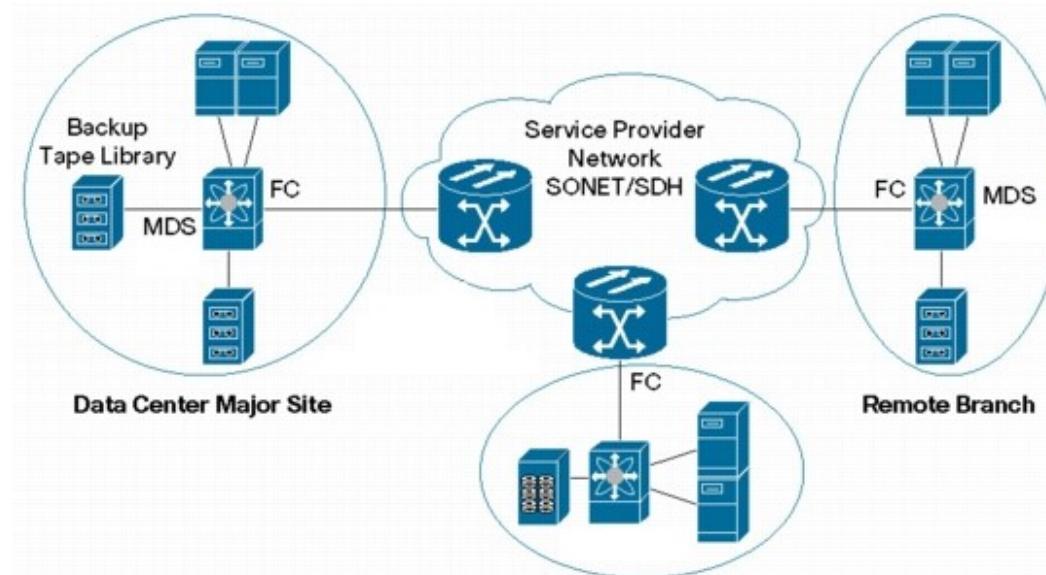
- Concatenated Payload
 - ◆ H1,H2,H3 tell us if there is concatenation
 - ◆ STS-3c has more payload than 3 STS-1s
 - ◆ STS-Nc payload = $N \times 260/3 \times 9$ bytes
 - ◆ Payload rates
 - ◆ OC-3c = 149.760 Mbps, OC-12c = 599.040 Mbps, OC-48c = 2.3961 Gbps, OC-192c = 9.5846 Gbps

- OC-Nc Concatenated Payload

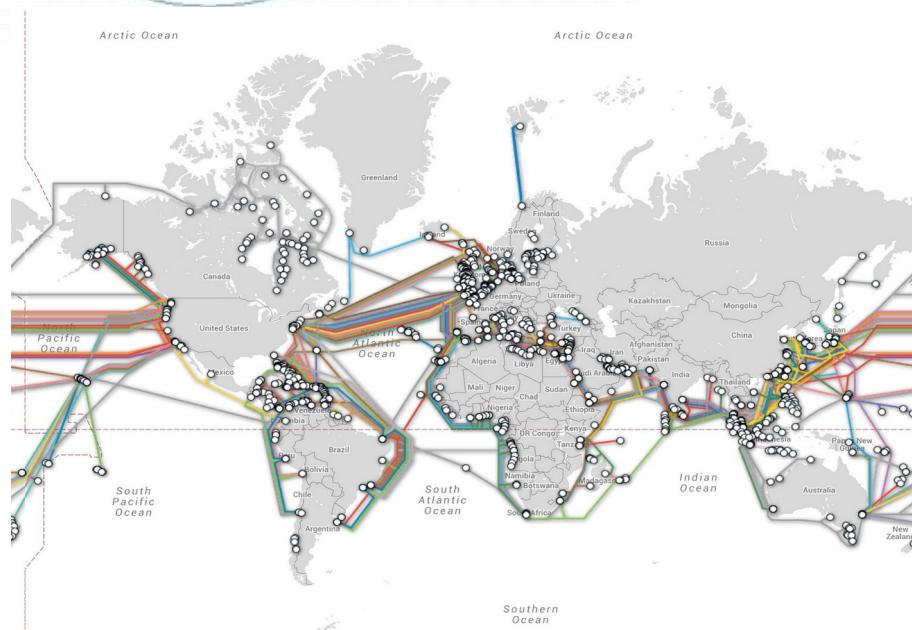


SONET/SDH Usage

Network/ISP Core

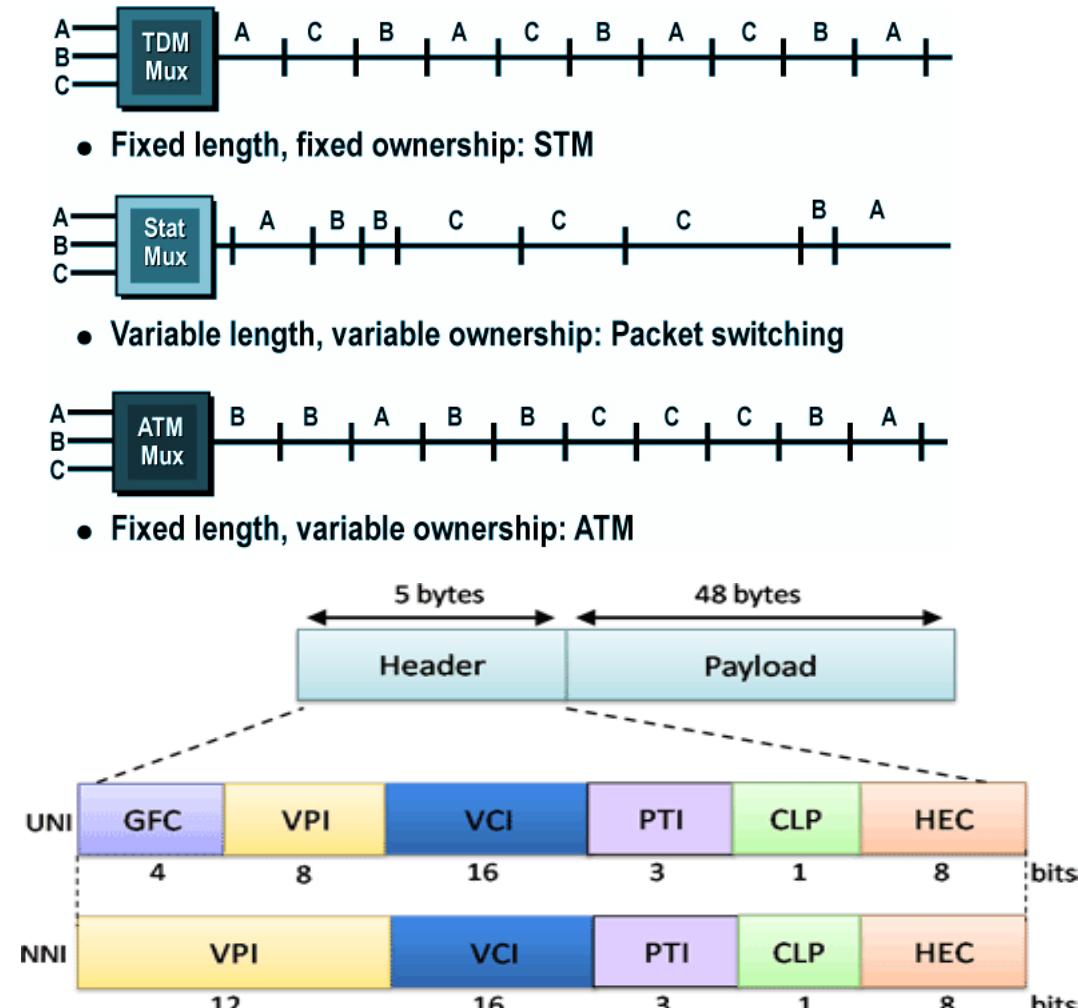


Long-range point-to-point links



Asynchronous Transfer Mode (ATM)

- ATM is a blend of Synchronous Transfer Mode (STM) and packet switching.
 - It has variable assignment, based on the arrival rate and delay sensitivity of the traffic.
 - However, after the assignment occurs, uses fixed-length time slots called cells.
 - Delay-sensitive traffic has immediate assignment
 - Data traffic can be temporarily buffered before being transmitted.
- Is a form of cell switching using small fixed-sized data units called cells.
 - 53 bytes: 5 header and 48 data.



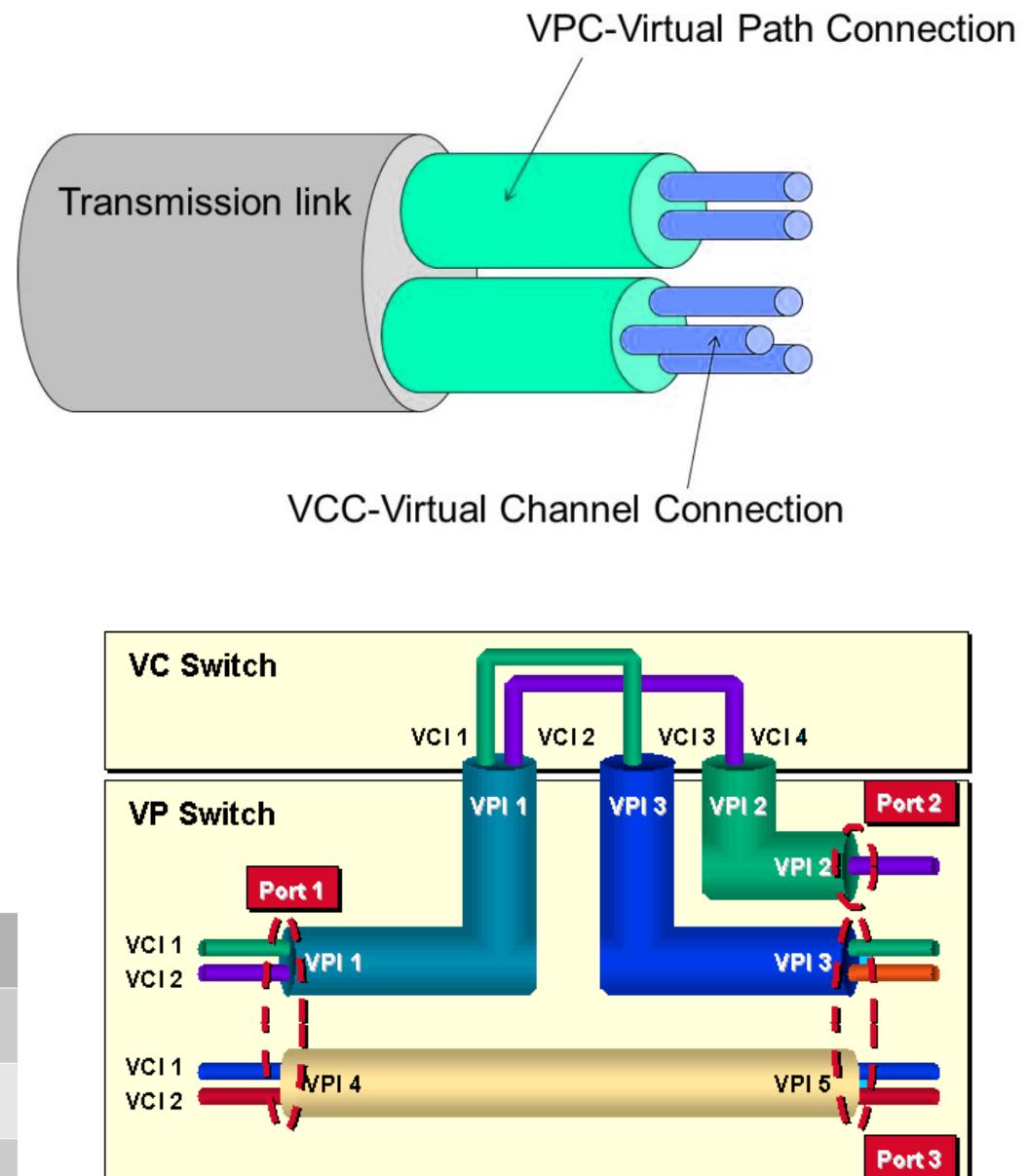
UNI (User-Network Interface).
NNI (Network-Network Interface).



