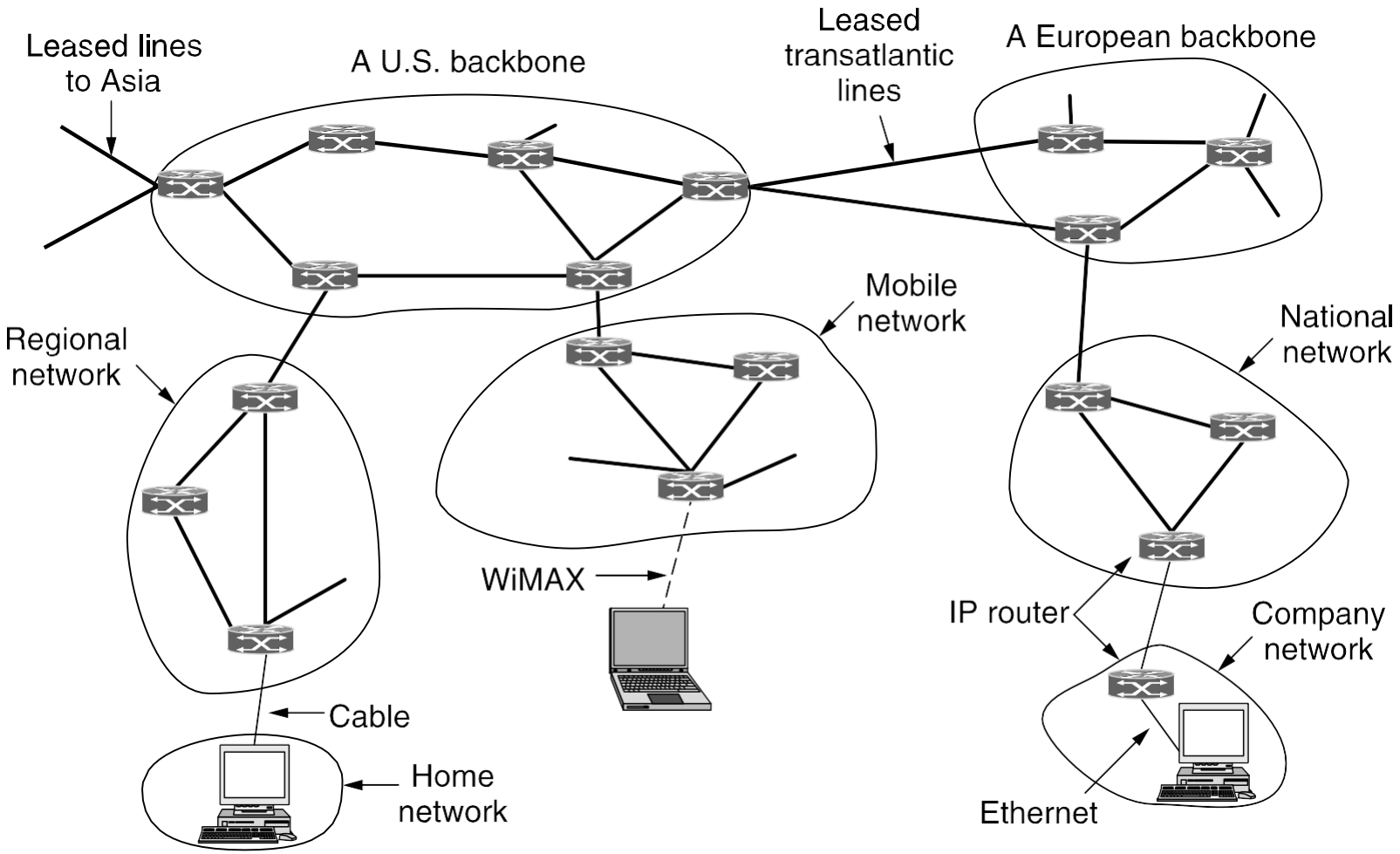
Internet Routing Protocols

The Internet Architecture

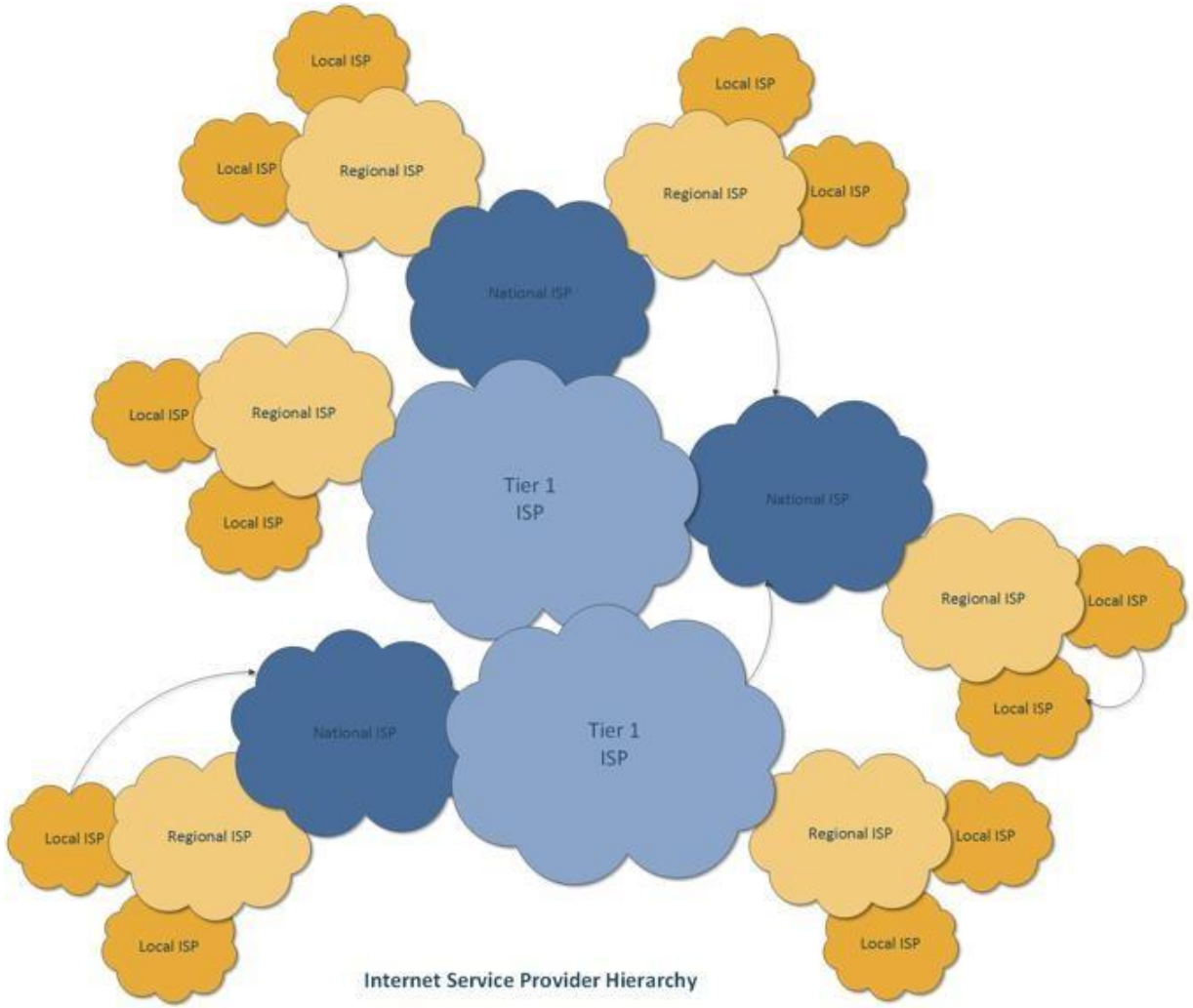


The glue that holds the whole Internet together is the network layer protocol, **IP** (**Internet Protocol**).

Internet Architecture

Autonomous Systems (AS) – A set of LANs for an administrative domain, identified by a unique AS number, and the routing policies are controlled by a single administrator.

Local Area Network (LAN) – A set of devices with a common layer 3 gateway



Routing Protocols So far

- **Link state routing:** *"tell about your neighbors to everyone".*
  - Each node collects information about all its neighbors and the link metrics.
  - This LSA (link state announcement) of every node is flooded through the entire network.
  - So, at the end of it, each node has complete view of the network graph.
  - Each node then independently runs Dijkstra's shortest path algorithm to get the shortest path to every destination, based on which it figures out the next hop for every destination.

- **Distance vector routing:** *"tell about everyone to your neighbors".*
  - Every node exchanges a distance vector containing its estimate of distance to each destination, with its neighbors.
  - Upon receiving a neighbor's distance vector, a node updates its distance vector by adding link cost to neighbor.
  - If a better path is found through the neighbor, it updates its best route.

