Backbone technologies

Mestrado em Engenharia de Computadores e Telemática

1º ano, 1º semestre, 2021/2022

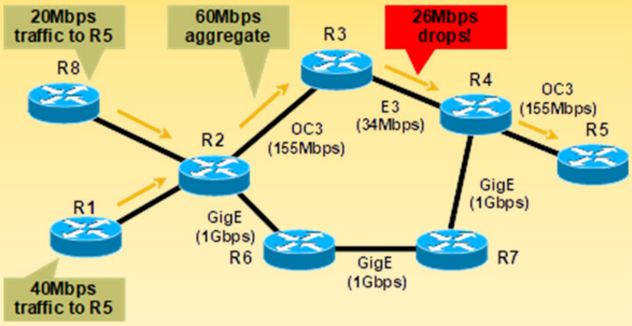
Traffic Engineering (TE)

- 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.

2

Example – avoiding congestion

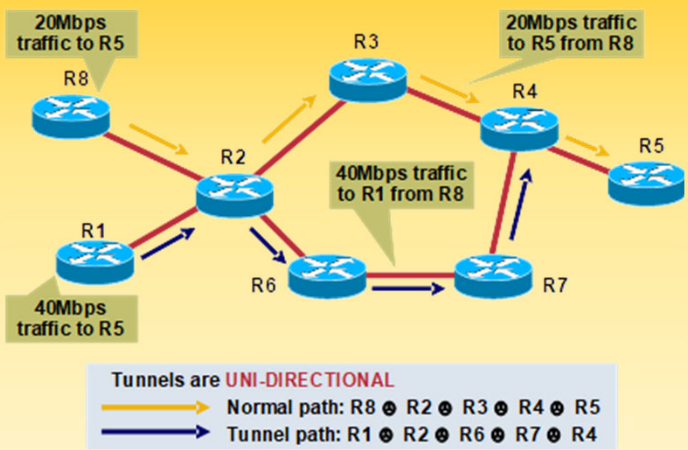
- On IP networks, *IntServ* and *DiffServ* are "routing independent architectures", retain the issues from routing
- IP network routing is based on the destination and does not allow to take the maximum possible advantage of the network resources
  - Shortest path will lead to congestion, even with available resources in the core
  - With **RIP or OSPF or ANY OTHER IGP** it is not possible to condition both flows.



3

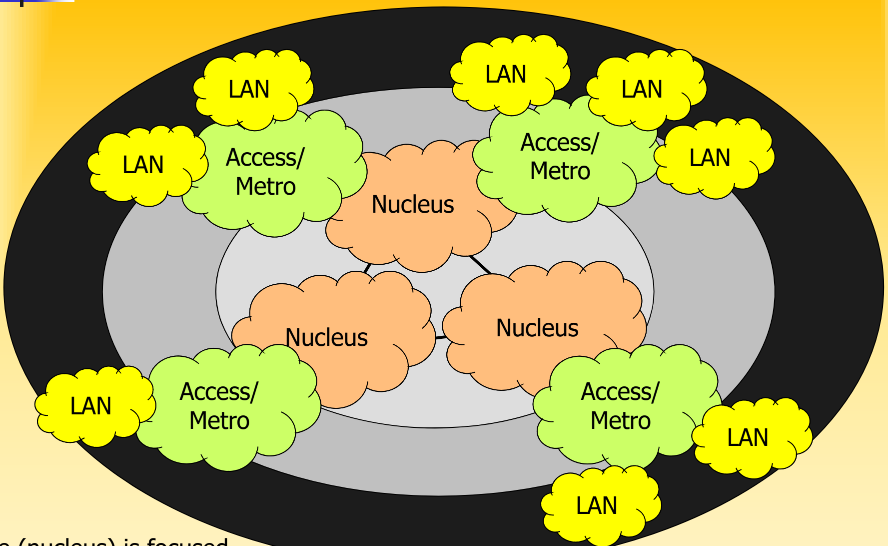
Using TE to solve… (source based routing)

- Tunnels (virtual entities) explore all capacity
  - Packets will transport, from their source, a list of routers' addresses that define their path to the destination (*Options* field of the IP datagram header)



4

Network architecture of a public operator



Core (nucleus) is focused
on high bit rate transfer and fast recovery, plus simple management

5

Leased lines are used to connect IP routers.
Network that offers leased line service is called "transport network" by telecom industry

Leased line

IP

Circuit technologies: time/frequency division multiplexing

PDH: T1, T3 switch: Digital Cross Connect (DCS)	SONET/SDH: OC3-OC768 Switch: SONET/SDH crossconnects	DWDM: OTU1-OTU3 Switch: optical WDM crossconnects

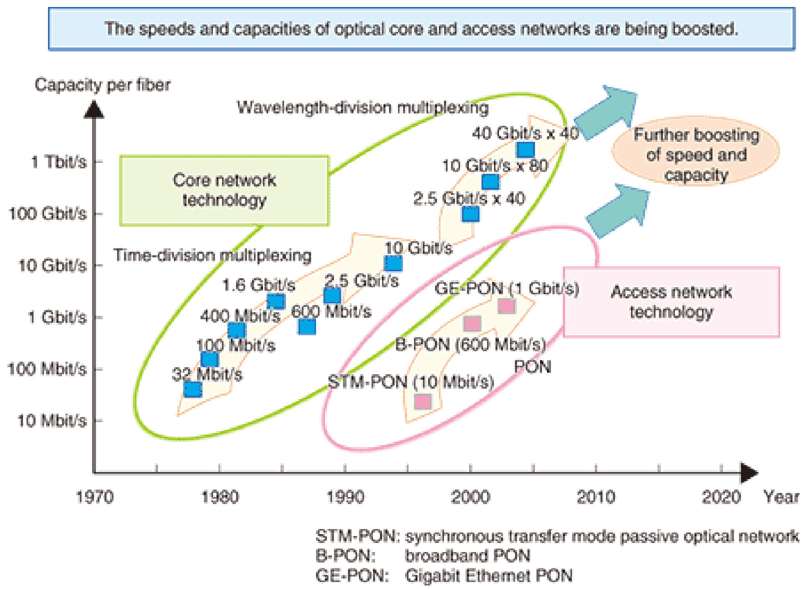
Packet technologies: virtual circuit switches

ATM	MPLS	Carrier-grade Ethernet

All the above technologies refer to data-plane technologies (control plane may be different)

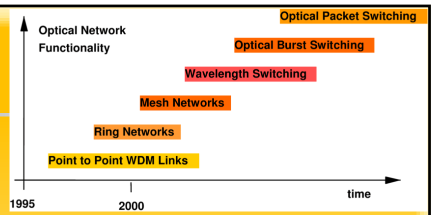
6

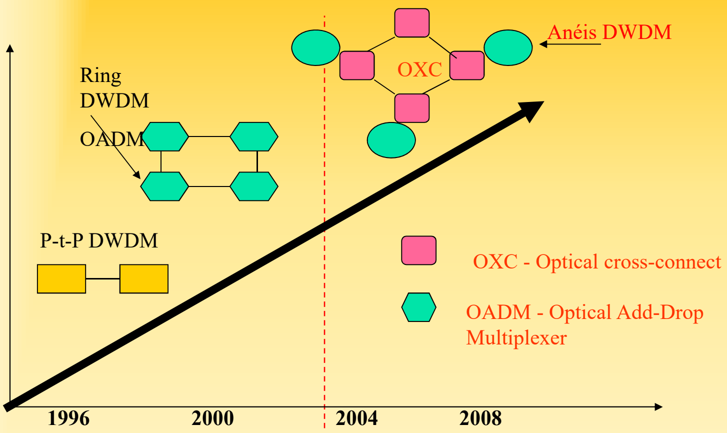
Optical Systems



7

Optical System Evolution



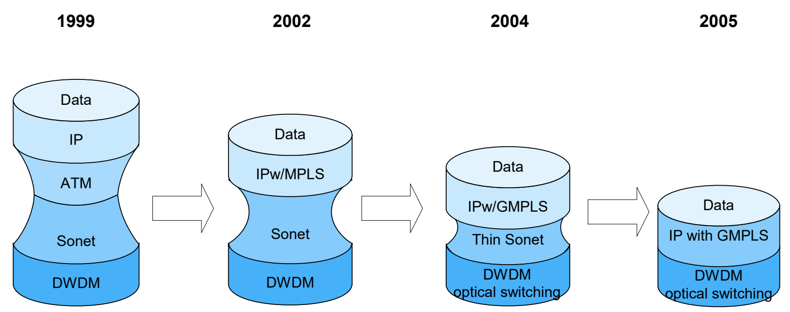


Improved optical infrastructure, increasingly optical switching.