ATM Connections and Switching

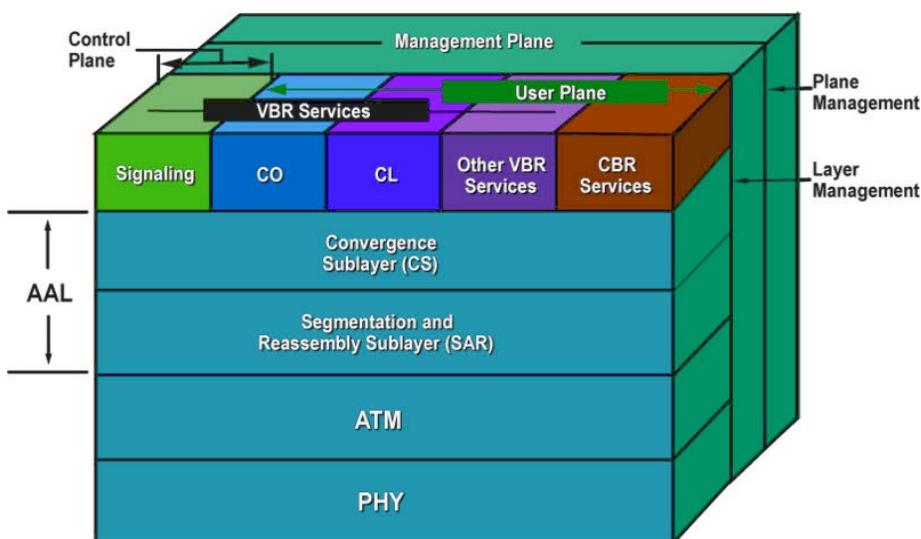
- ATM is connection-oriented.
 - ◆ A connection (an ATM channel) must be established before any cells are sent.
 - ◆ Two levels of ATM connections:
 - ◆ Virtual path connections.
 - ◆ Virtual channel connections.
 - ◆ Indicated by two fields in the cell header:
 - ◆ Virtual Path Identifier: VPI.
 - ◆ Virtual Channel Identifier: VCI.
- Switching based on VPI/VCI.

Port in	VPI/VCI	Port out	VPI/VCI
1	1/1	2	2/4
1	1/2	2	3/3
1	4/1	3	5/1
1	4/2	3	5/2



ATM Adaptation Layer (AAL)

- AAL is responsible for providing specific transport services to the higher layer protocols.
- The AAL is divided into:
 - ◆ Convergence Sublayer (CS) - manages the flow of data to and from SAR sublayer.
 - ◆ Segmentation and Reassembly Sublayer (SAR) - breaks data into cells at the sender and reassembles cells into larger data units at the receiver.
- ITU-T has defined four AAL service classes based on combinations of these three characteristics
 - ◆ Class A is a constant bit rate (CBR), delay-sensitive, connection-oriented service or a circuit emulation service.
 - ◆ Class B is a variable bit rate (VBR) service requiring time synchronization between sender and receiver (e.g., real-time compressed audio and video).
 - ◆ Classes C and D are delay-insensitive VBR services.
- Four AAL protocol types were defined to support the four service classes.
 - ◆ AAL 1 and AAL 5; And not in use anymore: AAL 2 and AAL 3/4.
 - ◆ Each type describes the format of the SAR-PDU (or the cell Payload field) and related operational procedures.

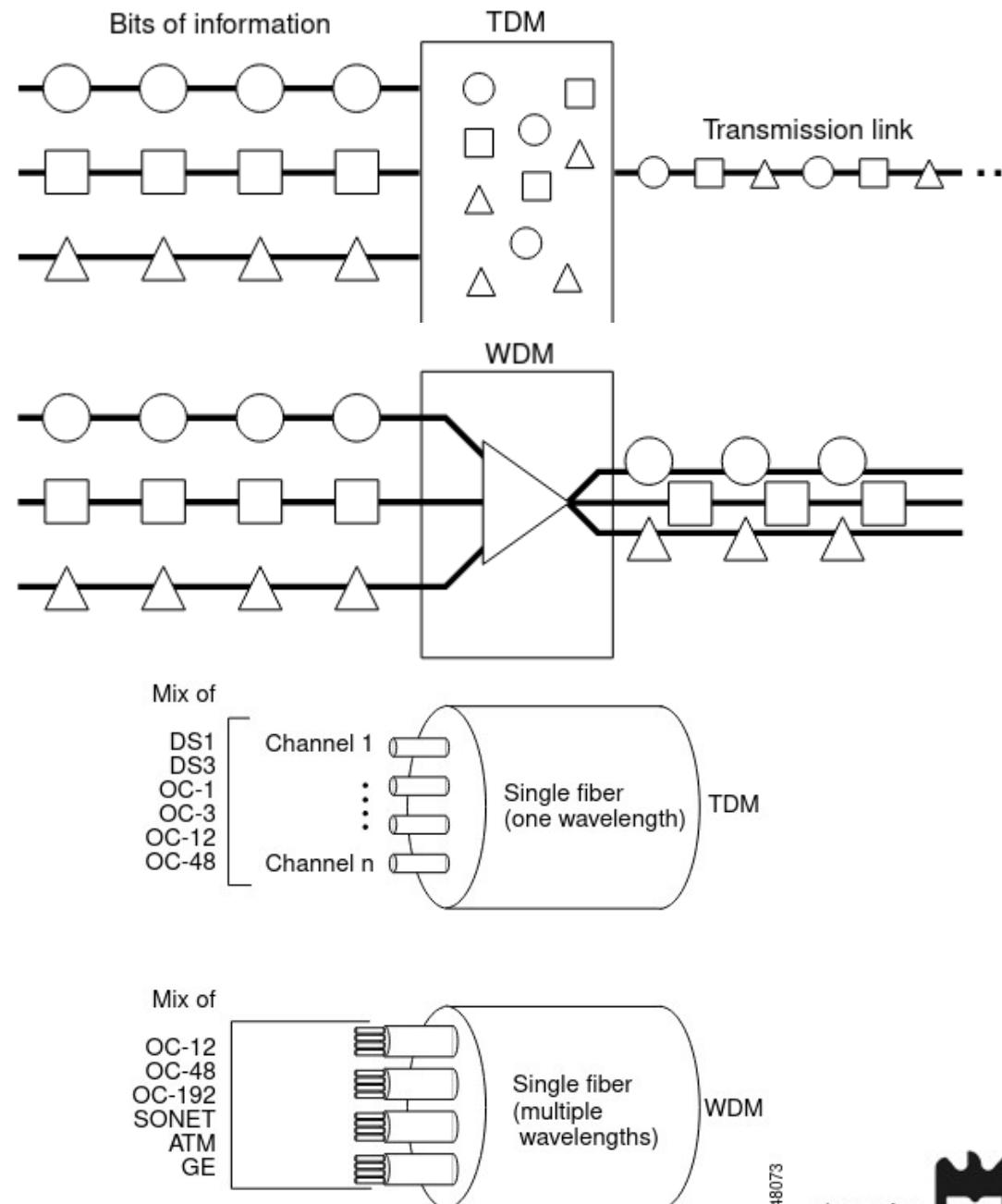


Service Class	A	B	C	D
Connection Mode	Connection-Oriented			Connectionless
Bit Rate	Constant	Variable		
End-to-End Timing Relationship	Required		Not Required	
Users	Circuit Emulation (e.g., Voice)	Packet Video and Compressed Voice	Connection-Oriented Data (e.g., Frame Relay)	Connectionless Data (e.g., SMDS, IP)
Suggested AAL Type	1	2	3/4, 5	3/4, 5



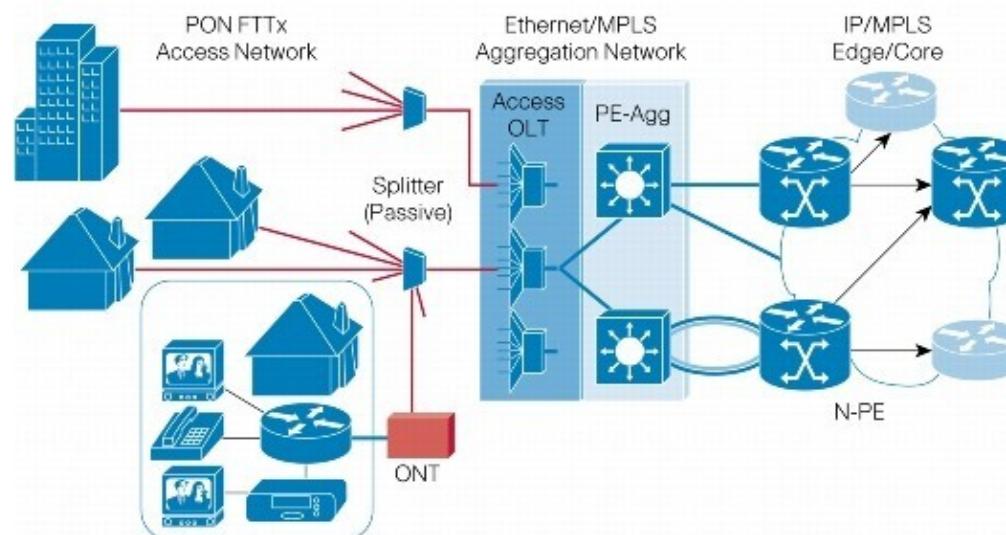
TDM, WDM and DWDM

- Time-division multiplexing (TDM).
 - ◆ E.g., SONET/SDH.
- Wavelength Division Multiplexing (WDM).
- Dense Wavelength Division Multiplexing (DWDM)
 - ◆ Optical fiber multiplexing technology that is used to increase the bandwidth of existing fiber networks.
 - ◆ Supports a higher number of wavelengths over the optical fiber.
 - ◆ Is a physical layer architecture, it can transparently support both TDM and data formats such as ATM, Gigabit Ethernet, etc...



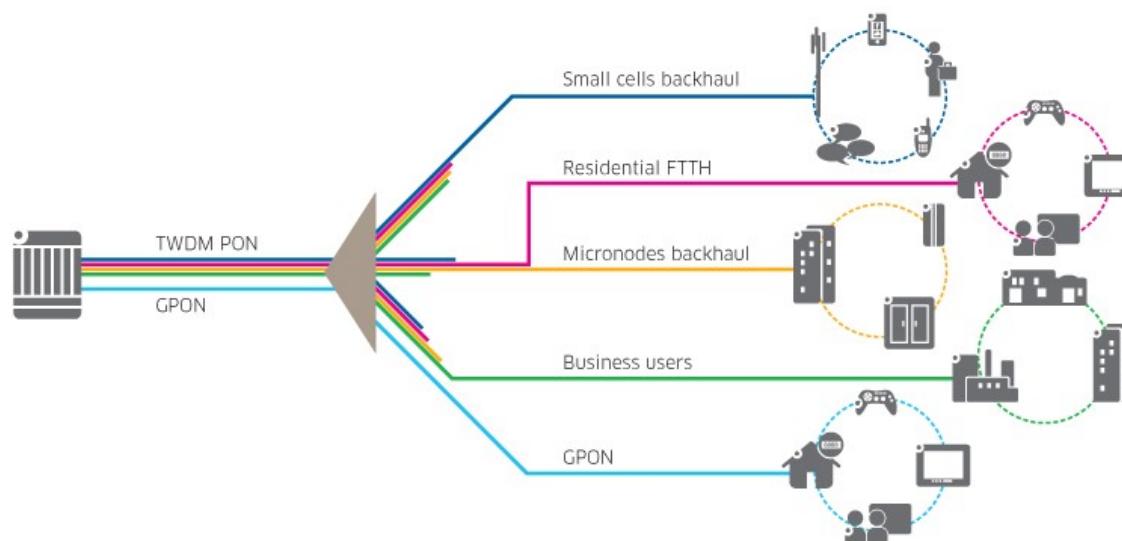
Passive Optical Network (PON)

- Is "passive" since it uses unpowered splitters to route data sent from a central location to multiple destinations.
- Based on TDM transmission.
- Variants
 - ◆ GPON - a "gigabit-capable PON" that supports 2.488 Gbps downstream and 1.244 Gbps upstream; follows the ITU G.984 standard.
 - ◆ EPON - the most popular PON implementation; transmits data as Ethernet frames at up to 10 Gbps downstream and upstream; also known as GEPON or the IEEE 802.3 standard.
- Focused on fiber connectivity to the home and other types of final network users (hotels, hospitals, and high-density residential buildings).