Routing in the Internet

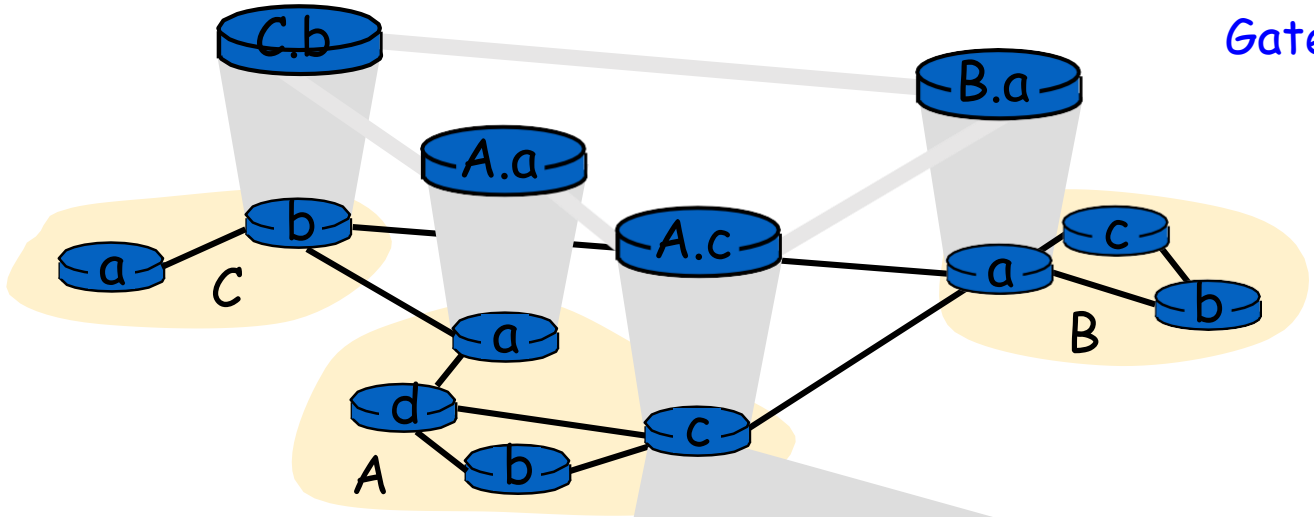
- Generic Routing Algorithms (Dijkstra / Bellman-Ford) – idealization
  - All routers are identical
  - Network is flat.
    - Not true in Practice

- Hierarchical routing
  - Internet = network of networks
  - Each network admin may want to control routing in its own routing network.
  - Hierarchical routing solves
    - Scale problems
    - Administrative autonomy.

Routing in theInternet

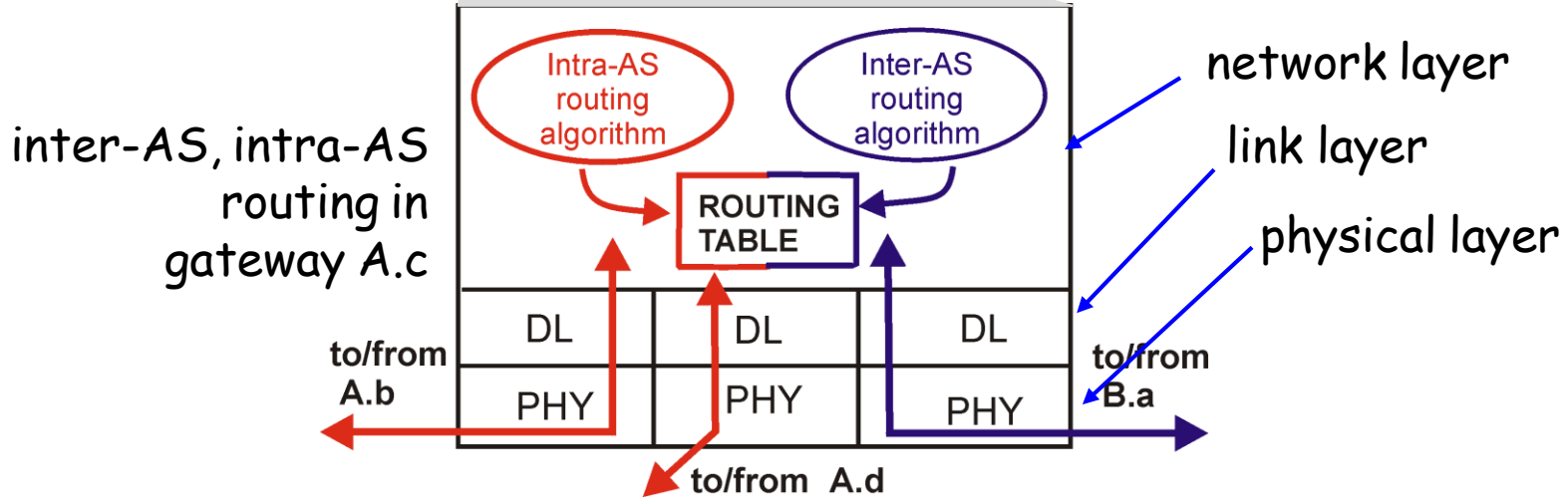
- **Intra-AS (Intra Domain) Routing Protocols:** Routing within an AS. Sometimes, they are called as **Interior Gateway Protocols (IGP)**
  - Routing Information Protocol (RIP)
  - Open Shortest Path First (OSPF)
  - IGRP: Interior Gateway Routing Protocol (Cisco propr.)
- **Inter-AS (Inter Domain) Routing Protocols:** Routing among the various ASes based on peering relationship. These are known as **Exterior Gateway Protocols (EGP)**
  - Border Gateway Protocol (BGP).