8

Core network evolution

Increased simplification of the optical infrastructure

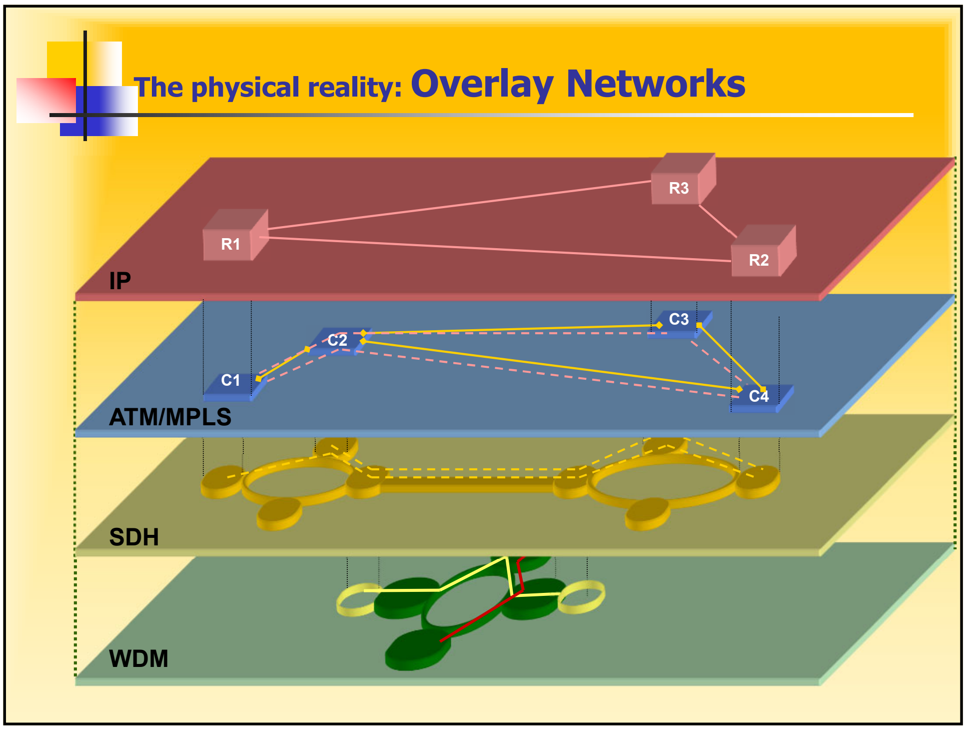


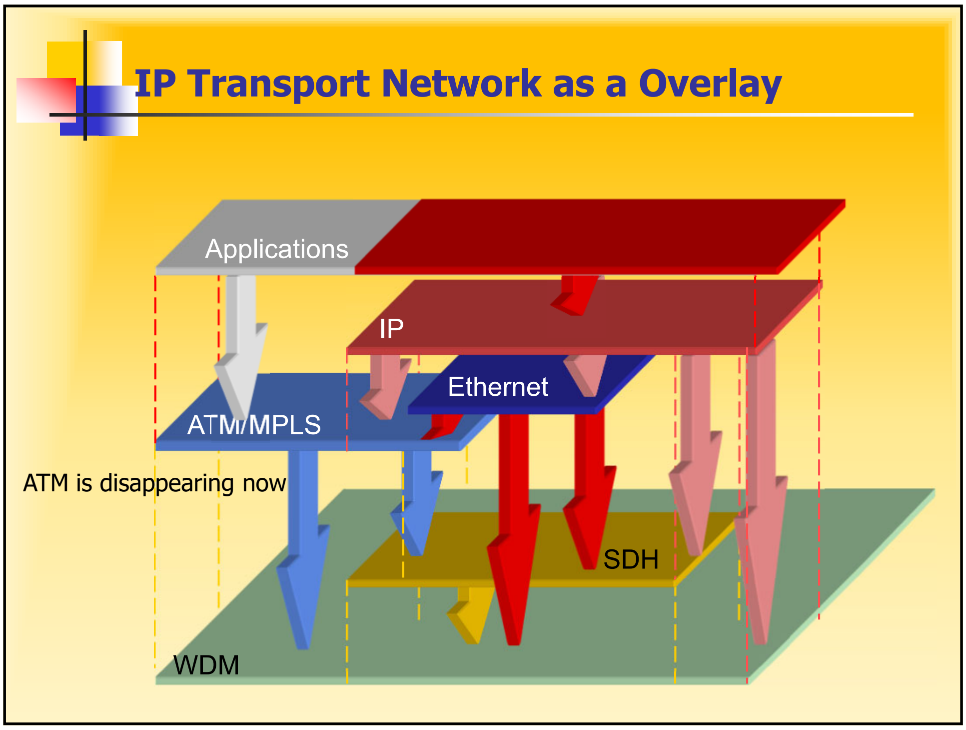
"GMPLS: An Overview of Routing and Management Enhancements, Ayan Banerjee et al"

9

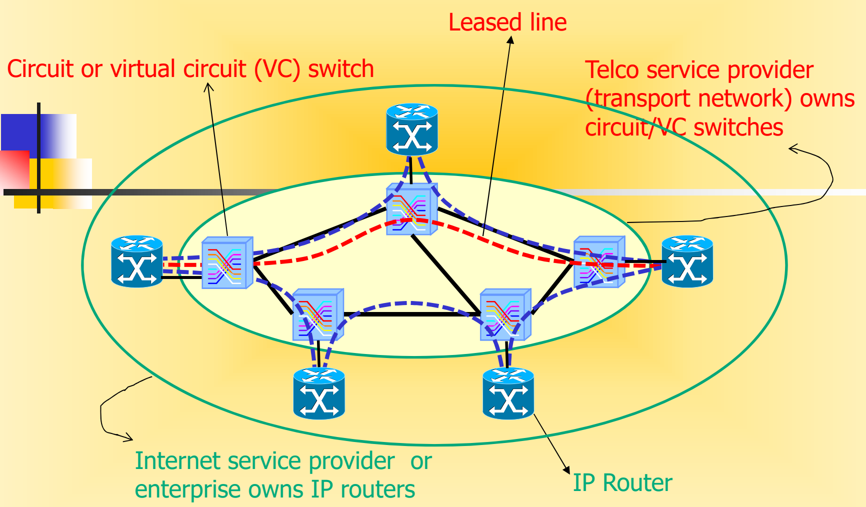
Why Migrating Core networks towards packet-based networks

- **Substitution of SDH/SONET circuit-switching transport network:**
  - Removing:
    - Inflexible bandwidth allocation of circuit switching
  - Keeping:
    - Connection Oriented
    - Deterministic performance and QoS
    - High resiliency and availability based on:
      - In-band OA&M
      - Protection with 50 ms recovery
    - Security
    - Simple, centralized operational model

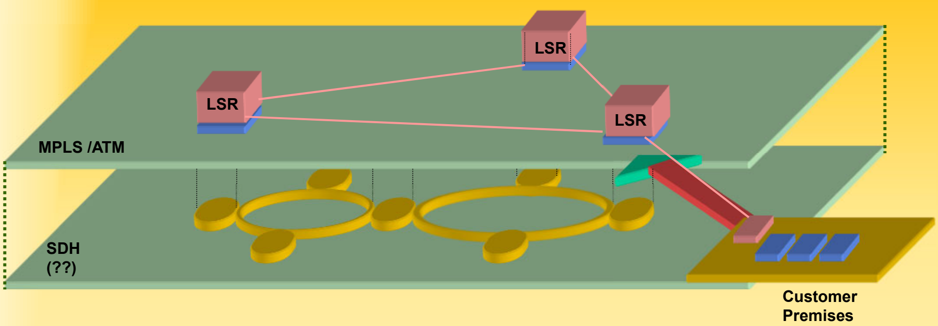

The physical reality: Overlay Networks
R3
R1
R2
IP
C3
C2
C1
C4
ATM/MPLS
SDH
WDM


IP Transport Network as a Overlay
Applications
IP
Ethernet
ATM/MPLS
ATM is disappearing now
SDH
WDM

IP and leased line service deployment

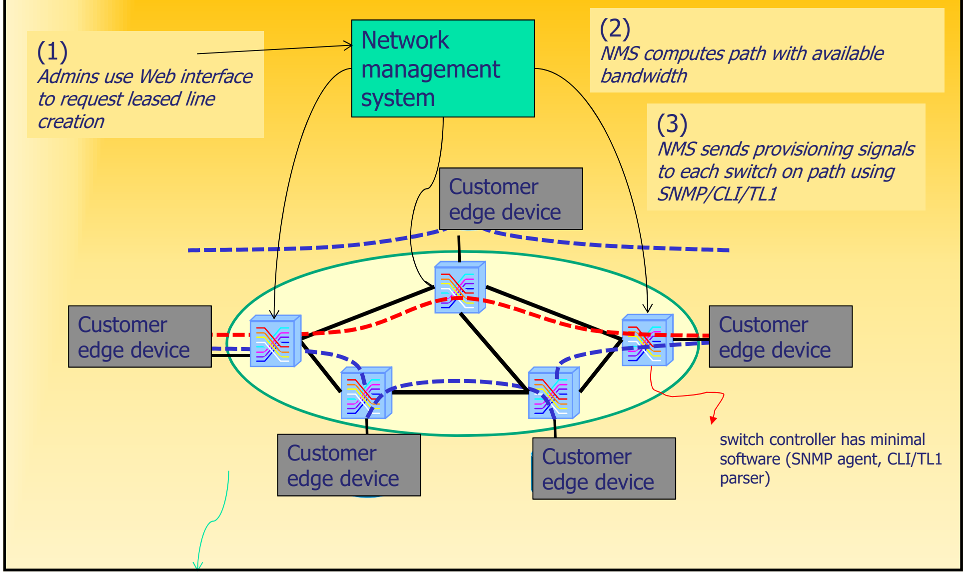


Switching Capacity not an issue any more, but some technology (ATM,MPLS) still needed



- Quality of Service
- Evolved VPN
- Traffic Engineering, Protection
- Multicast

Management plane (in transport network)

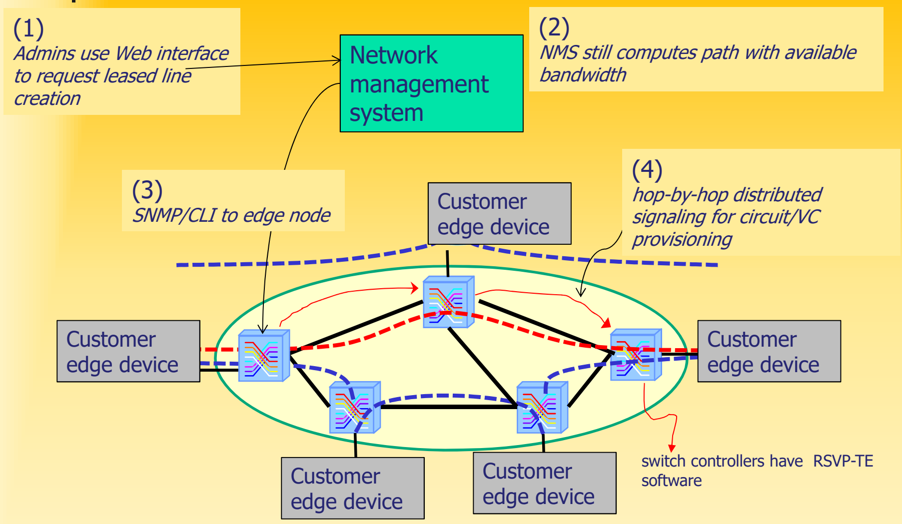


Migration towards packet-based networks

- **IP over MPLS/ATM**

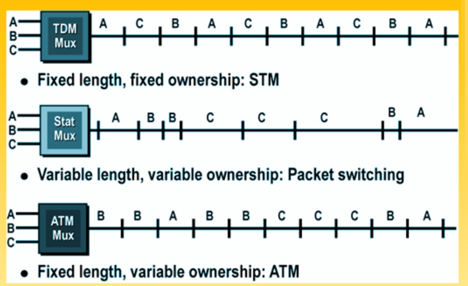
  improves/guarantees packet-switching performance and survivability:
  - Removing:
    - IP Connectionless;
    - IP best-effort delivery with limited Quality of Service (QoS);
  - Keeping:
    - IP Statistical Multiplexing;
    - IP Dynamic Control Plane;
  - Offering:
    - Connection Oriented, Client agnostic;
    - Efficient bandwidth management with QoS;
    - Provisioning using centralized management;
    - Resiliency based on control-plane OA&M mechanisms and MPLS fast reroute;
    - High Scalability;

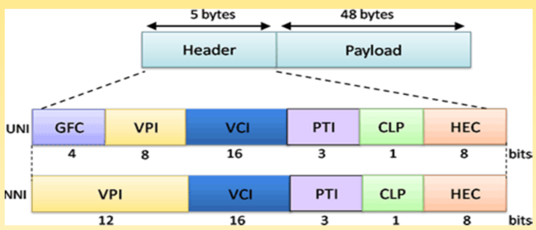
Management plane + control plane



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.
  - 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
  - Developed to overcome switch eletronics limitations at the time