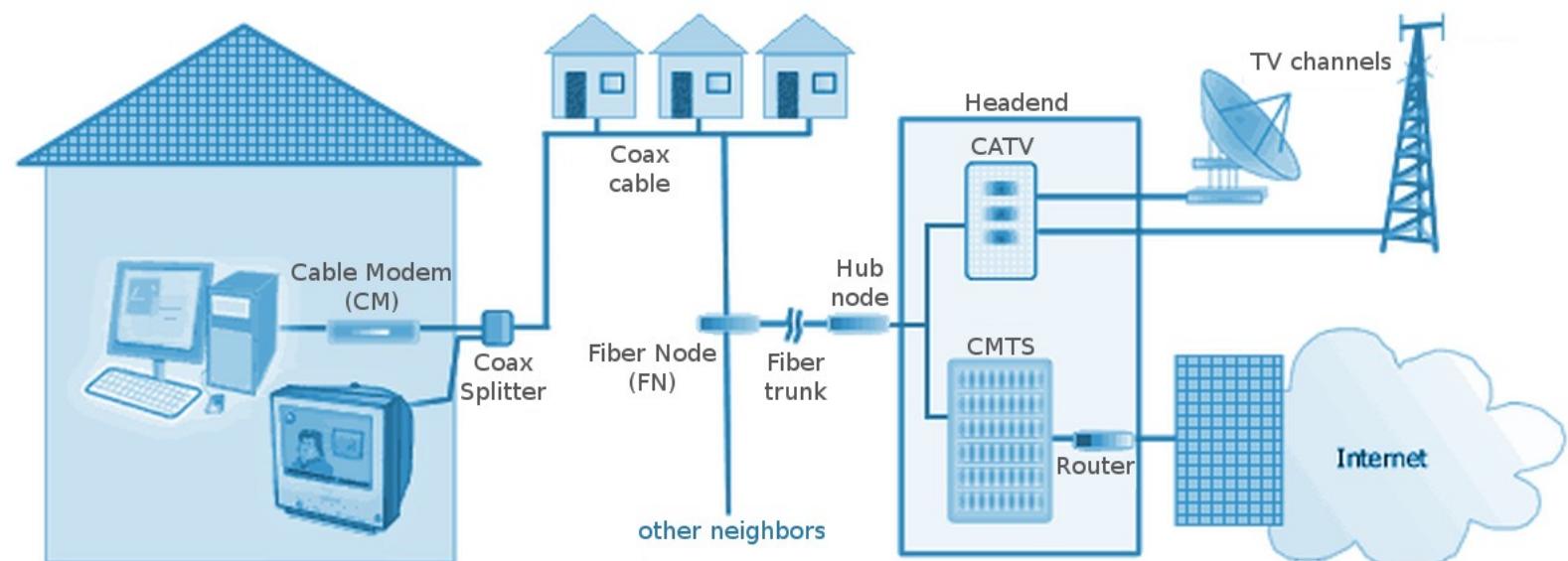
DWDM-PON and TWDM-PON

- The evolution of PON technologies are:
 - Time and Wavelength Division Multiplexed Passive Optical Network (TWDM-PON).
 - Dense Wavelength Division Multiplexed Passive Optical Network (DWDM-PON).
 - Adds flexibility by supporting the overlay of multiple services, user groups or organizations on the same fiber.
 - Can coexist with, and expand on, current PON deployments.
 - Ensures that operators' investments will keep providing value in the long term.



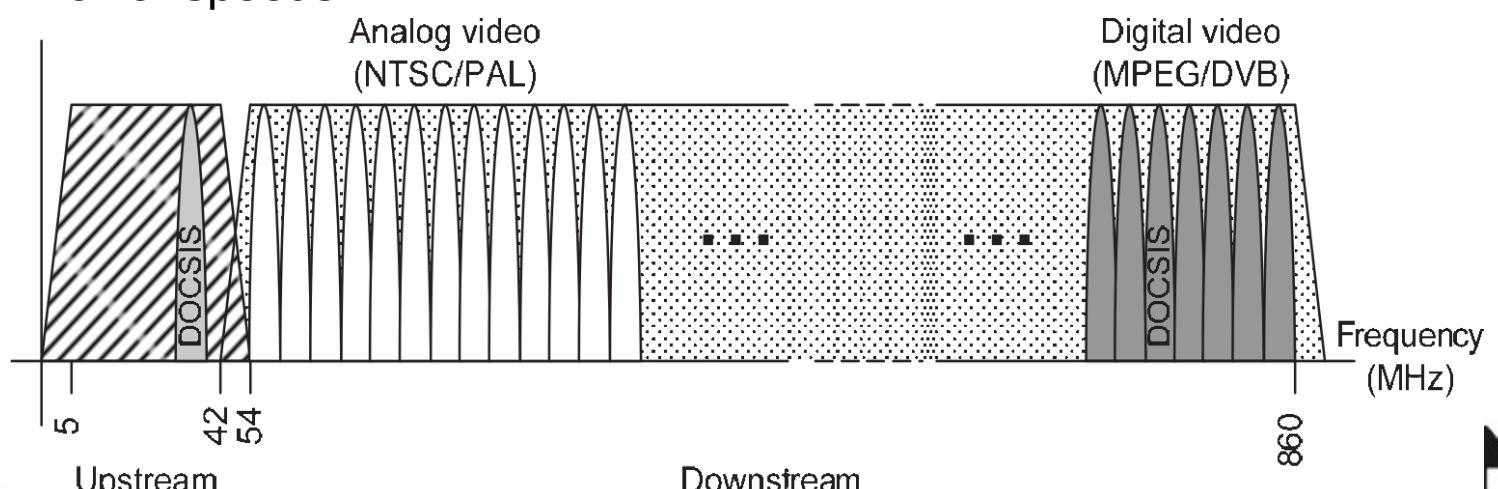
Community Access Television (CATV) “Cable” TV

- Hybrid Fiber Access (HCF)
 - ◆ Copper/Coax from Fiber Nodes and clients.
 - ◆ Fiber core (FN-Hub + Hub-Hub + some Hubs-Headend).
- Cable Modem (CM)
- Cable Modem Terminating System (CMTS)
- Fiber Node (FN)



Data Over Cable Service Interface Specification (DOCSIS)

- Versions 1.0 and 1.1
 - ◆ D/U: up to 50Mbps/9Mbps. Speed in Europe, 8MHz channels.
- Version 2.0
 - ◆ Adds A-TDMA which is a direct extension of the DOCSIS 1.x concepts and new synchronous CDMA (S-CDMA) → Upstream speed improvement.
 - ◆ D/U: up to 50Mbps/27Mbps.
- Version 3.0
 - ◆ Adds bonding of individual physical channels.
 - ◆ Using 4 channels – D/U: up to 200Mbps/108Mbps.
 - ◆ Using 8 channels – D/U: up to 400Mbps/108Mbps.
- In US, 6MHz channels → Lower speeds.
- Spectrum allocation



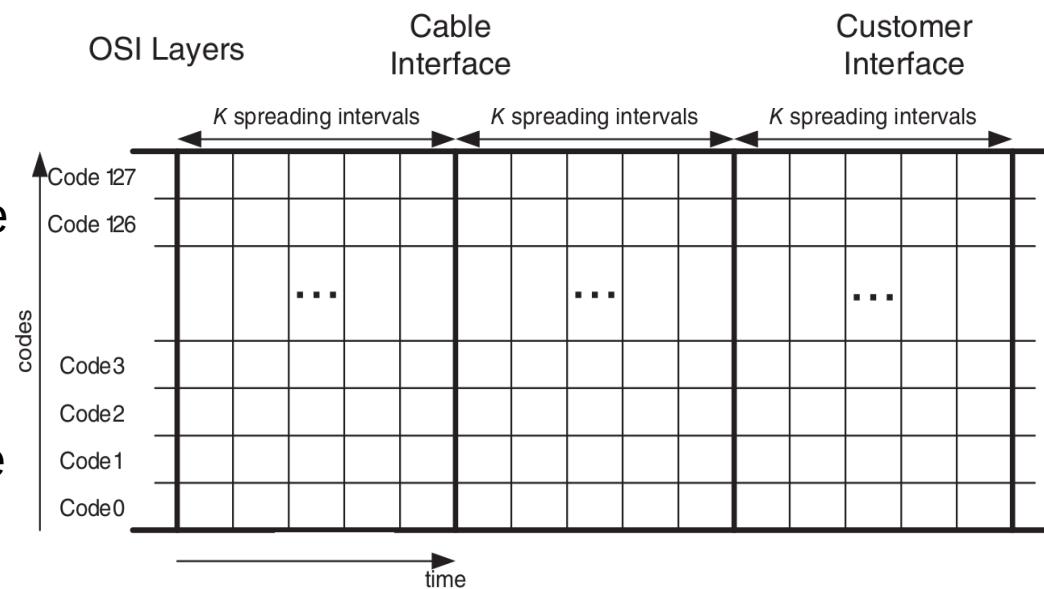
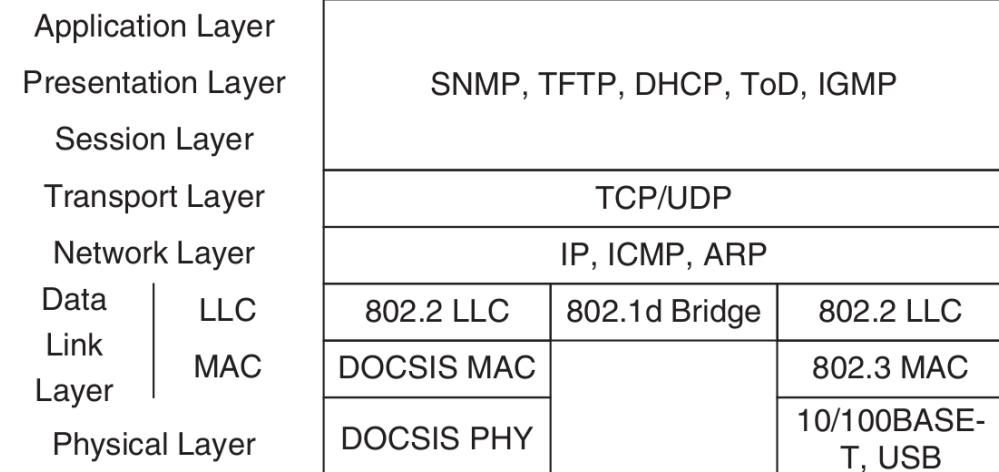
DOCSIS Transmission

- Downstream Transmission

- Digital Video Broadcast (DVB) standards.
- Signal is a continuous stream of 188-byte long MPEG packets, that contain:
 - An MPEG video payload, or
 - a DOCSIS MAC payload.

- Upstream Transmission

- TDMA Transmission Mode (from version 1.0).
 - Transmissions are separated only in time.
- Synchronous CDMA (S-CDMA) Transmission Mode (from version 2.0).
 - Transmissions are separated by both time and CDMA spreading code.
 - CDMA - Code division multiple access.
 - DOCSIS MAC payloads transmitted in one or more mini-slots (time/code divided mini slots).