Intra-AS and Inter-ASRouting



Gateways:

- perform inter-AS routing amongst themselves
- perform intra-AS routers with other routers in their AS



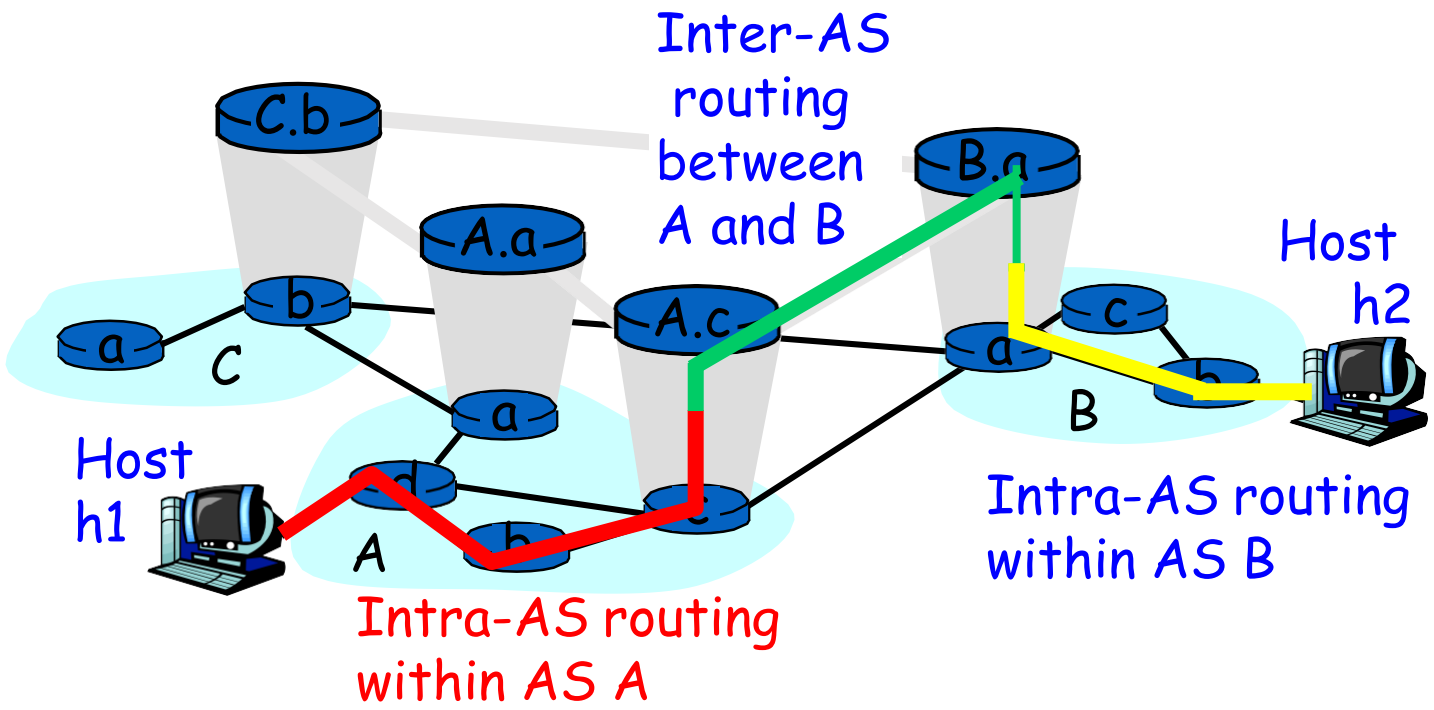
inter-AS, intra-AS routing in gateway A.c

network layer

link layer

physical layer

Intra-AS and Inter-AS Routing



Why different Intra- and Inter-AS routing?