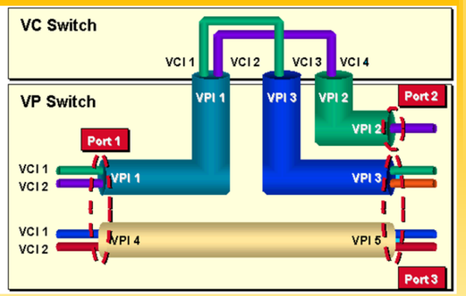
UNI (User-Network Interface).
NNI (Network-Network Interface)

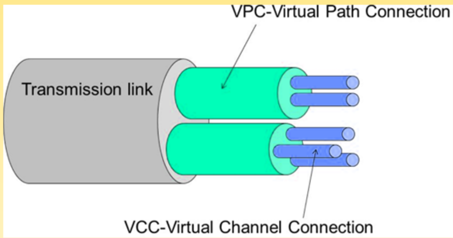
18

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 (very simple to do here) is based on VPI/VCI

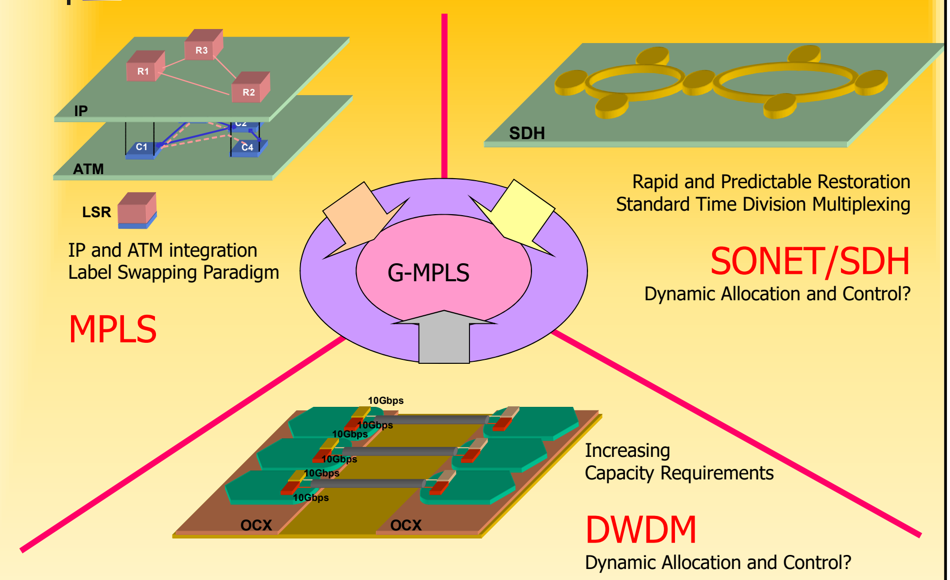
Port in	VPI/VCI	Port out	VPI/VCI
1	1/1	2	2/4
1	1/2	2	3/3
1	4/1	3	5/1
3	4/2	3	5/2





19

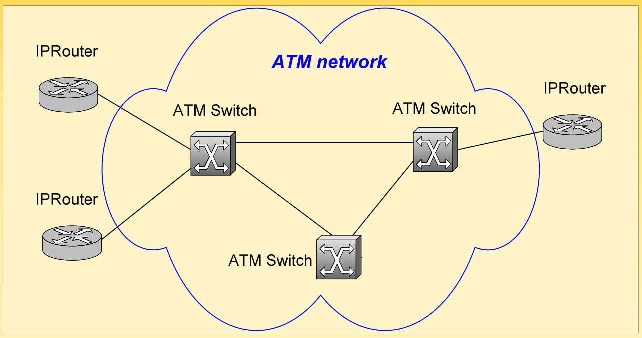
Main trends



MPLS

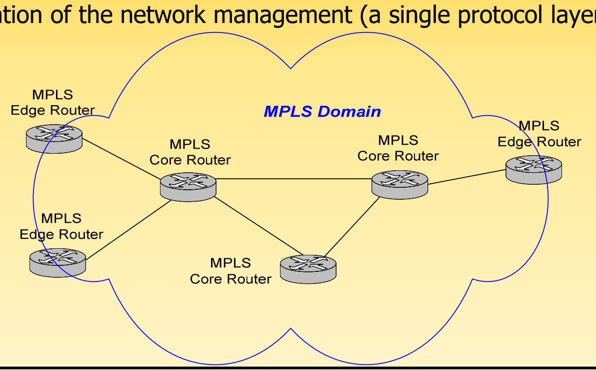
IP networks over ATM

- IP routers are interconnected by an ATM network
- Connections between IP routers are implemented through virtual circuits (VCCs) or virtual paths (VPCs) on the ATM network
- It is necessary to manage two protocol layers (ATM is not available anymore on new networks)



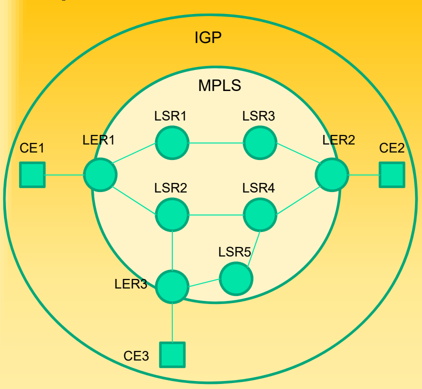
MPLS networks

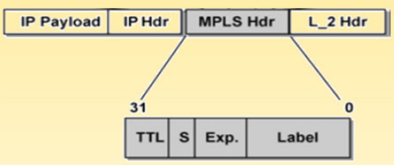
- Packets are labeled at the source with the label of the first hop
- Routers route packets based on their labels, just like ATM does with the VCI and VPI fields
- Advantages
  - Simplification of the packet routing process on routers
  - Traffic engineering capability equivalent to ATM
  - Simplification of the network management (a single protocol layer)



23

IP/MPLS



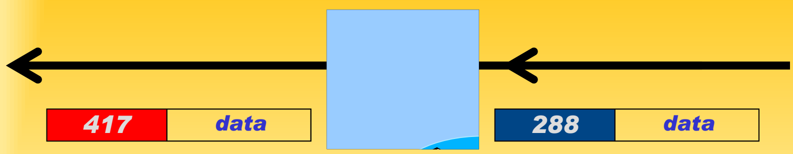


- **FEC – Forwarding Equivalence Class**
  - Identifies group of MPLS packets that are handled in the same way;
- **LABEL**
  - Local identifier based on the concept of FEC;
- **LSR – Label Switching Router**
  - LSRs in the same domain form an MPLS domain, using label forwarding;
- **LER – Label Edge Router**
  - LERs interacts with devices outside the domain, using IP technologies;

IP/MPLS

- **LSP – Label Switched Path**
  - Path along which FEC travels through an MPLS network;
- **LDP – Label Distribution Protocol**
  - The protocol used by MPLS for control that classifies FECs, distributes labels, and establish and maintains LSPs;
- **LFIB – Label Forwarding Information Base**
  - Forwarding table created by routing protocols and labels;
- **LSP tunneling**
  - Explicit route LSP between two LSRs that are not directly connected (indicated for TE);
- **Multi-level label stack**
  - Label stack carried by MPLS packets organized as LIFO (last-in first-out) stack;

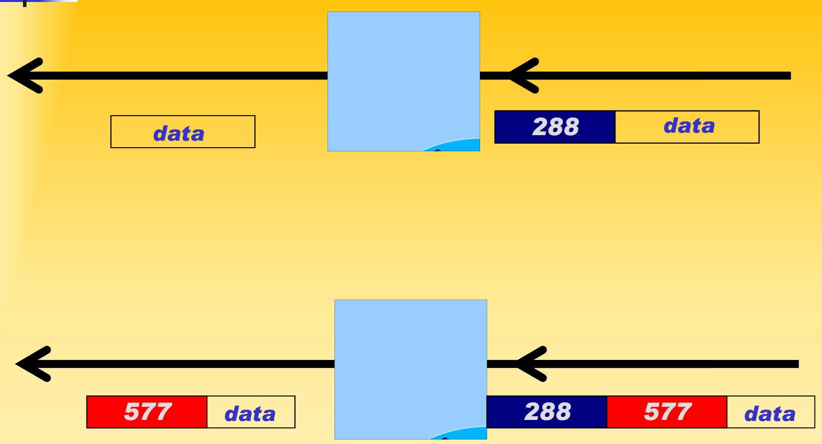
Forwarding via Label Swapping



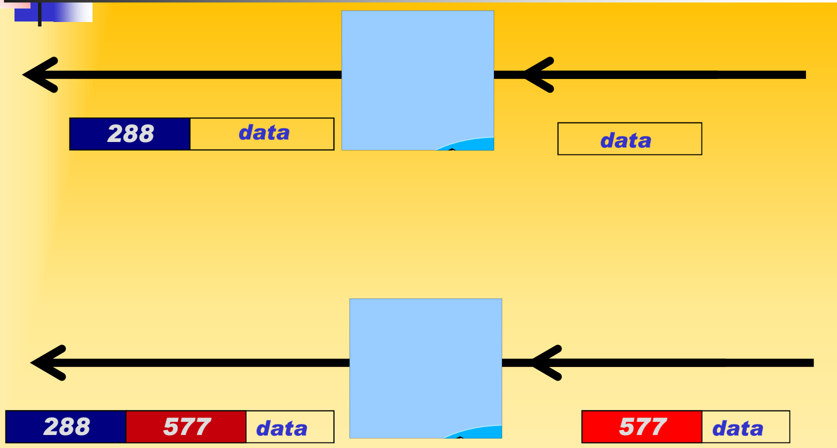
Labels are short, fixed-length values.