Mobile Networks

• 2G:

- ◆ GSM (Global System for Mobile)
- ◆ GSM Packet Radio System (GPRS)
- ◆ Enhanced Data-rates for GSM Evolution (EDGE)
- ◆ Based on TDMA

• 3G:

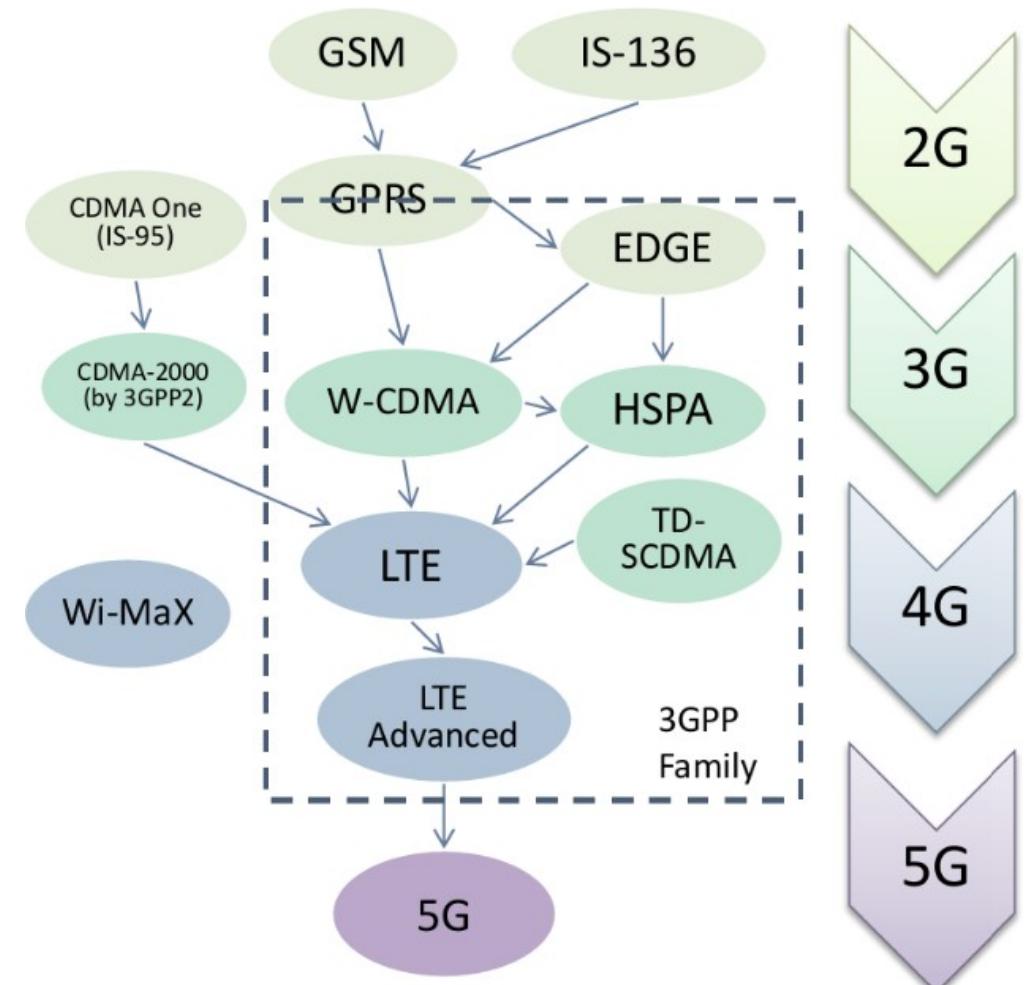
- ◆ Universal Mobile Telecommunication System (UMTS)
- ◆ Based on Wideband-CDMA (W-CDMA)
- ◆ High Speed Packet Access (HSPA)
- ◆ High-Speed Downlink Packet Access (HSDPA)
- ◆ High-Speed Uplink Packet Access (HSUPA)
- ◆ CDMA2000

• 4G:

- ◆ LTE
- ◆ LTE-Advanced
- ◆ IEEE 802.16e (WiMax) and IEEE 802.16m
- ◆ Based on OFDMA and MIMO

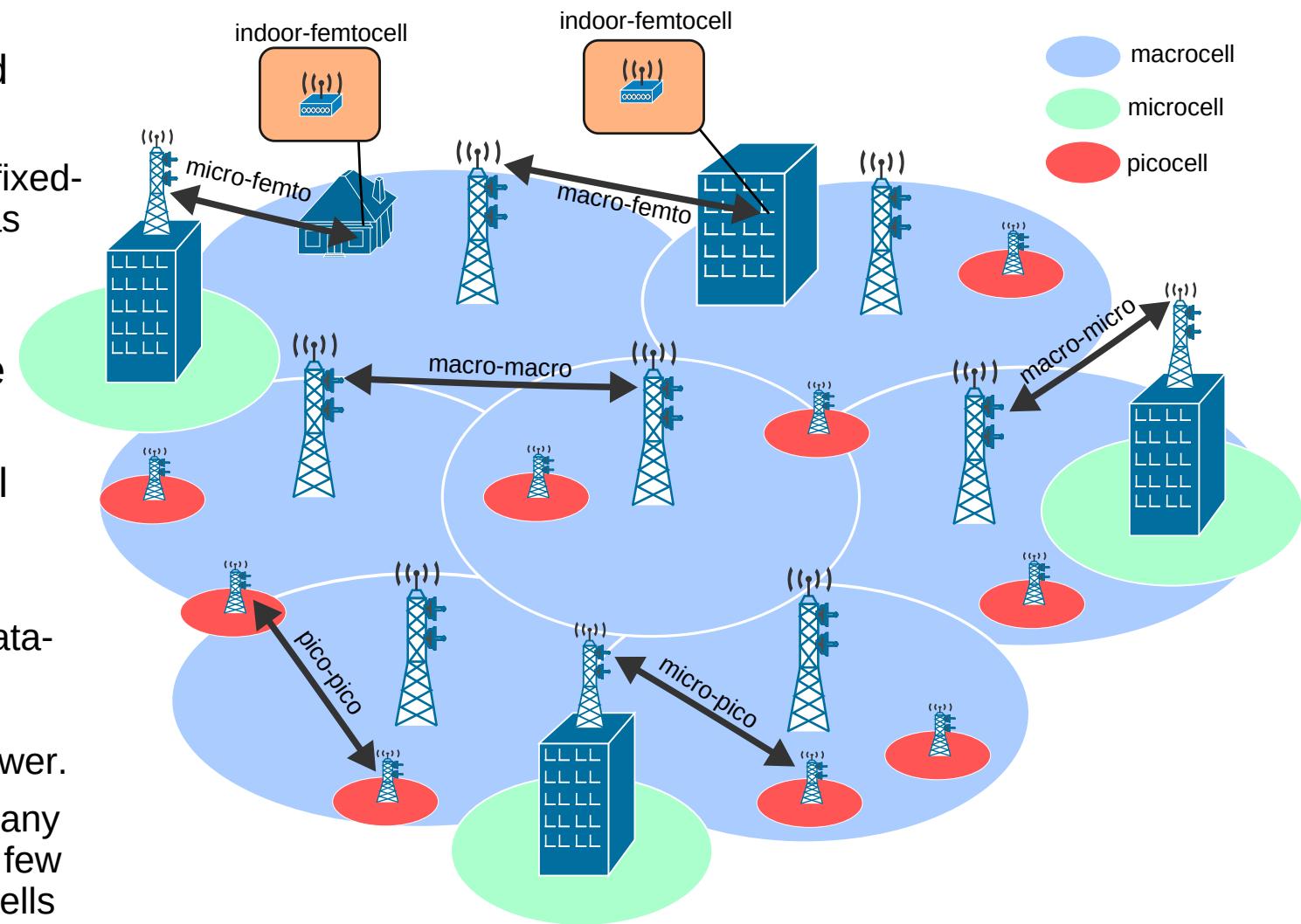
• 5G:

- ◆ Based on MIMO
- ◆ Small cells
- ◆ NFV Core
- ◆ Integrated Wired and Wireless IP networks

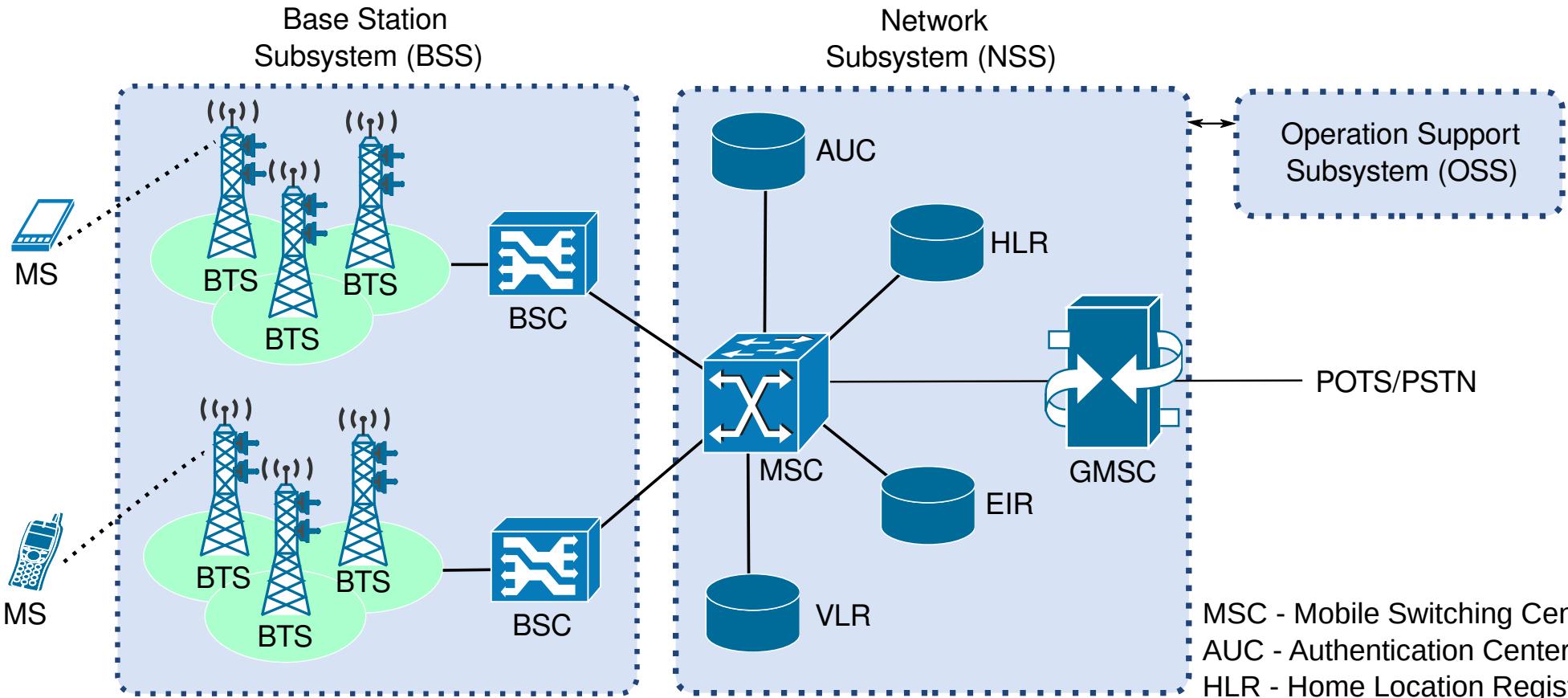


Cellular Network Concept

- Concept used on Public Land Mobile Networks (PLNM).
- Network is distributed over land areas called cells.
 - ◆ Each served by at least one fixed-location transceiver, known as base station.
- Macrocells are mainly used to provide a widespread coverage area.
- Smaller micro, pico or femtocell structures can be used for high data-rate.
 - ◆ Able to sustain high speed data-traffic by reducing the propagation distance, hence reducing the transmission power.
 - ◆ Micro/picocells can handle many devices within the range of a few hundred meters while femtocells are mostly used for indoor or home area.