Policy:

- Inter-AS: admin wants control over how its traffic routed, who routes through its net
- Intra-AS: single admin, so no policy decisions needed

Scale:

- hierarchical routing saves table size, reduced update traffic

Performance:

- Intra-AS: can focus on performance
- Inter-AS: policy may dominate over performance

Routing Information Protocol

- Distance vector algorithm
- Included in BSD-UNIX Distribution in 1982
- Distance metric: # of hops (max = 15 hops)
  - *Can you guess why?*
- Distance vectors: exchanged every 30 sec via Response Message (also called **advertisement**)
- Each advertisement: route to up to 25 destination nets

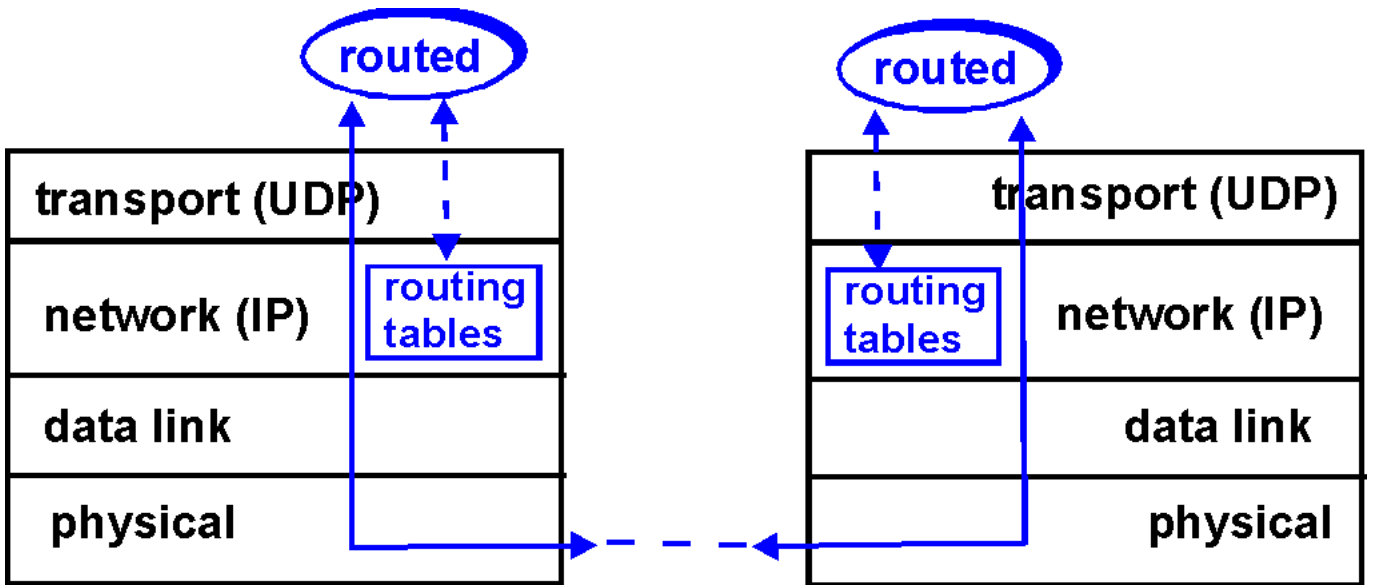
RIP: Link Failure and Recovery

If no advertisement heard after 180 sec --> neighbor/link declared dead

- routes via neighbor invalidated
- new advertisements sent to neighbors
- neighbors in turn send out new advertisements (if tables changed)
- link failure info quickly propagates to entire net
- Split horizon is used to prevent ping-pong loops (infinite distance = 16 hops)