26

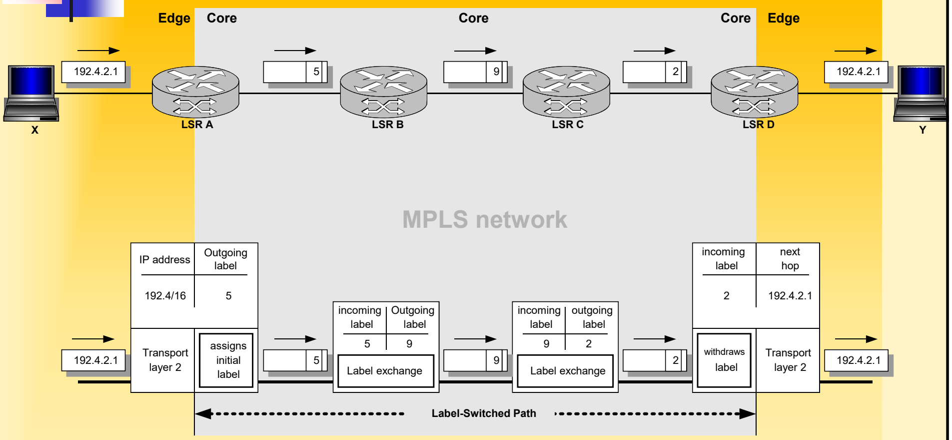
Popping Labels



Pushing Labels



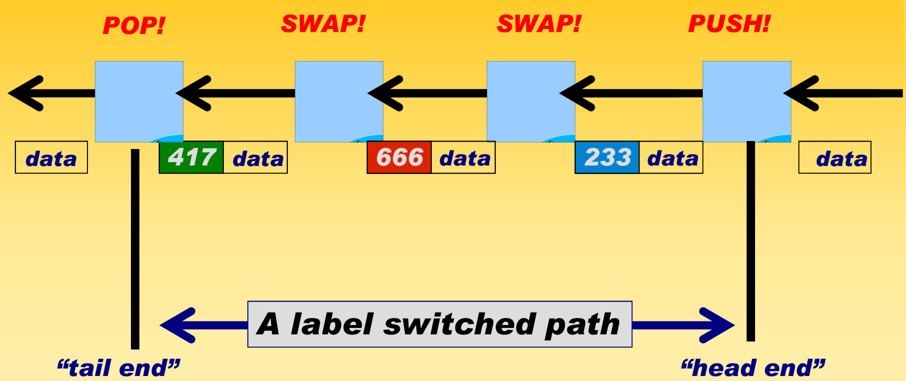
Label Switched Path: virtual circuit with label switching



- Networks are organized in domains
- Border routers insert/withdraw labels
- Labels have local meaning (can be reused on other links)
- Label distribution is made by an appropriate protocol

29

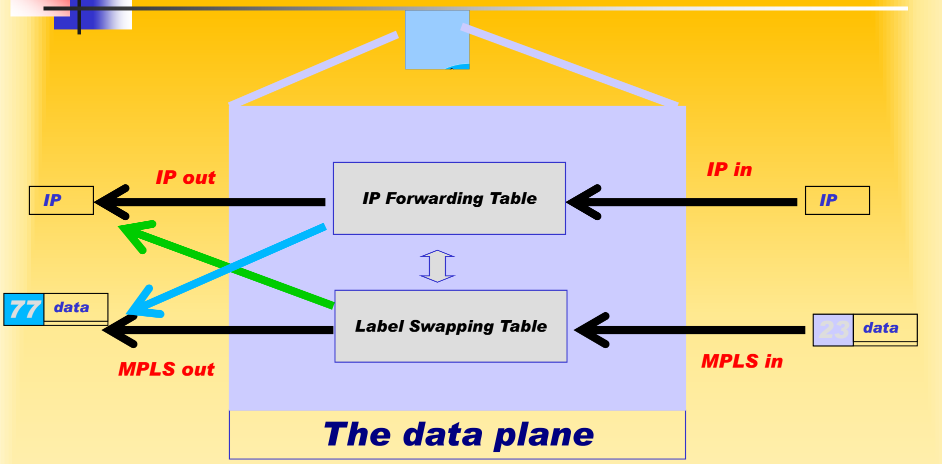
A Label Switched Path (LSP)



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

30

Label Switched Routers

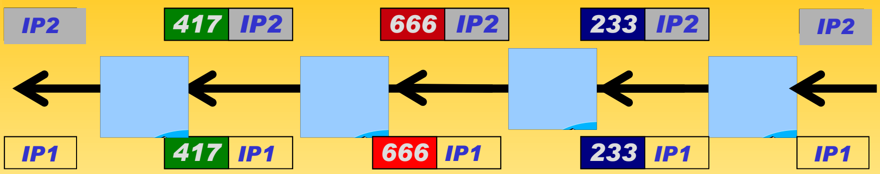


represents IP Lookup + label push

represents label pop + IP lookup

31

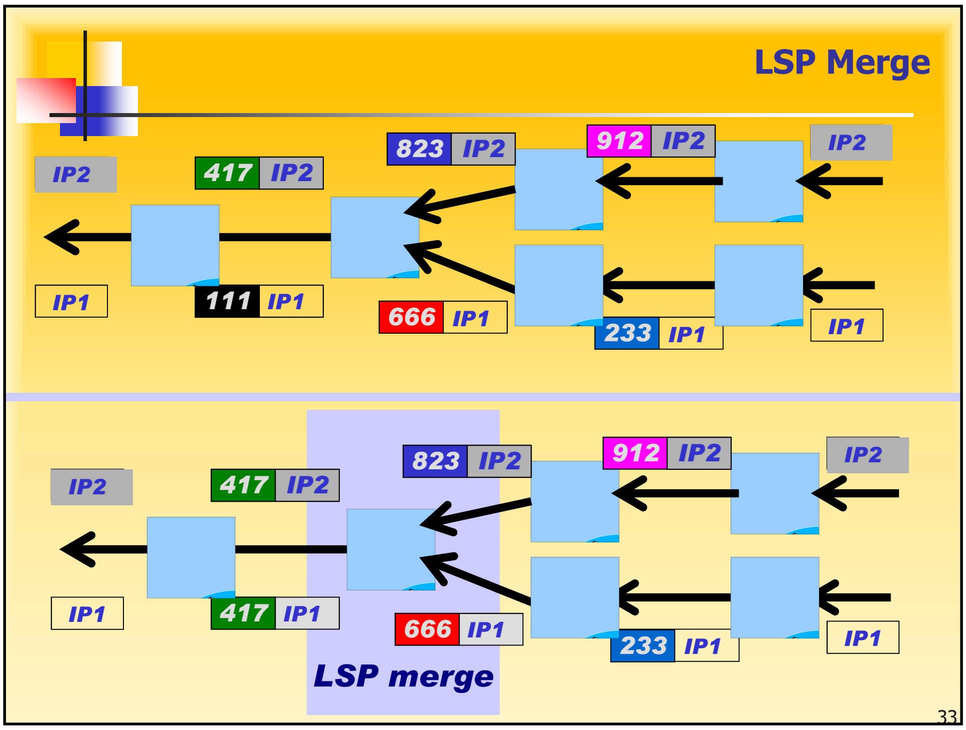
Forwarding Equivalence Class (FEC)

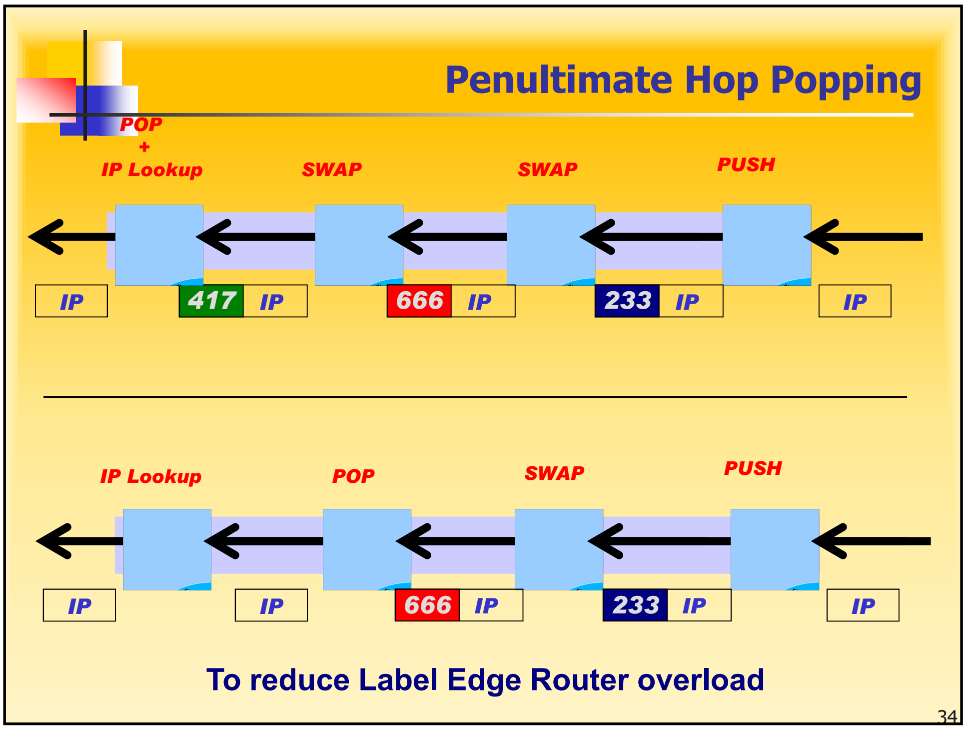


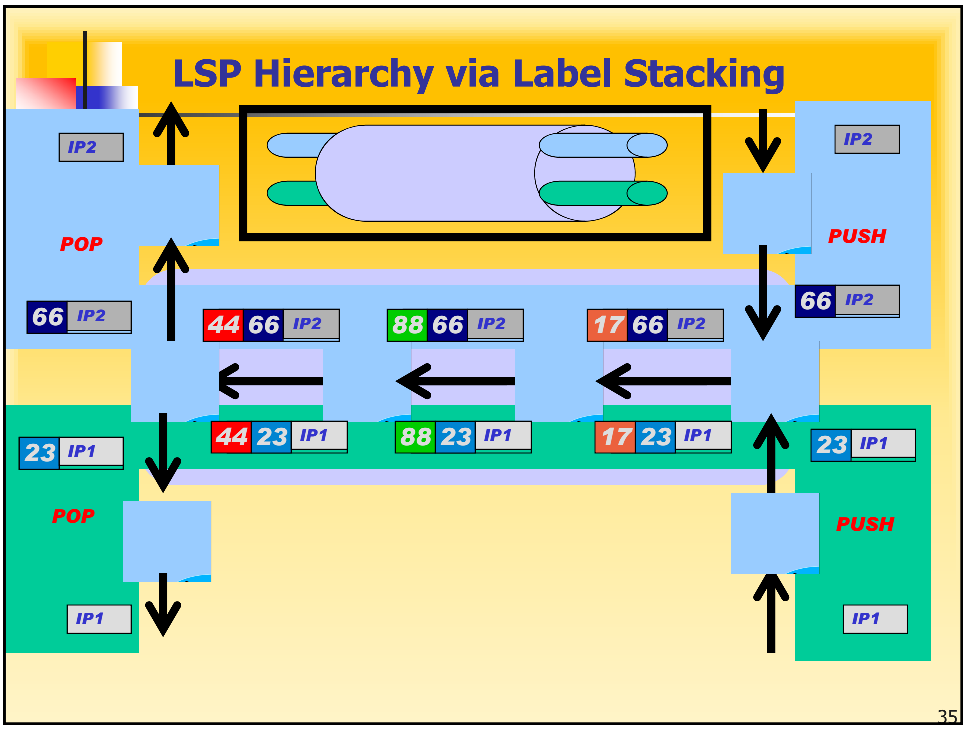
Packets IP1 and IP2 are forwarded in the same way --- they are in the same FEC.

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.