Global System for Mobile (GSM)



BTS - Base Transceiver Station

BSC - Base Station Controller

MS - Mobile Station

MSC - Mobile Switching Center

AUC - Authentication Center

HLR - Home Location Register

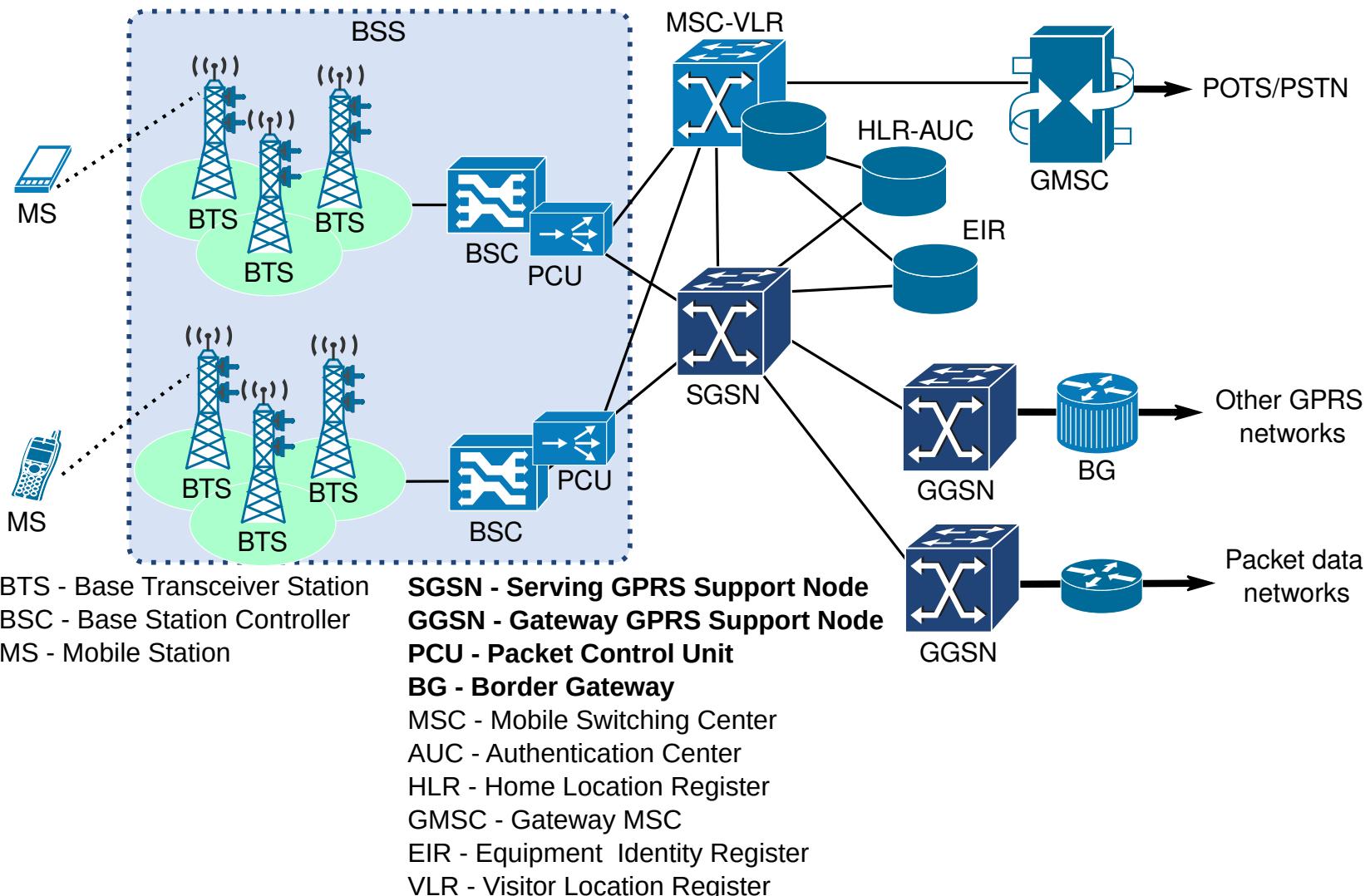
GMSC - Gateway MSC

EIR - Equipment Identity Register

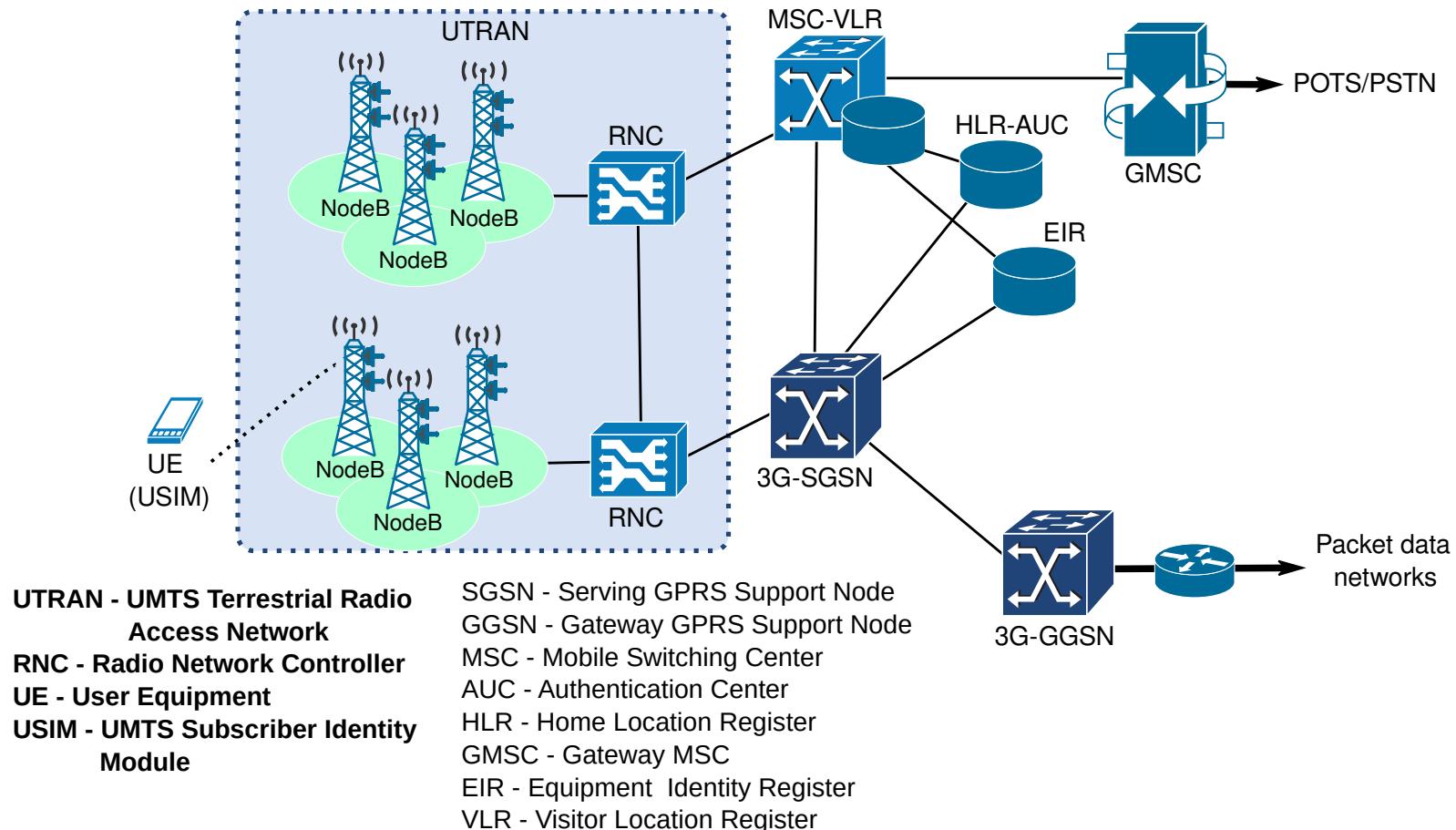
VLR - Visitor Location Register



GSM Packet Radio System (GPRS)



Universal Mobile Telecommunication System (UMTS)



- 3rd Generation Partnership Project (3GPP) standard.
- Novel radio access network called Universal Terrestrial Radio Access Network (UTRAN)
- Core network remains largely unchanged from GPRS/EDGE.

High Speed Packet Access (HSPA)

- Upgrade to W-CDMA networks to provide **higher bit rates** and **lower delays**.
- High-Speed Downlink Packet Access (HSDPA)
 - ◆ To be able to make faster decisions on radio channel allocation (adapting to varying channel quality) and reduces delays, new functions were added closer to the radio interface (NodeB):
 - ◆ Scheduling, select which UE(s) is/are to use the radio resources at each Transmission Time Interval (TTI), where one TTI is 2 ms.
 - ◆ Link adaptation, setting of channel coding rate and modulation (QPSK or 16QAM), in order to utilize the resources effectively.
- High-Speed Uplink Packet Access (HSUPA)
 - ◆ Uses a packet scheduler that operates on a request-grant principle where the UEs request a permission to send data and the scheduler decides when and how many UEs will be allowed to do so.
 - ◆ However, unlike HSDPA, uplink transmissions are not orthogonal to each other.
- Evolved High Speed Packet Access (HSPA+)
 - ◆ Further increase bit rates.
 - ◆ New functions are added:
 - ◆ Higher order modulation 64QAM (DL) and 16QAM (UL),
 - ◆ Multiple Input Multiple Output (MIMO) used only in the DL.

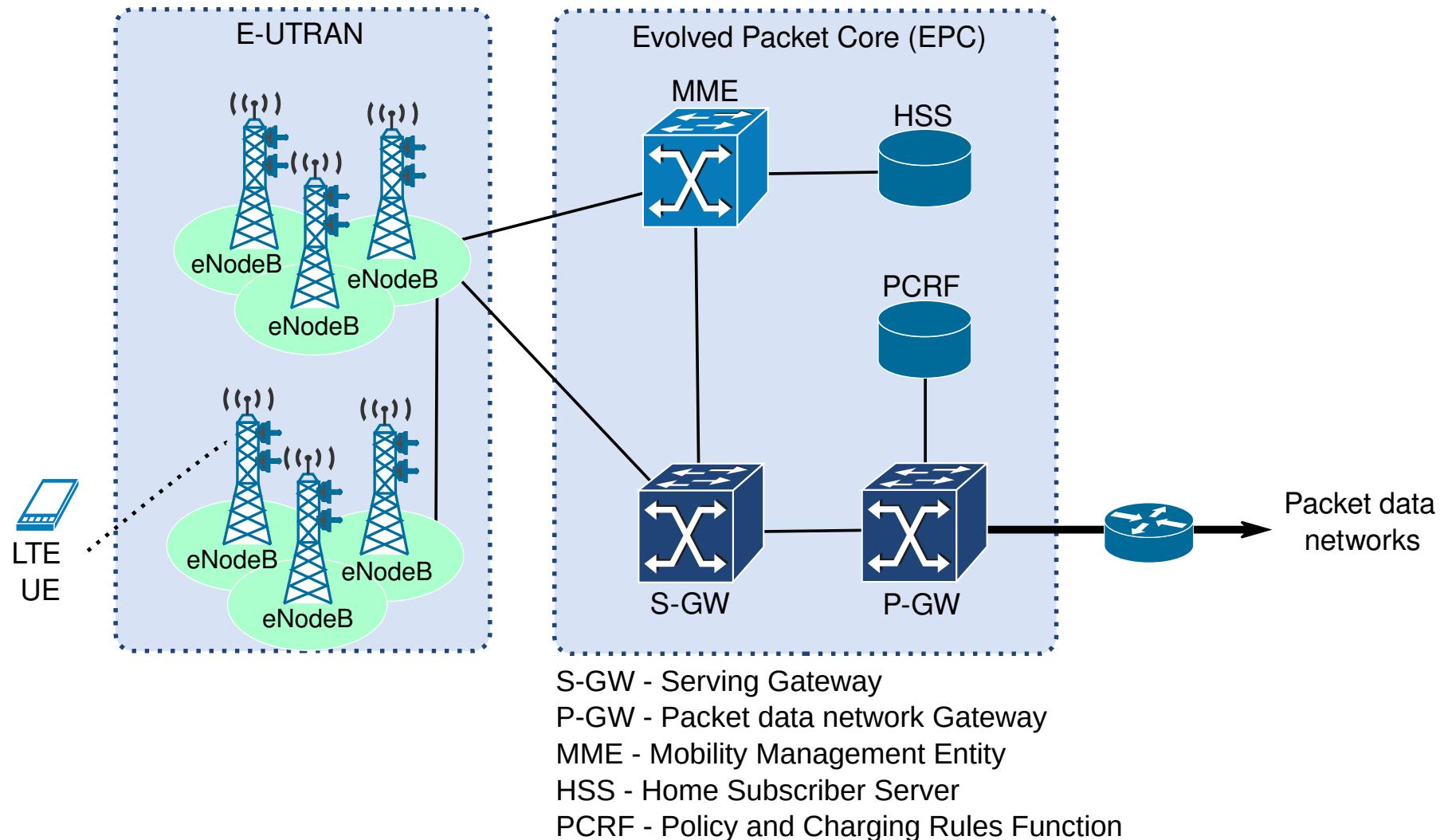


Long Term Evolution (LTE)