RIP Table Processing

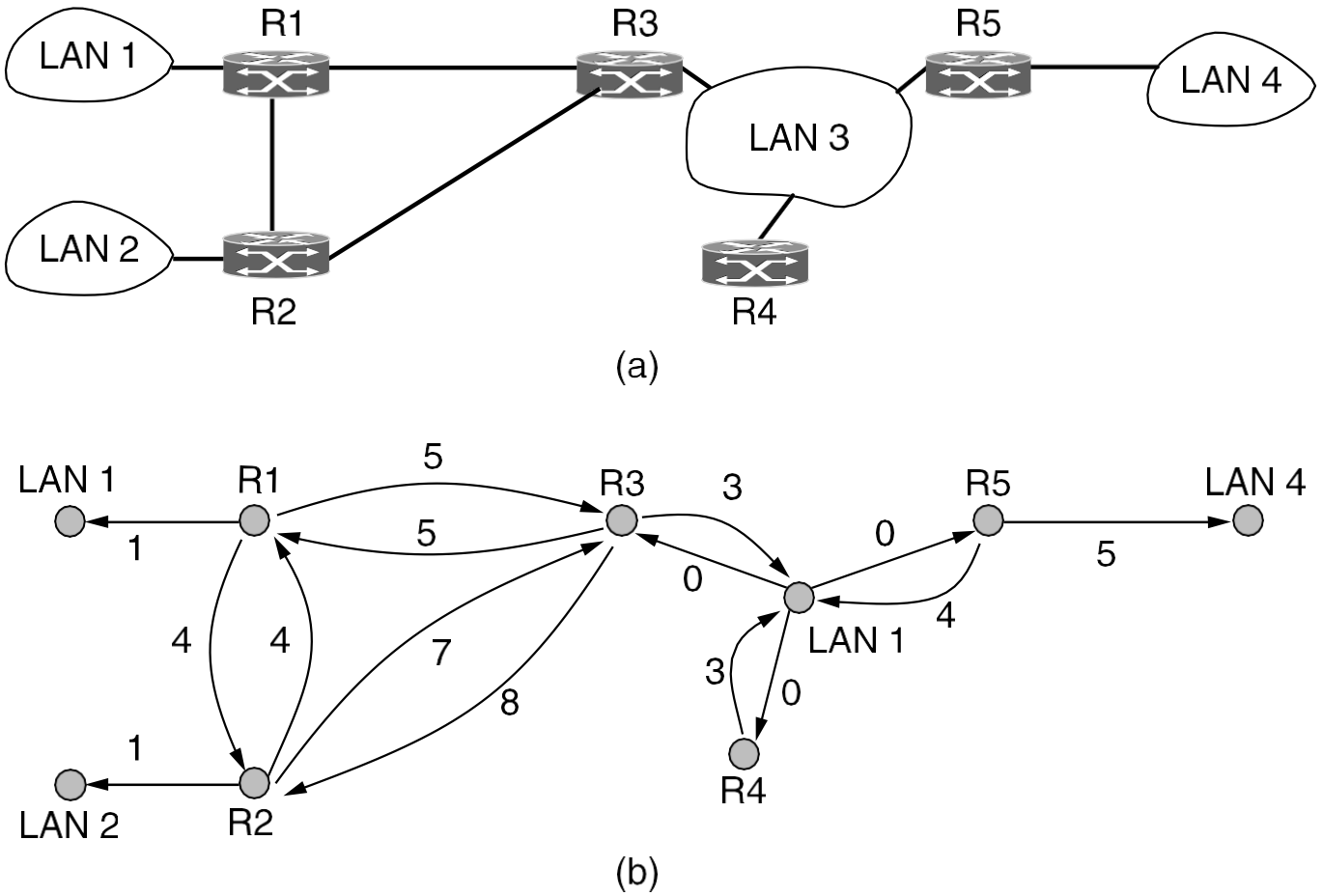
- RIP routing tables managed by **application-level** process called route-d (daemon) advertisements sent in UDP packets, periodically repeated
- RIP uses the UDP as its transport protocol, and is assigned the reserved port number 520.



Open Shortest Path First(OSPF)

- Open: publicly available
- Dynamic algorithm, adapts to changes in the topology automatically and quickly.
- Supports routing based on type of service and a variety of distance metrics, including
  - Physical distance, delay, etc.
  - For each link, multiple cost metrics for different TOS (eg, satellite link cost set "low" for best effort; high for real time)
- Multiple same-cost paths allowed (only one path in RIP): achieves load balancing, splitting the load over multiple lines
- Hierarchical OSPF in large domains
- Security: all OSPF messages authenticated (to prevent malicious intrusion); TCP connections used. Supports *tunneling*.
- Integrated uni- and multicast support:
  - Multicast OSPF (MOSPF) uses same topology data base as OSPF

OSPF Operation



(a) An autonomous system. (b) A graph representation of (a).