32


LSP Merge
IP2
417 IP2
823 IP2
912 IP2
IP2
IP1
111 IP1
666 IP1
233 IP1
IP1
IP2
417 IP2
823 IP2
912 IP2
IP2
IP1
417 IP1
666 IP1
233 IP1
IP1
LSP merge
33


Penultimate Hop Popping
POP
+
IP Lookup
SWAP
SWAP
PUSH
IP
417 IP
666 IP
233 IP
IP
IP Lookup
POP
SWAP
PUSH
IP
IP
666 IP
233 IP
IP
To reduce Label Edge Router overload
34



IP/MPLS Network Establishment

- **Discovery link and topology**
  - IP routing table is built
    - LSRs and LERs use routing protocols to discover network topology eg. OSPF, ISIS, (BGP);
    - CEs advertise their addresses using routing protocols into MPLS cloud;
  - Forwarding Information Base (FIB) is built, initially without label information
- **Label Assignment**
  - FECs creation
    - LSRs classify with the same FEC all packets handled on the same way;
  - Allocate Labels
    - Every LSR allocates locally labels for every destination in the IP routing table (LIB and LFIB setup);

Label Distribution Protocols

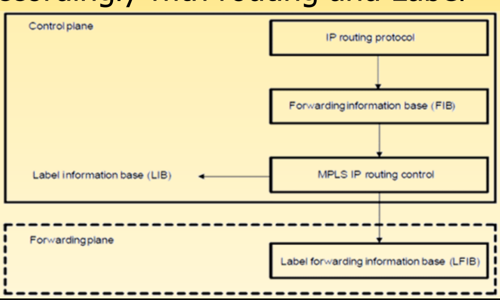
- Unconstrained routing
  - Label Distribution Protocol (LDP).
  - Path is chosen based on IGP shortest path.
- Constrained routing
  - Constrained by explicit path definition and/or performance requirements (e.g., available bandwidth).
  - Resource Reservation Protocol with Traffic Engineering (RSVP-TE).
    - 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.

37

IP/MPLS Network Establishment

Label distribution operation and LSP Establishment

- Discovery
  - Basic Discovery – LSRs send LDP link Hellos UDP (multicast) for directly connected peers.
  - Extended Discovery – LSRs send LDP targeted Hellos UDP for a specific (remote) IP peer.
- Session Establish and Maintenance
  - TCP session is established and it is maintained through periodically Keep-Alive messages
- LFIBs are established accordingly with routing and Label tables.



Label Distribution Protocol (LDP)

- 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).

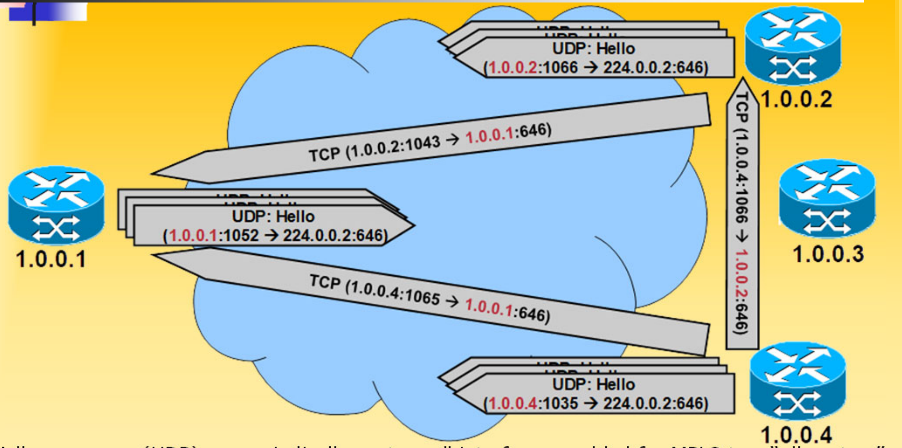
39

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.

40

LDP Neighbour Discovery

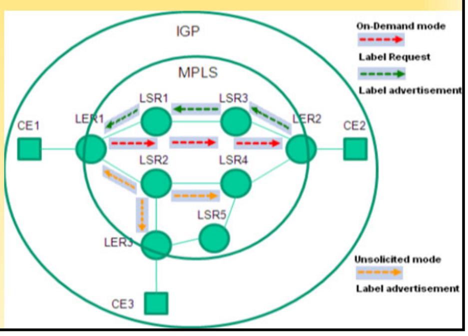


- Hello messages (UDP) are periodically sent on all interfaces enabled for MPLS to a "all-routers" multicast address (224.0.0.2).
- If there is another router on that interface it will respond by trying to establish a LDP/TCP session with the source of the hello messages.
- Both TCP and UDP messages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

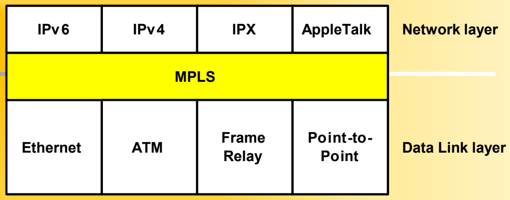
41

IP/MPLS Network Establishment

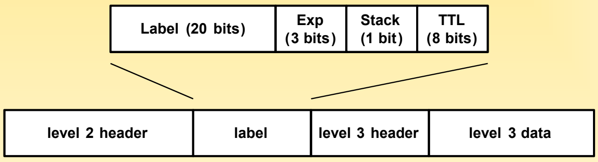
- LSP Establishment and Maintenance
  - Downstream On-demand mode– Upstream LSR sends a label request message (with FEC description) to its Downstream:
    - Ordered mode – a LSR only sends label (response) to its Upstream when it receives the label from its Downstream;
    - Independent mode – a LSR sends the label (response) when receives any request label;
  - Downstream Unsolicited mode – a Downstream LSR advertises label binding information to its Upstream LSR unsolicited after session to be established, without request ;
  - Label retention mode
    - Conservative mode
      - LSRs keeps only the labels from next hops
      - Indicated for limited label space
    - Liberal mode
      - LSRs keeps any labels, even if those are not from next hops
      - Indicated for quick adaptation of route changes



Labels

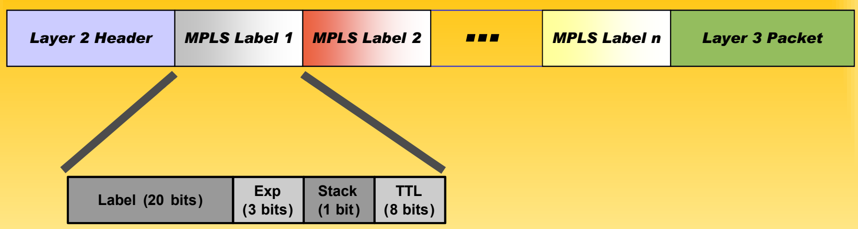


- On some Data Link (level 2) technologies, label is given by the appropriate fields of their header
  - ATM technology : VPI (Virtual Path ID) and VCI (Virtual Channel ID) fields
  - Frame Relay technology: DLCI (Data Link Connection Identifier) field
- On other Data Link technologies (Point-to-Point, Ethernet), the label is inserted between layer 2 and layer 3 headers



43

Generic MPLS Encapsulation



RFC 3032. MPLS Label Stack Encoding

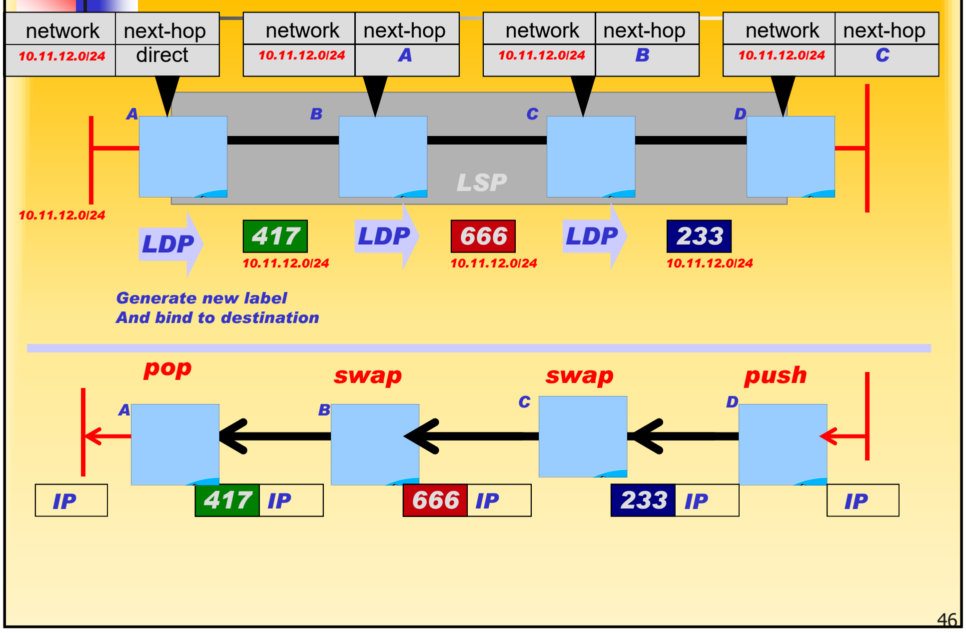
- ***Label:*** ***Label Value, 20 bits***
- ***Exp:*** ***Experimental, 3 bits***
- ***S:*** ***Bottom of Stack, 1 bit***
- ***TTL:*** ***Time to Live, 8 bits***

44

Multiprotocol Label Switching (MPLS) – TE usage

Engenharia de Computadores e Telemática
4º ano, 1º semestre, 2020/2021

LDP and Hop-by-Hop routing



46

Constrained based Routing

- A cost is associated to each link
- Each link has a further set of attributes that represent performance metrics

The routing objective is to determine the lowest cost path that does not violate the restrictions that were assigned

- Restrictions can be associated to a set of performance characteristics, like for example, bandwidth, delay, priority, etc.
  - For the bandwidth case, the restriction that is imposed to the routing algorithm is that the path must have, on each connection it traverses, a bandwidth higher than a certain threshold.
  - In this case, the connection attribute used is the available bandwidth.

47

Constraint Based Routing?

Basic components

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

1.Specify path constraints
2.Extend topology database to include resource and constraint information
Extend Link State Protocols (IS-IS, OSPF)
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)
Extend RSVP or LDP or both!
6.Map ingress traffic to the appropriate LSPs

Note: (3) could be offline (e.g. network operational center – NOC - systems) or online (perhaps an extension to OSPF)

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

48

Resource Reservation + Label Distribution

Two emerging/competing/dueling approaches:

Add label distribution and explicit routes to a resource reservation protocol

RSVP-TE

RSVP

CR-LDP

LDP

Add explicit routes and resource reservation to a label distribution protocol

Constraint-Based LSP Setup using LDP
draft-ietf-mpls-cr-lpd-05.txt

RSVP-TE: Extensions to RSVP for LSP Tunnels
draft-ietf-mpls-rsvp-lsp-tunnel-08.txt

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 signalling protocols

49

LSPs establishing protocols

- RSVP-TE (*Resource Reservation Protocol – Traffic Engineering*)
  - Extension of the RSVP protocol
- CR-LDP (*Constrained based Routing – Label Distribution Protocol*)
  - Extension of the LDP protocol, deprecated
- Both protocols enable:
  - The specification of a route to a LSP
  - To chose the labels on each link of the route
  - To make resources reservation for the LSP
- Routes are previously determined:
  - By management (Traffic engineering), in a NOC
  - By a *Constrained based Routing* type protocol

50

RSVP-TE vs. CR-LPD

RSVP-TE

- ***Soft state periodically refreshed***
- ***IntServ QoS model***

CR-LDP

- ***State maintained incrementally***
- ***New QoS model derived from ATM models***

And the QoS model determines the additional information attached to links and nodes and distributed with extended link state protocols...