- LTE standard has been developed by 3GPP
 - ◆ Extension of UMTS (based on 3GPP standard)
 - ◆ and CDMA200 1xEV-DO (based on 3GPP2 standard).
- Designed for high speed data applications both in the uplink and downlink.
 - ◆ Offers about 300Mbps data rate in the downlink and about 75 Mbps in the uplink.
- LTE is an all IP based network, supporting both IPv4 and IPv6.
 - ◆ Possibility of supporting voice over LTE (VoLTE).
- Uses a different form of radio interface from UMTS.
 - ◆ Instead of CDMA it uses OFDMA (Orthogonal Frequency Division Multiple Access is used in the downlink; and SC-FDMA(Single Carrier - Frequency Division Multiple Access) is used in the uplink.
- Uses MIMO (Multiple Input Multiple Output).
 - ◆ Requires the use of multiple antennas (antenna matrices).
- LTE has been defined to accommodate both FDD and TDD operation.



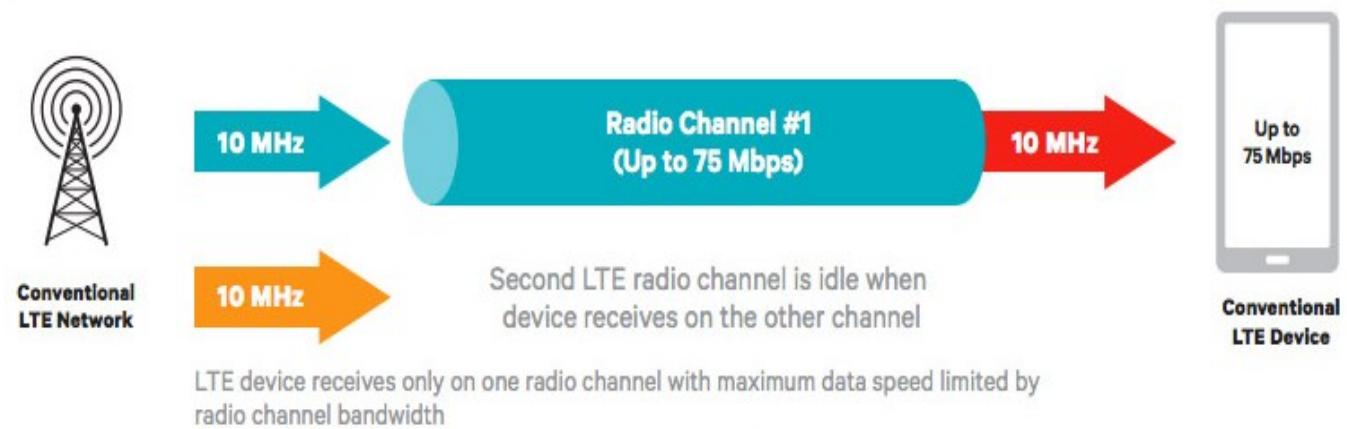
Long Term Evolution (LTE)



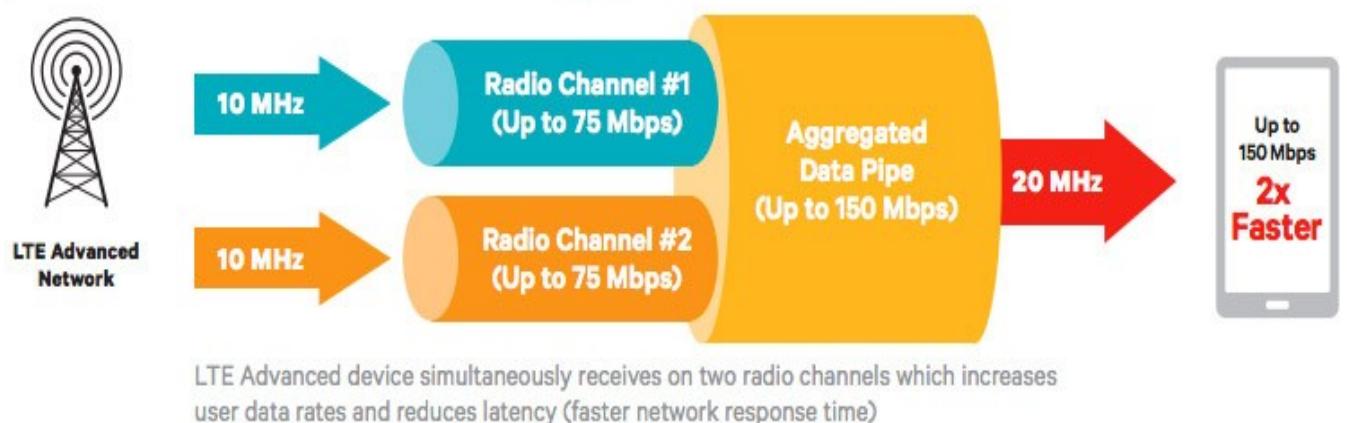
LTE-Advanced

- LTE-Advanced is the upgraded version of LTE.
 - ◆ Increases the peak data rates to about 1GBPS in the downlink and 500MBPS in the uplink.
- Utilizes higher number of antennas and added carrier aggregation feature.
 - ◆ Carrier aggregation can be used for both FDD and TDD.

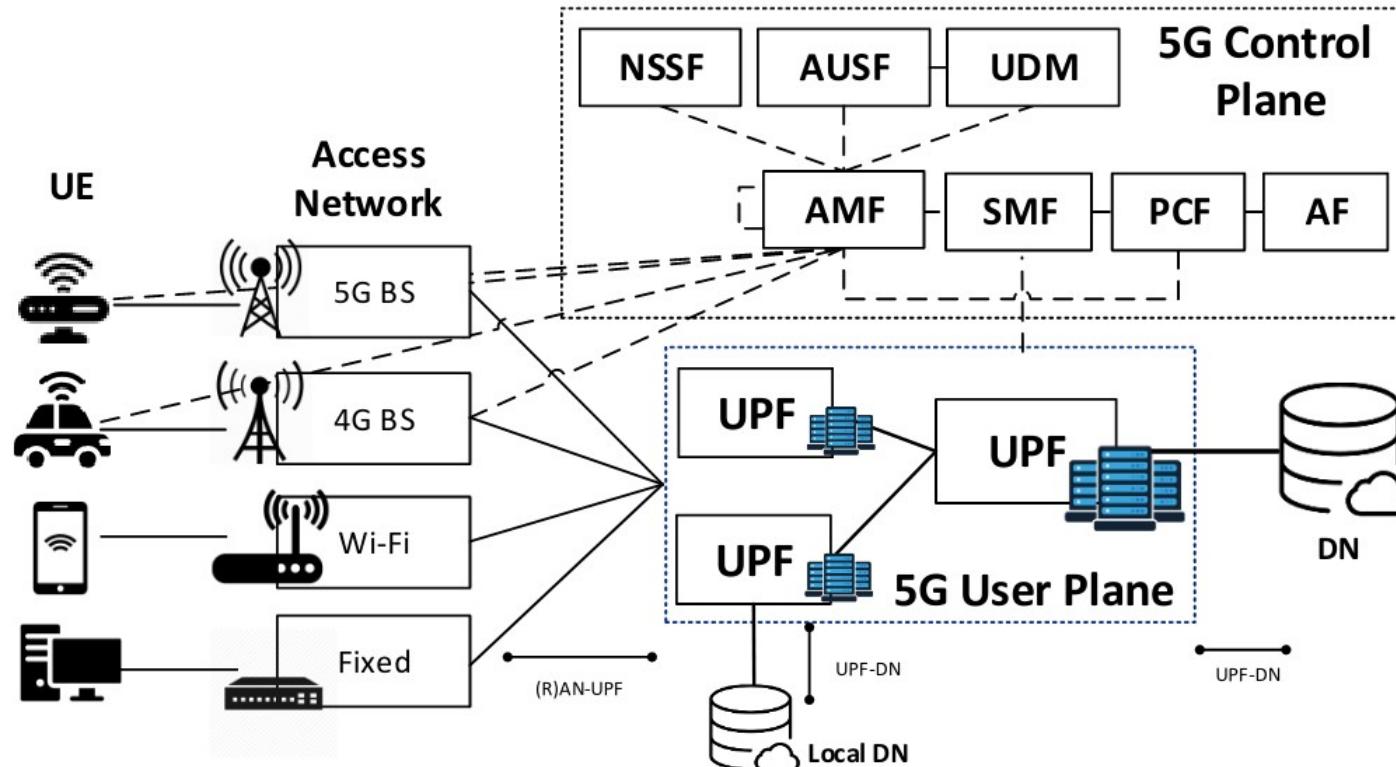
Conventional LTE Network: Single channel approach to data transfer