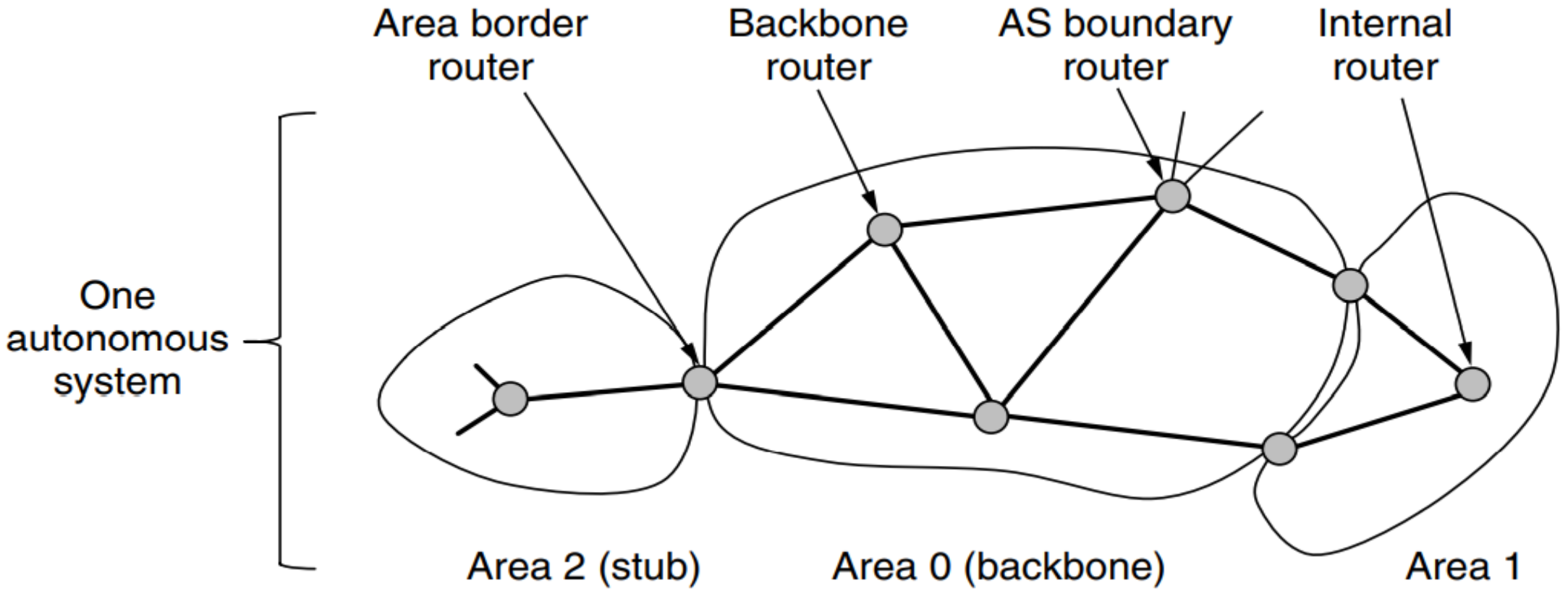
OSPF Operation

- OSPF represents the actual network as a graph and then use the link state method to have every router compute the shortest path from itself to all other nodes.
- Multiple paths may be found that are equally short.
- In this case, OSPF remembers the set of shortest paths and during packet forwarding, traffic is split across them. This helps to balance load.
- It is called ECMP (Equal Cost MultiPath).

OSPF Operation – Large ASes

- Many of the ASes in the Internet are themselves large and nontrivial to manage.
- To work at this scale, OSPF allows an AS to be divided into numbered areas, where an area is a network or a set of contiguous networks.
- An area is a generalization of an individual network.
- Outside an area, its destinations are visible but not its topology. This characteristic helps routing to scale.
- Areas do not overlap but need not be exhaustive, that is, some routers may belong to no area.
- Routers that lie wholly within an area are called internal routers.

OSPF Operation – Large ASes



The relation between ASes, backbones, and areas in OSPF.

OSPF Operation – Large ASes

- Every AS has a backbone area, called **area 0**.
- The routers in this area are called **backbone routers**.
- All areas are connected to the backbone, so it is possible to go from any area in the AS to any other area in the AS via the backbone.
- As with other areas, the topology of the backbone is not visible outside the backbone.
- Each router that is connected to two or more areas is called an area border router. It must also be part of the backbone.
- The job of an area border router is to summarize the destinations in one area and to inject this summary into the other areas to which it is connected.