And what about that other Internet QoS model, diffserv?

51

MORE ON LATER:
ReSerVation Protocol (RSVP)

- ReSerVation Protocol (RSVP) was developed to communicate resource needs between hosts and network devices
  - Associated to the Intserv QoS model
- RSVP allows:
  - The source to 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 fulfil the requested reservation.
- Encapsulated on IP (protocol type = 46 (0x2E))
- Signalling is based on 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).

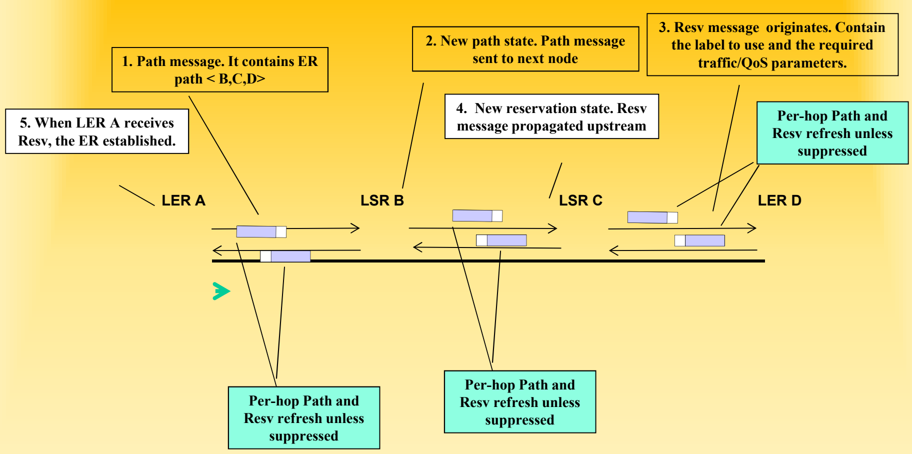
54

Resource Reservation Protocol with Traffic Engineering (RSVP-TE)

- Evolution of RSVP
  - RFC 3209: RSVP-TE: Extensions to RSVP for LSP Tunnels. (12/2001)
  - RFC 5151: Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions. (2/2008)
- To map traffic flows onto the physical network topology through label switched paths, resource and constraint network information are required
  - 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)

55

Establishment of a LSP with the RSVP-TE protocol



56

REVIEW: MPLS - Major Drivers

- Provide IP VPN Services *"the leased line"*
  - Scalable IP VPN service – Build once and sell many
  - Managed Central Services – Building value added services and offering them across VPNs
- Managing traffic on the network using MPLS Traffic Engineering
  - Providing tighter SLA/QoS (Guaranteed BW Services)
  - Protecting bandwidth - Bandwidth Protection Services
- Integrating Layer 2 & Layer 3 Infrastructure
  - Layer 2 services such as ATM (or Frame Relay) over MPLS
  - Mimic layer 2 services over a highly scalable layer 3 infrastructure

57

Terminology, 1/2

- RR—Route Reflector
  - A router (usually not involved in packet forwarding) that distributes BGP routes within a provider's network
- PE—Provider Edge router
  - The interface between the customer and the MPLS-VPN network; only PEs (and maybe RRs) know anything about MPLS-VPN routes
- P—Provider router
  - A router in the core of the MPLS-VPN network, speaks LDP/RSVP but not necessarily VPNv4
- CE—Customer Edge router
  - The customer router which connects to the PE; does not know anything about labels, only IP (most of the time)
- LDP—Label Distribution Protocol
  - Distributes labels with a provider's network that mirror the IGP, one way to get from one PE to another
- LSP—Label Switched Path
  - The chain of labels that are swapped at each hop to get from one PE to another

77

Terminology, 2/2

- VPN—Virtual Private Network
  - A network deployed on top of another network, where the two networks are separate and never communicate
- VRF—Virtual Routing and Forwarding instance
  - Mechanism in IOS used to build per-interface RIB and FIB
- VPNv4
  - Address family used in BGP to carry MPLS-VPN routes
- RD
  - Route Distinguisher, used to uniquely identify the same network/mask from different VRFs (i.e., 10.0.0.0/8 from VPN A and 10.0.0.0/8 from VPN B)
- RT
  - Route Target, used to control import and export policies, to build arbitrary VPN topologies for customers

78

Interconnecting networks

External Routing

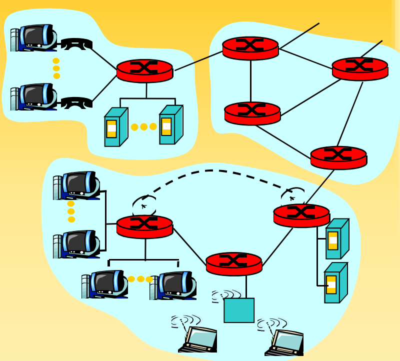
What you should learn…

- Notion of Autonomous System
- The role and diferences of External Routing protocols
- Features of BGP
- Attributes and advanced usage of BGP

80

Networks: service vision

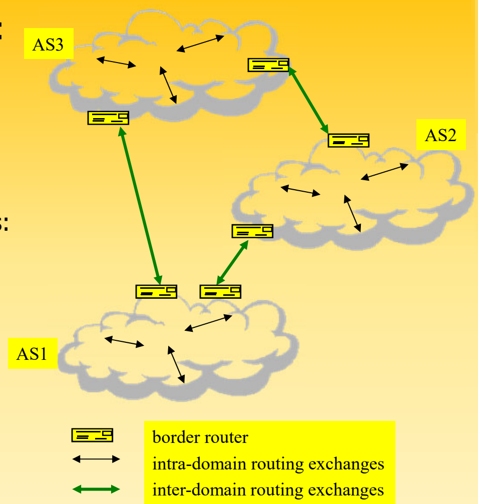
- Distributed communications infrastructure supporting applications, also potentially distributed
  - WWW, email, games, e-commerce, databases, voting
- Communications services supporting:
  - Connection-oriented
  - Connection-less
- Service platforms for millions of devices: *hosts, end-systems*
  - Pc's, workstations, servers
  - PDA's, phones, fridges…



81

Network structure

- Administrative borders define:
  - Autonomous systems(AS)
    - **Intradomain routing**
    - Internal policies
    - Different metrics can be used on different domains protocols: RIPv2, OSPFv2
  - Interconnection of ASs
    - **Interdomain routing**
    - Connectivity information protocols: BGP

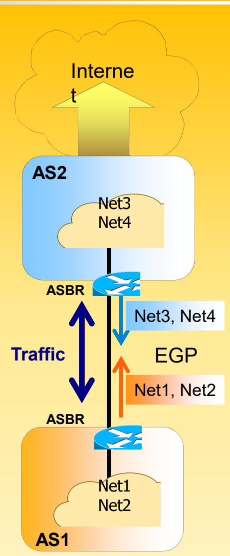


82

Single-homed (or Stub) AS

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

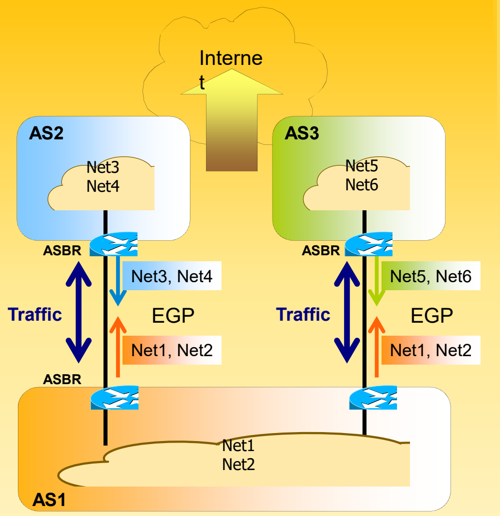
This also reflects the typical home network connection.



83

Multi-homed Non-transit AS

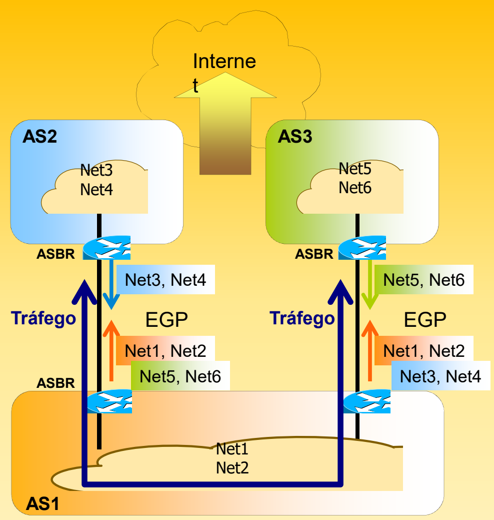
- AS has more than one border router (ASBR)
  - Multiple Internet accesses
  - Multiple ISP
- Does not transport traffic from other AS



84

Multi-homed Transit AS

- AS has more than one border router (ASBR)
  - Multiple Internet accesses
  - Multiple ISP
- Transports traffic from other AS
- Announces networks cross-domains



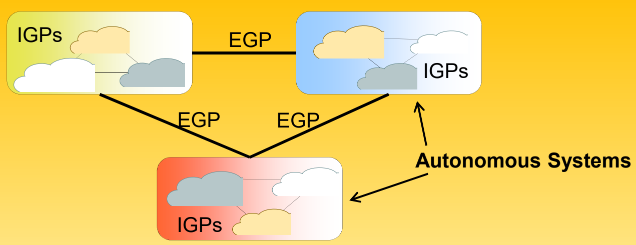
85

External Routing (BGP and MP-BGP)

Purpose:

If you are an operator (ISP) how do you control how your traffic flows inward (and outward) your domain?