LTE Advanced Network: Carrier Aggregation effectively doubles data rates



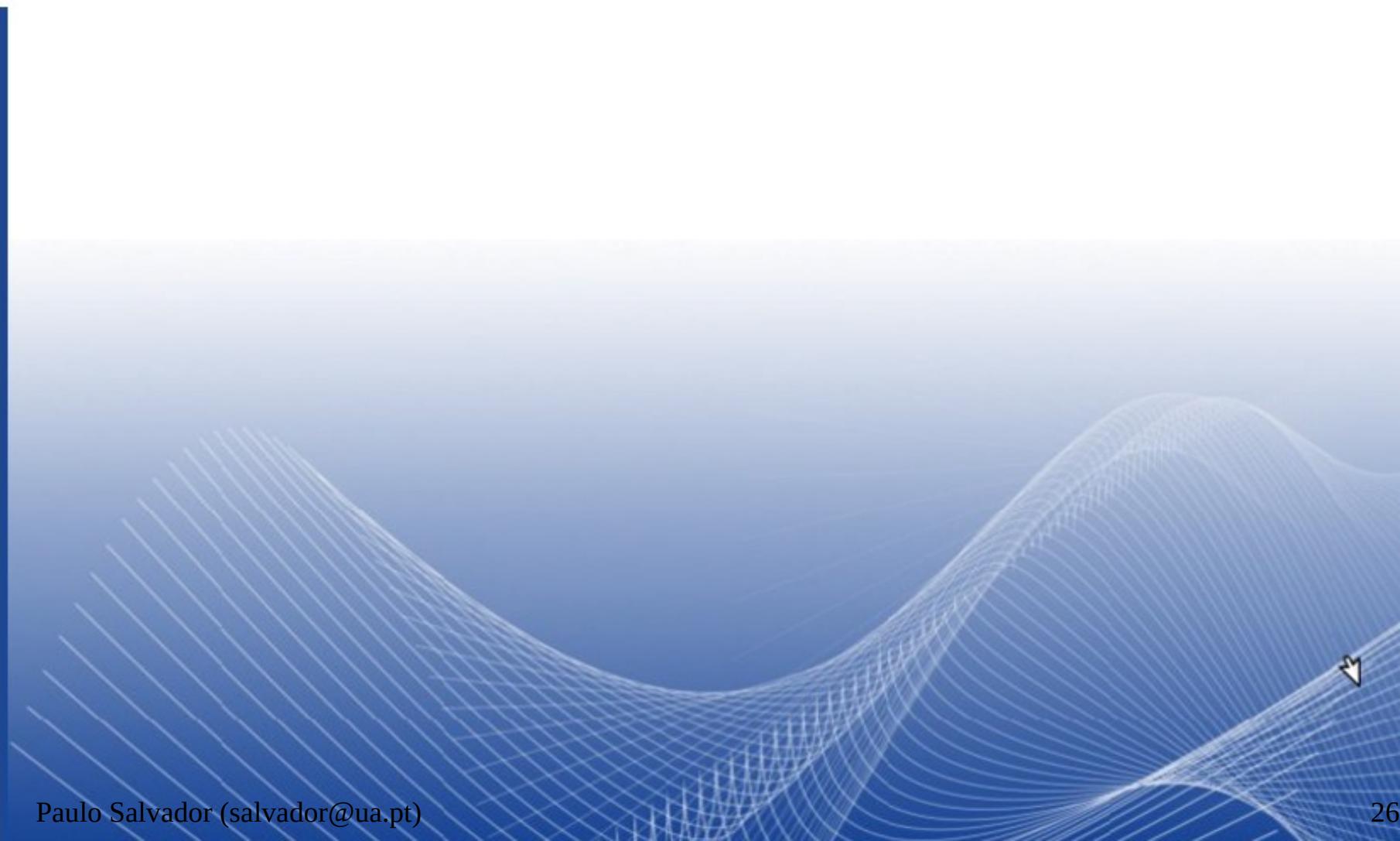
5G



- Architecture incorporates
 - ◆ Network Function Virtualization (NFV) at the core,
 - ◆ Edge Computing (EC),
 - ◆ Software Defined Networks (SDNs).
- Uses a high frequency range (30 GHz and 300 GHz) of the radio spectrum,
 - ◆ Higher frequency → Higher bandwidth, Lower range → Smaller cells.
- Integrated Wired and Wireless IP networks.

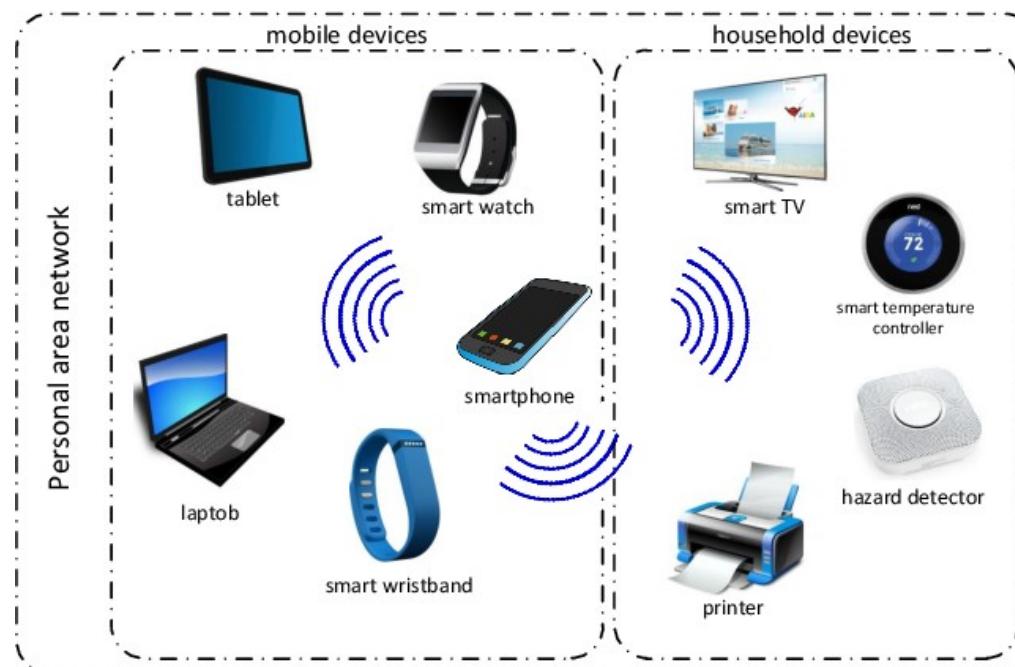


WPAN and Sensor Networks



Wireless Personal Area Network (WPAN)

- Span a small area (e.g., a private home or an individual workspace)
 - ◆ Communicate over a short distance.
 - ◆ Low-powered communication.
 - ◆ Primarily uses ad-hoc networking.
 - ◆ Could be wireless or wired.



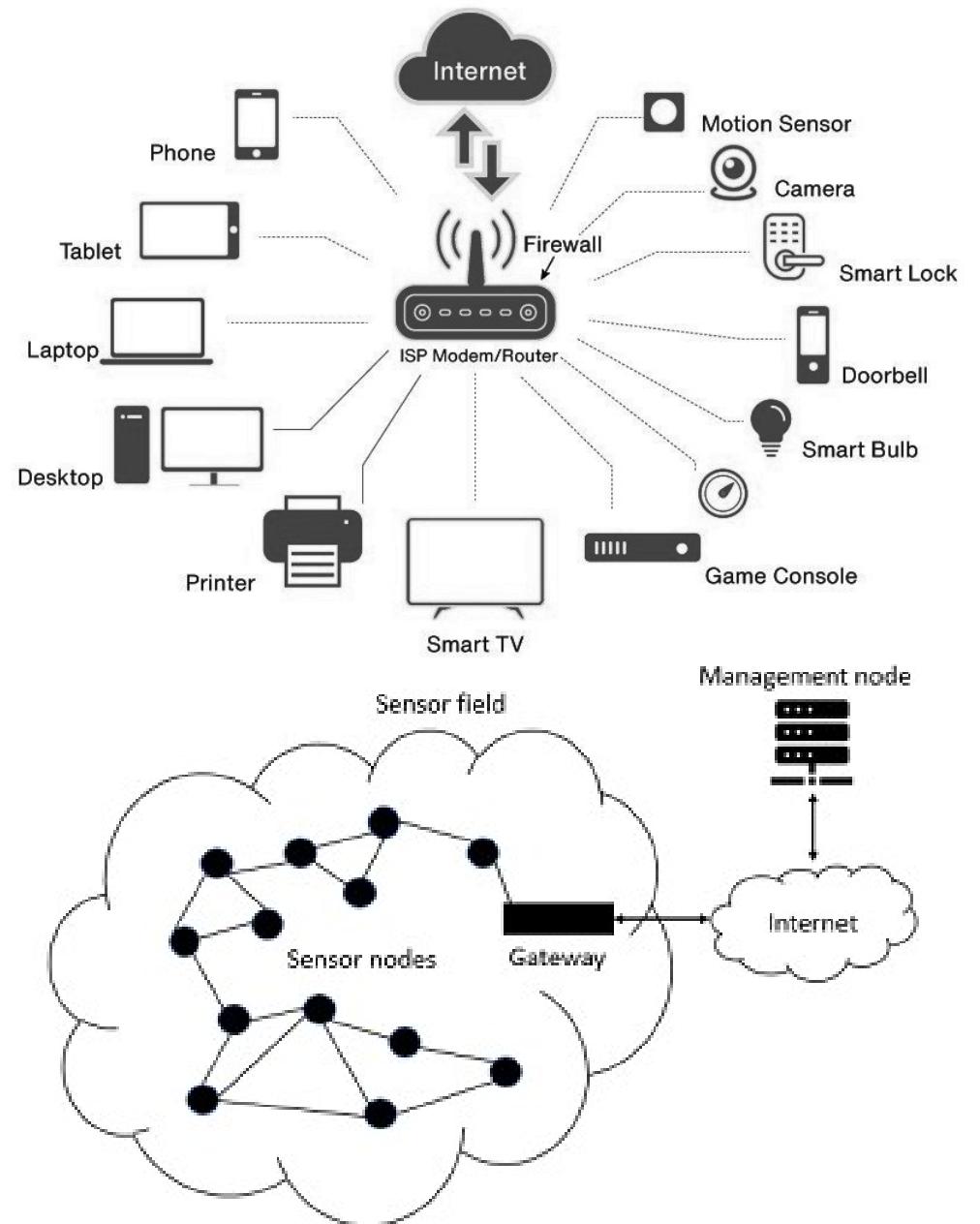
Internet of Things / Sensor Networks

- Composed by small and medium devices

- Usually battery powered.
- May not allow battery replacement.
- Low computational resources.

- Network requirements

- Simplicity
 - Easy to deploy and with low computational requirements.
 - Low cost devices.
- Security
 - Node access should be controlled.
 - Data should be encrypted.
- Reliability
 - Limited failures and integrated recovery features.
- Efficiency (low-power)
 - Battery life should be measured in months or years.
- Scalability
 - Should support an high number of connected devices.



IEEE 802.15

- Standard for low-data-rate physical and medium access control layer specifications for wireless personal area networks (WPAN).
- Evolved over time:
 - ◆ IEEE 802.15.4-2003 ; IEEE 802.15.4-2006, IEEE 802.15.4-2011 IEEE 802.15.4-2015.
- IEEE 802.15.4 is a wireless access technology for
 - ◆ Low-cost and low-data-rate devices.
 - ◆ Devices powered by batteries.
 - ◆ Enables easy installation using a compact protocol stack.
 - ◆ Several network communication stacks use this technology in both the consumer and business markets.



Communication Standards

- Wi-Fi

- Range: ~50 meters
- Data Rate: 23-144Mbps
- Frequency: 2.4GHz/5GHz
- Max. Devices: 250



- Bluetooth

- Range: 10 meters (class 2/3), 100 meters (class 1)
- Data Rate: 1-3Mbps
- Frequency: 2.4GHz
- Max. Devices: 7



- Zigbee

- Range: 50-70 meters
- Data Rate: 20-250 kbps
- Frequency: 915MHz to 2.4GHz
- Max. Devices: ~1000 (realistically)

- Z-Wave

- Range: ~100 meters
- Data Rate: 100 kbps
- Frequency: 915MHz
- Max. Devices: 232

- Thread (newest and trending)

- Range: ~30 meters
- Data Rate: 250 kbps
- Frequency: 2.4GHz
- Max. Devices: 300

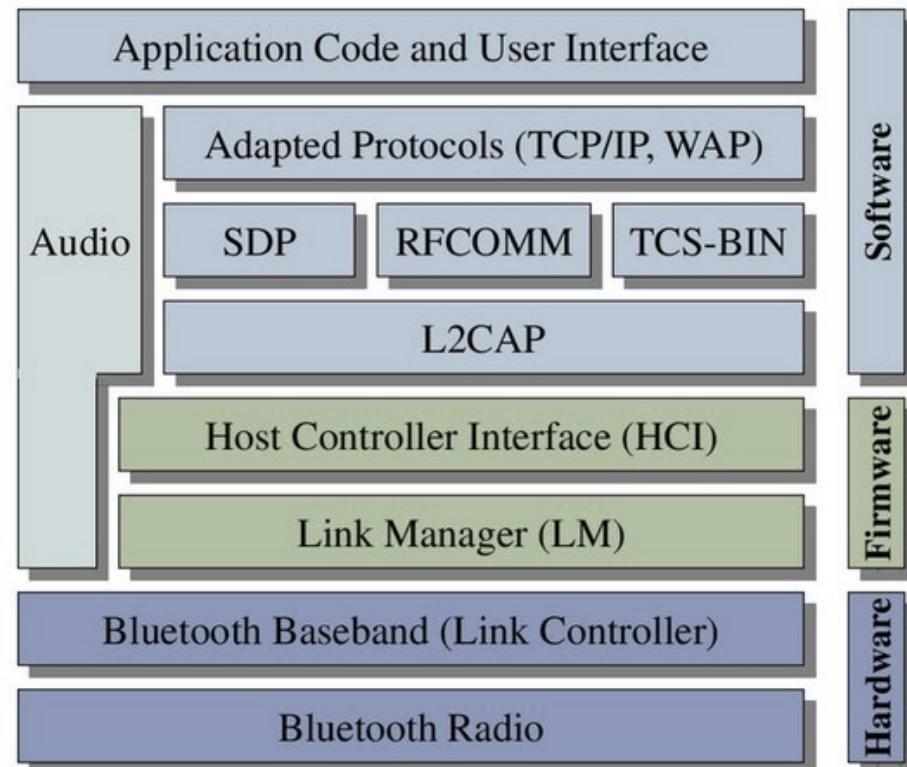
Wi-Fi

- Star topology.
- Current consumption: ~250mA (very high)
- Wi-Fi is an alternative only for always or frequently powered devices.