Border Gateway Protocol (BGP)

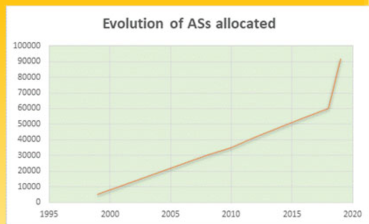


- Border Gateway Protocol Version 4 of the protocol (BGP4) was deployed in 1993 and currently is the protocol that assures Internet connectivity
- BGP is mainly used for routing between Autonomous Systems.
  - It has a very clear objective: provide a tool that ISPs may trust to build the Internet
  - Hint: traffic flow across links in AS correspond to different costs and profits, according with finantial agreements.
- Autonomous System (AS) is a network under a single administration
  - One (rarely more than one) network operators with a common well defined global routing policy

87

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
    - 92000+ have already been allocated
    - We would have 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 potential AS numbers (more than 54K used in 2016)
  - As of January 1, 2009, all new Autonomous System numbers issued were 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

88

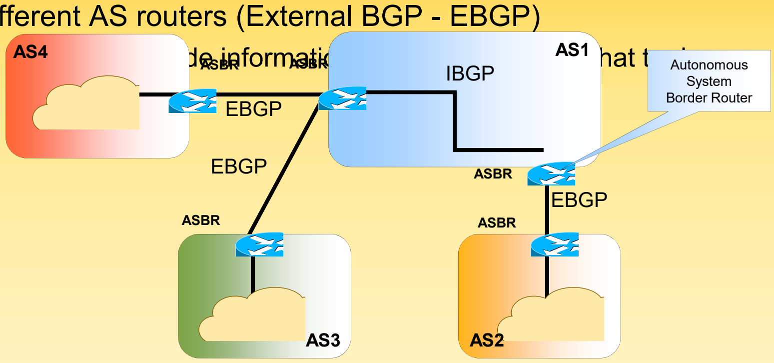
BGP Neighbor Relationships

- Often called **peering**
  - Usually manually configured into routers by the administrator
- Each neighbor session runs over TCP (port 179)
  - Ensures reliable data delivery
  - Reflect contracts!!!!
- 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

89

Internal BGP (IBGP) & External BGP (EBGP)

- Neighbor relations can be established between
  - Same AS routers (Internal BGP – IBGP)
    - Why? To provide information (inside the AS) about its multiple borders.
  - Different AS routers (External BGP - EBGP)



90

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 internally an IGP, such as OSPF
    - For all the routers in the AS, not only border routers.
  - Additional methods can be used to reduce IBGP Mesh complexity
    - Route reflectors, private AS, ...

91

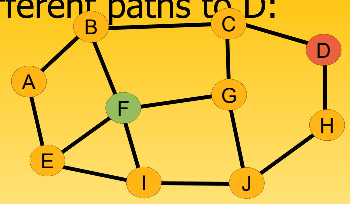
Path-vector

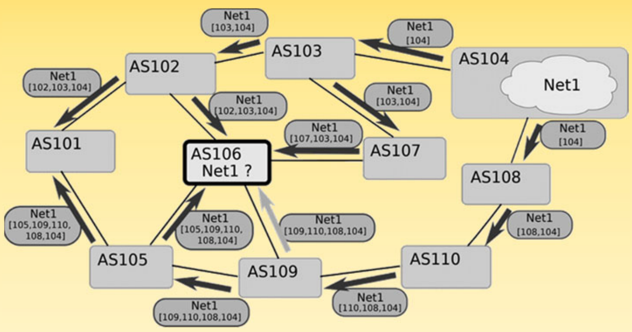
- BGP is a path-vector protocol
- Although it is essentially a distance-vector protocol that carries a list of the AS traversed by the route
  - Provides loop detection
- An EBGP speaker adds its own AS to this list before forwarding a route to another EBGP peer
- An IBGP speaker does not modify the 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

92

Path vectors (attribute AS_PATH)

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

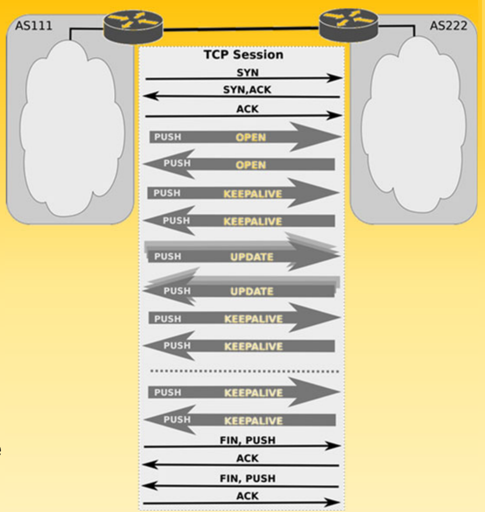




93

BGP Messages

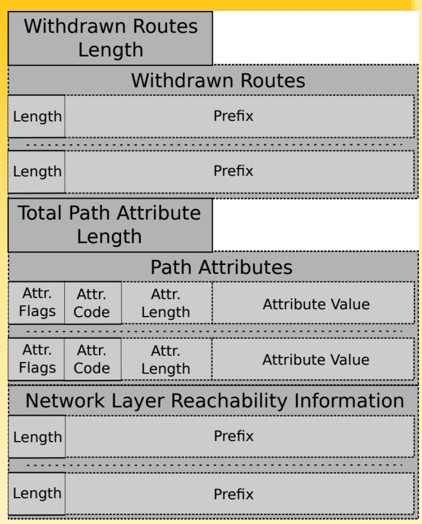
- OPEN messages are used to establish the BGP session
- UPDATE messages are used to send routing prefixes, along with their associated BGP attributes (such as the AS-PATH)
- NOTIFICATION messages are sent whenever a protocol error is detected, after which the BGP session is closed
- KEEPALIVE messages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



94

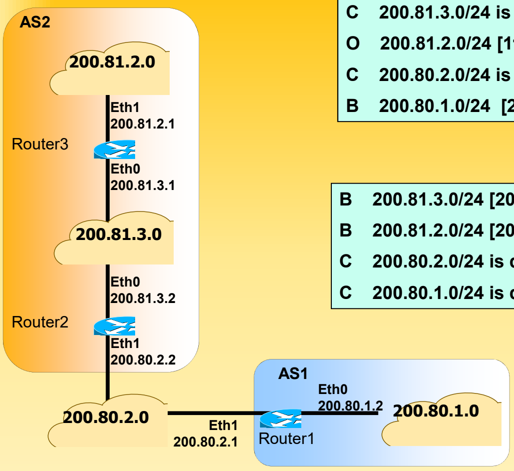
Update Message

- Widrawn 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



95

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

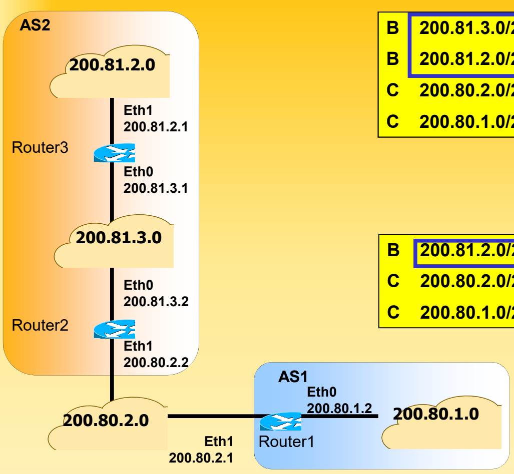
Router 2's routing table

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

96

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

Route aggregation is associated with the IP address allocation management globally in the Internet by ISOC.

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

98

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.
    - –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.

99

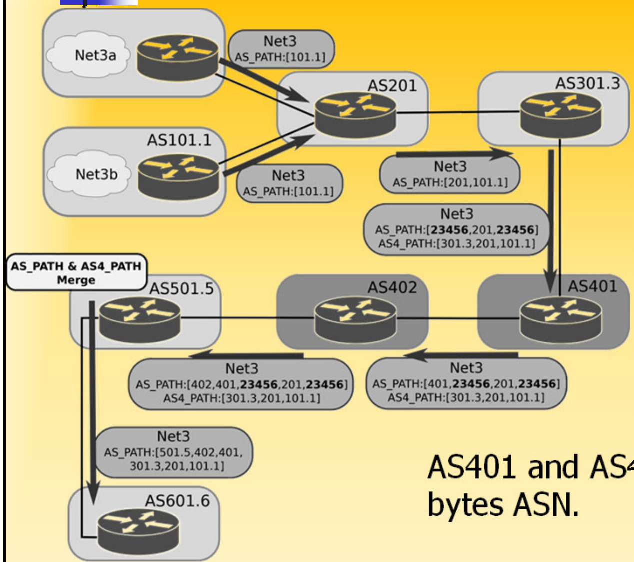
AS4_PATH & AS4_AGGREGATOR

- 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

100

4-bytes AS Operational Example

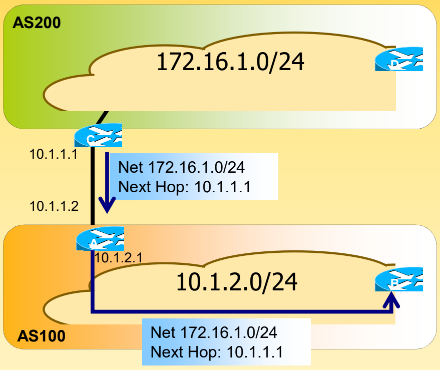


AS401 and AS402 do not support 4-bytes ASN.

101

Next-Hop Attribute

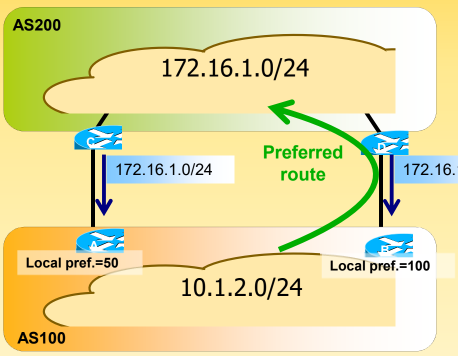
- The EBGP next-hop attribute is the IP address that is used to reach the advertising router
- For EBGP, the next-hop address is the IP address of the connection between the peers
- For IBGP, the EBGP next-hop address is carried into the local AS
  - By configuration the AS border router can be the next-hop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



103

Local Preference Attribute

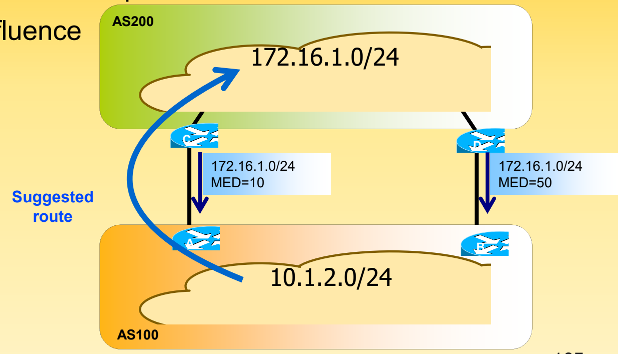
- The local preference attribute is used to choose an exit point from the local autonomous system (AS)
- The local preference attribute is propagated throughout the local AS
- If there are multiple exit points from the AS, the local preference attribute is used to select the exit point for a specific route



104

Multi-Exit Discriminator Attribute (MED)

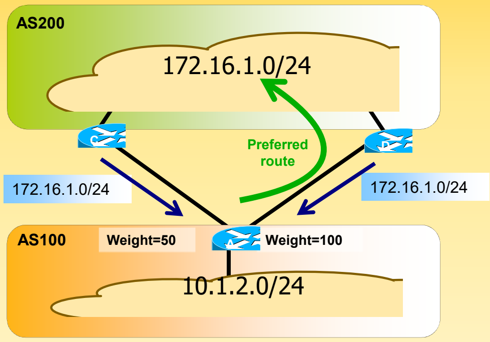
- The multi-exit discriminator (MED) or metric attribute is used as a suggestion to an external AS
- The external AS that is receiving the MEDs may be using other BGP attributes for route selection
- The **lower value** of the metric is preferred
- MED is designed to influence incoming traffic



105

Weight Attribute

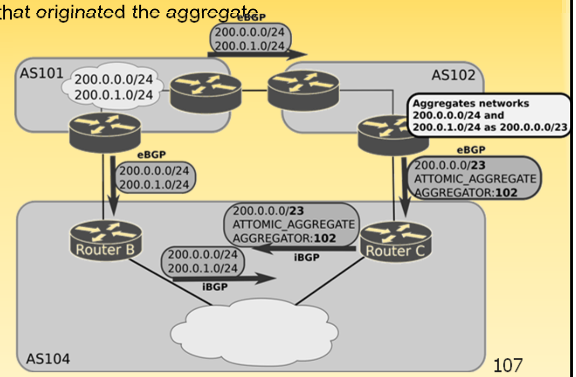
- Weight is a Cisco-defined attribute that is local to a router.
- The weight attribute is not advertised to neighboring routers.
- If the router learns about more than one route to the same destination, the route with the highest weight will be preferred.