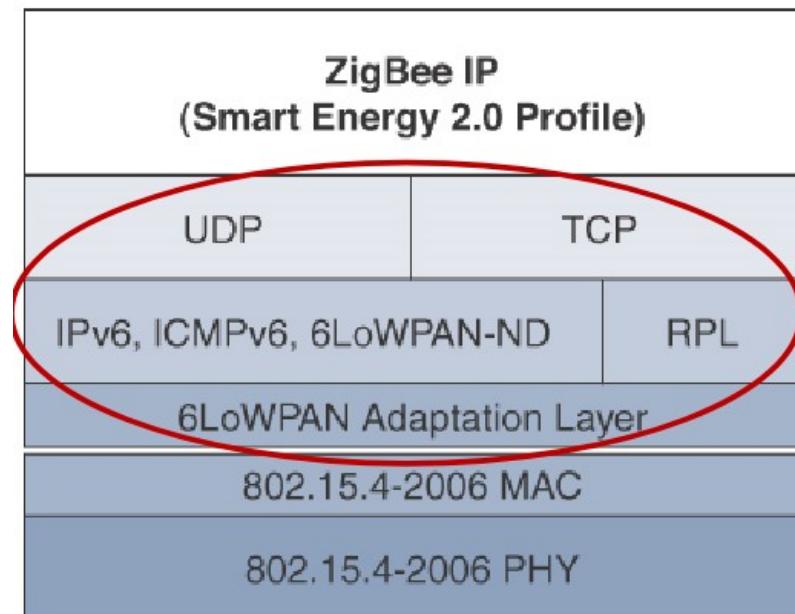
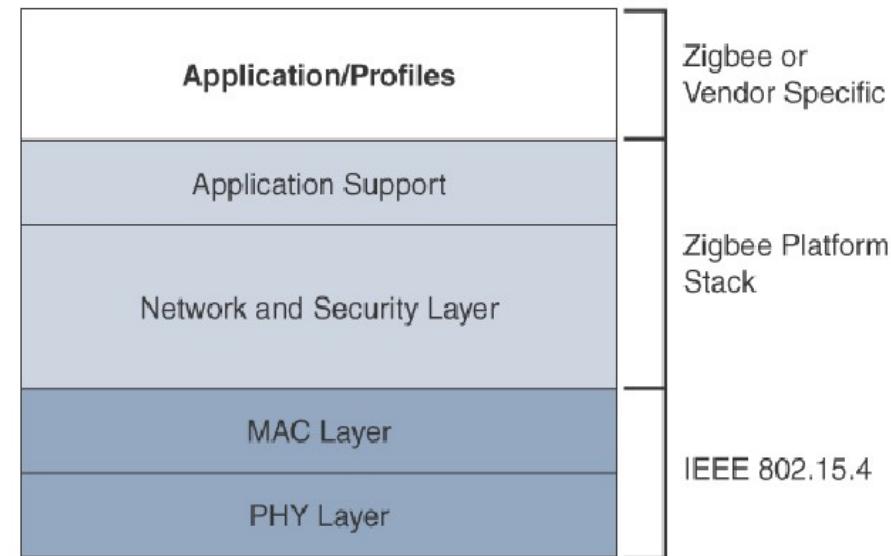
Bluetooth

- Mesh and Star topology.
- Current consumption:
 - ◆ Bluetooth: ~30mA.
 - ◆ Bluetooth LE: less than 15mA.
- Bluetooth has classes that define indicate the power output and wireless range of a device:
 - ◆ Class 1: 100 mW (20 dBm), 100 meter
 - ◆ Class 2: 2.5 mW (4 dBm), 10 meter
 - ◆ Class 3: 1 mW (0 dBm), 1 meter
- Bluetooth Low Energy (LE) is a power-conserving variant of PAN technology.
- Frequency Hopping Spread Spectrum



ZigBee

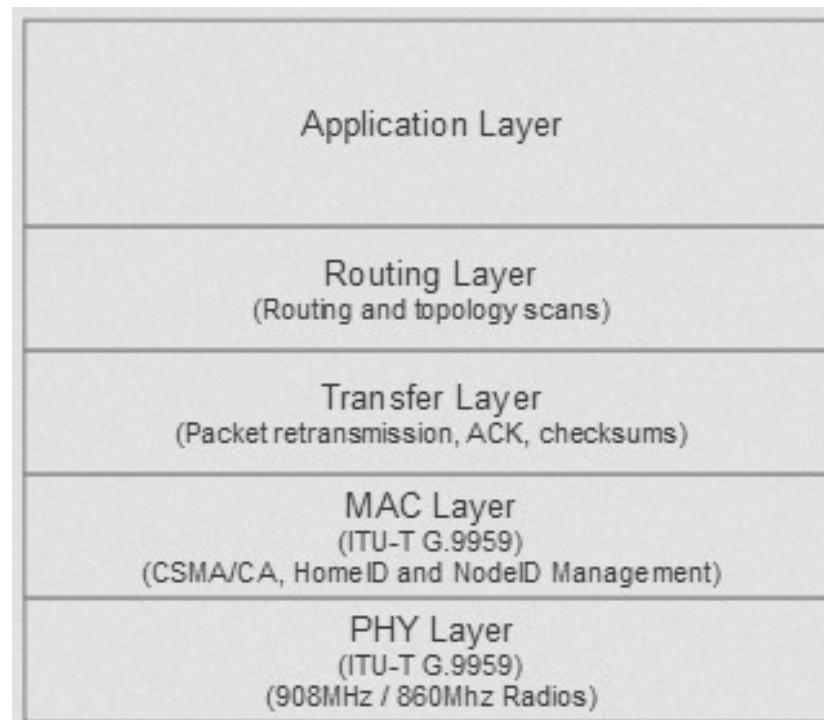
- Star, Tree or Mesh topology.
- Current consumption: ~50mA.
- ZigBee has not provided interoperability with other IoT solutions or open standards
- ZigBee IP was created to embrace the open standards at the network and transport layers
- Based on IEEE 802.15.4.
- And, based on 6LoWPAN
 - ◆ Defines encapsulation and header compression to send and receive IPv6 packets over IEEE 802.15.4 networks.





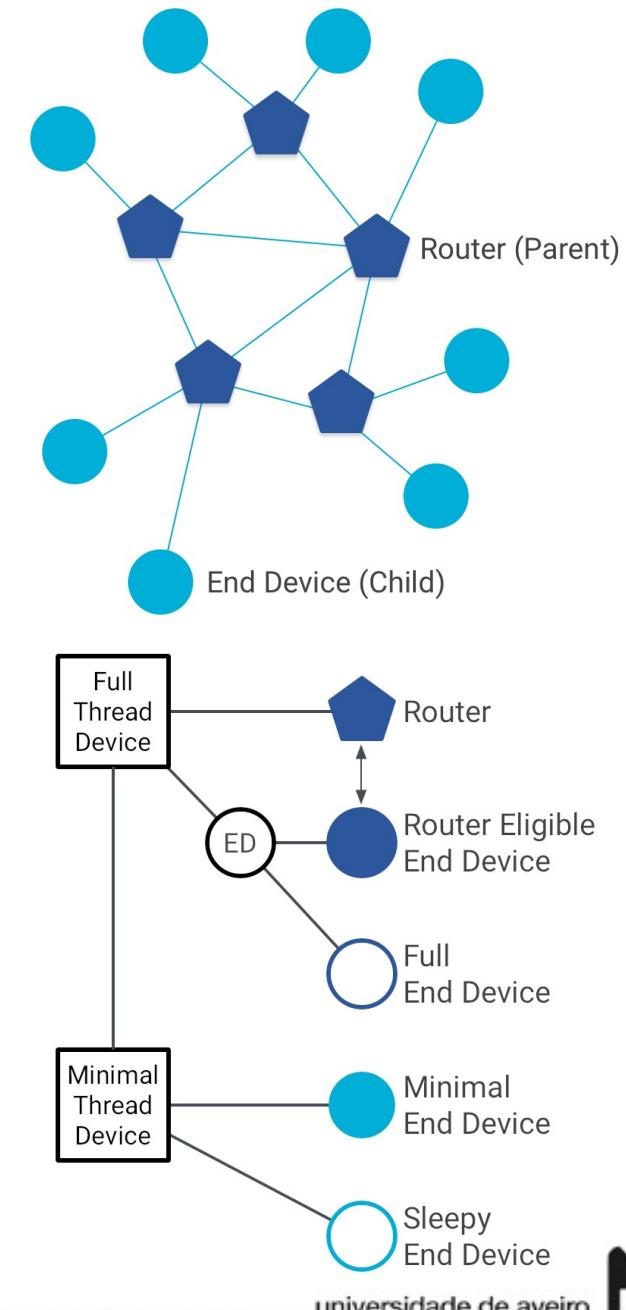
Z-Wave

- Mesh topology.
- Current consumption: ~2.5mA.
- Defined by ITU-T G.9959 Standard.
 - ◆ Closed standard until 2020.



Thread

- Defined for all but the application layer and all of the layers are pre-existing protocols.
 - At the physical and link layer, IEEE 802.15.4 protocol is used just like with ZigBee.
 - At the network and transport layers, Thread uses a combination of IPv6, 6LowPAN, UDP, and DTLS (Datagram Transport Layer Security).
- Mesh topology.
- IPv6-based networking protocol.
- Independent of other mesh networking protocols, such as ZigBee, Z-Wave, and Bluetooth LE.
- Nodes are split into two forwarding roles: router and end-device.
- Nodes comprise a number of types:
 - Full Thread Device - always has its radio on and subscribes to the all-routers multicast address
 - Router, Router Eligible End Device (REED), Full End Device (FED)
 - Minimal Thread Device - does not subscribe to the all-routers multicast address
 - Minimal End Device (MED) - radio always on
 - Sleepy End Device (SED) – radio normally disabled, wakes on occasion to poll for messages from its parent



Low Power Wide Area Network (LPWAN)

- Wireless telecommunication wide area network designed to allow long-range communications at a low bit rate, low power consumption and low cost.
- SigFox
 - ◆ Supports millions of end devices.
 - ◆ Proprietary.
 - ◆ Access infrastructure (built with operators) and software.
 - ◆ Open market for the endpoints.
 - ◆ 30-50km range in rural areas, and 3-10km range in urban areas.
 - ◆ Ultra narrow band, 868MHz (EU) or 902Mhz (US).
 - ◆ Low energy consumption.
 - ◆ Dedicated network.
- LoRaWAN
 - ◆ Stands for “Long Range”.
 - ◆ To be used in long-lived battery-powered devices scenarios.
 - ◆ Semi-proprietary
 - ◆ Parts of the protocol are well documented, others not
 - ◆ They own the radio part (but sub-licensing is on the way)
 - ◆ You can install your own gateways
 - ◆ LoRa usually means two different things:
 - ◆ LoRa: a physical layer that uses Chirp Spread Spectrum (CSS) modulation.
 - ◆ LoRaWAN: a MAC layer protocol.