106

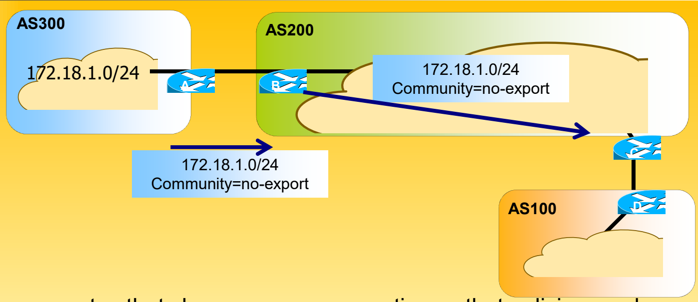
Atomic Aggregate and Aggregator Attributes

- 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.



107

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 *(transitive property)*
  - e.g. 300:1, 200:38, etc...

108

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

109

What you should learn...

- Notion of Autonomous System
- The role and diferences of External Routing protocols
- Features of BGP
- Attributes and advanced usage of BGP

110

BGP AS Routing Policies

```
aut-num:    AS15525
as-name:    PTPRIMENET
descr:      PT Prime Autonomous System
descr:      Corporate Data Communications Services
descr:      Portugal
import:     from AS1930 action pref=100;
        accept AS-RCCN # RCCN
import:     from AS3243 action pref=200;
        accept AS-TELEPAC # Telepac
import:     from AS5516 action pref=100;
        accept AS5516 # INESC
import:     from AS5533 action pref=100;
        accept AS-VIAPT # Via NetWorks Portugal
import:     from AS8657 action pref=300;
        accept ANY # CPRM
import:     from AS12305 action pref=100;
        accept AS12305 # Nortenet
import:     from AS1897 action pref=100;
        accept AS1897 AS9190 AS13134 AS15931 # KPN Qwest
import:     from AS13156 action pref=100;
        accept AS13156 # Cabovisao
import:     from AS8824 action pref=100;
        accept AS8824 AS15919 # Eastecnica
```

```
export:     to AS1897 announce RS-PTPRIME # KPNQwest
export:     to AS1930 announce RS-PTPRIME # RCCN
export:     to AS3243 announce RS-PTPRIME # Telepac
export:     to AS5516 announce {0.0.0.0/0} # INESC
export:     to AS5533 announce RS-PTPRIME # Via NetWorks Portugal
export:     to AS8657 announce RS-PTPRIME # CPRM
export:     to AS8824 announce RS-PTPRIME  # Eastecnica
export:     to AS8826 announce {0.0.0.0/0} # Siemens
export:     to AS9186 announce RS-PTPRIME # ONI
export:     to AS12305 announce RS-PTPRIME # Nortenet
export:     to AS12353 announce RS-PTPRIME # Vodafone Portugal
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

111

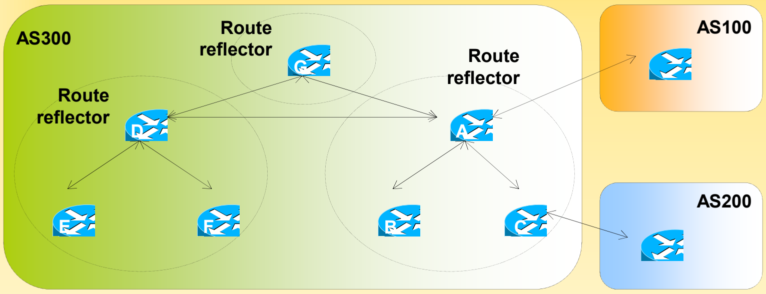
Problem: BGP Synchronization

- Synchronization states that, if your AS passes traffic from another AS to a third AS, BGP should not advertise a route before all the routers in your AS have learned about the route via IGP.
  - This is a problem for large ISPs, with many ASBG
- BGP waits until IGP has propagated the route within the AS.
  - Then, BGP advertises the route to external peers.

112

Solution: 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



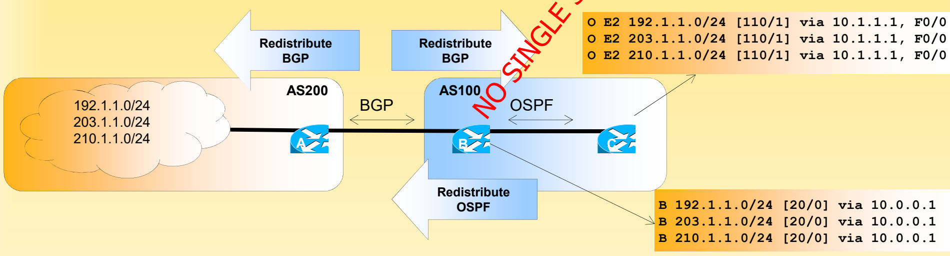
113

Potential Problem: Route Redistribution

NO SINGLE SOLUTION IS ADEQUATE EVERYWHERE

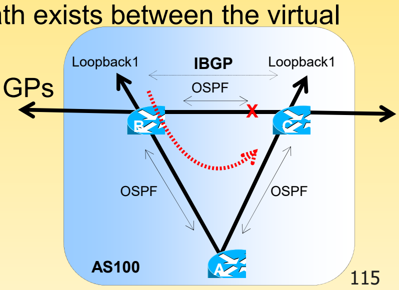
A. Redistributing IGP routes by BGP will: FROM AS → Intern
   - Simplify BGP configuration (advantage?)
   - And BGP will announce only internal networks with connectivity (advantage?)

B. Redistributing BGP routes by IGP protocols will: FROM Internet → A
   - Make internal routers 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?)



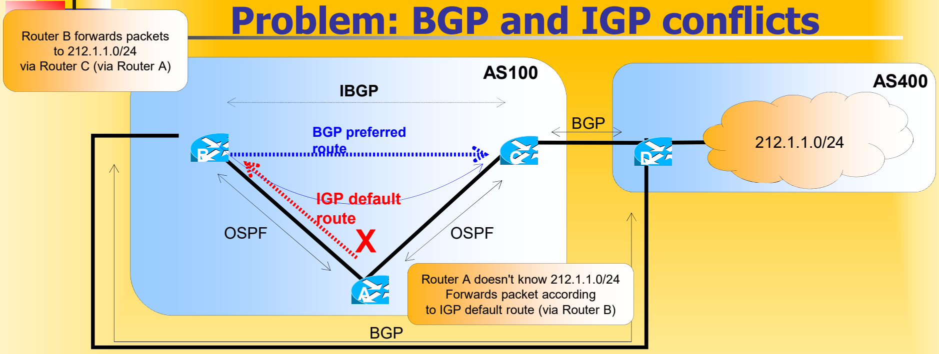
Problem: BGP Neighborhood Resilience

- BGP neighbor relations between physical interfaces are dependent on interface stability/status
- (Virtual) neighbor relations using Loopback interfaces/addresses
  - Loopback interfaces are virtual and software based
    - If the router is active Loopback interfaces are always active
  - Neighbor relation is active while a path exists between the virtual networks
    - (Alternative) Routing provided by IGPs



115

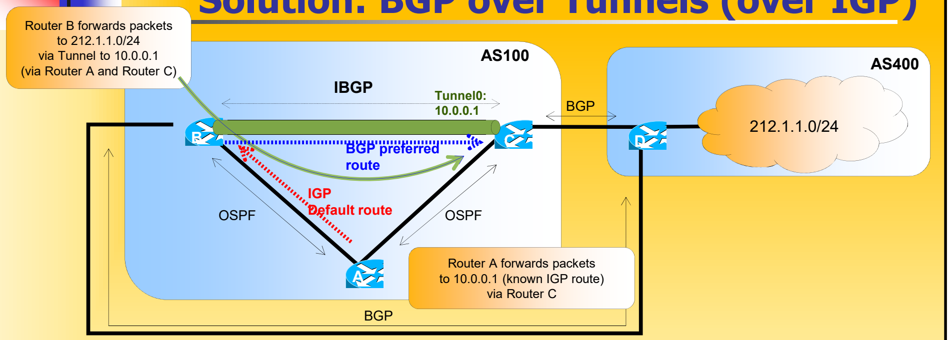
Problem: BGP and IGP conflicts



- Routing conflicts may arise with
  - Internal routers without BGP
  - No redistribution of BGP routes by IGP
  - IGP default routes
  - BGP preferred routes (with no agreement with IGP default routes)
- Solutions
  - Adjust IGP default routes
  - Adjust BGP preferred routes (e.g. with local preference)
  - BGP neighborhood and Internal routing via IP-IP tunnels

116

Solution: BGP over Tunnels (over IGP)



- IP-IP tunnels to solve BGP/IGP routing conflicts
  - Tunnels manually configured
    - Between physical or Loopback interfaces
  - BGP neighborhood via Tunnel
  - BGP routes learned via Tunnel (next hop is remote Tunnel end-point)
  - Tunnel "network" distributed internally via IGP
- In Router A, to any packet destined to an outside network it's forwarded via Tunnel
  - A new IP header is added, new IP destination address is the remote Tunnel end-point
  - Internally, packet is routed according to the new IP header (Tunnel end-points IP addresses)

117

BGP Filtering and Route Maps

- Sending and receiving BGP updates can be controlled by using a number of different filtering methods.
- BGP updates can be filtered based on
  - Route information
  - Path information
  - Communities
- Route maps are used with BGP to
  - Control and modify routing information
  - Define the conditions by which routes are redistributed between routing domains

118