OSPF Operation – Large ASes

- The summary includes cost information but not all the details of the topology within an area.
- Passing cost information allows hosts in other areas to find the best area border router to use to enter an area.
- Not passing topology information reduces traffic and simplifies the shortest-path computations of routers in other areas.
- However, if there is only one border router out of an area, even the summary does not need to be passed, just the instruction ‘‘Go to the border router.’’
- This kind of area is called a **stub area**.

OSPF Operation – Large ASes

- The last kind of router is the AS boundary router. It injects routes to external destinations on other ASes **into the area**.
- The external routes then appear as destinations that can be reached via the AS boundary router with some cost.
- An external route can be injected at one or more AS boundary routers.
- One router may play multiple roles, for example, a border router is also a backbone router.

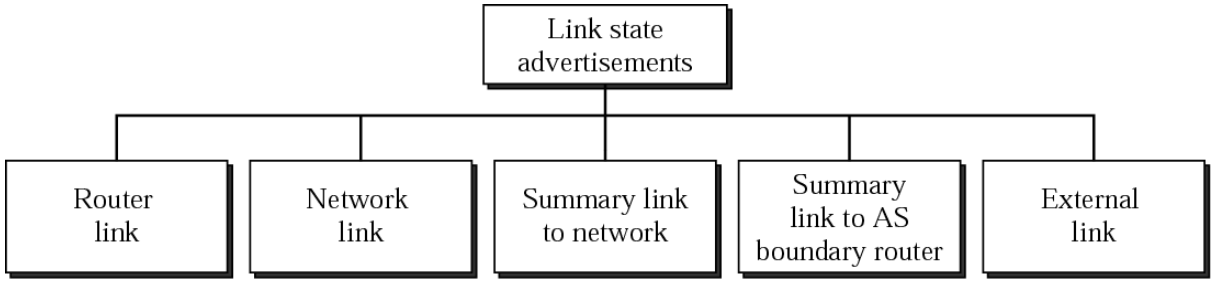
OSPF Operation – Large ASes

- When a router boots, it sends HELLO messages on all of its point-to-point lines and multicasts them on LANs to the group consisting of all the other routers.
- From the responses, each router learns who its neighbors are. Routers on the same LAN are all neighbors.
- During normal operation, each router periodically floods LINK STATE UPDATE messages to each of its adjacent routers. These messages gives its state and provide the costs used in the topological database.
- The flooding messages are acknowledged, to make them reliable. Each message has a sequence number, so a router can see whether an incoming LINK STATE UPDATE is older or newer than what it currently has.
- Routers also send these messages when a link goes up or down or its cost changes.

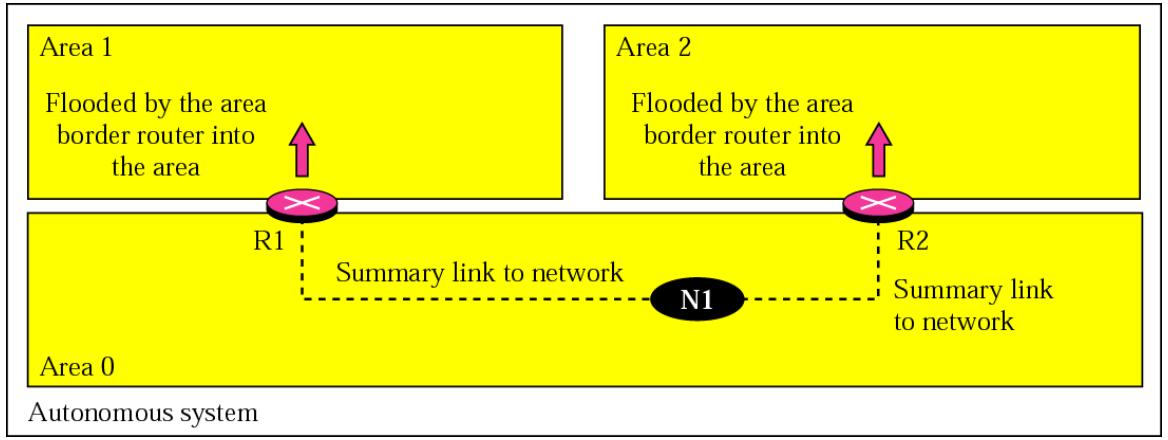
OSPF Operation – Large ASes

- **DATABASE DESCRIPTION** messages give the sequence numbers of all the link state entries currently held by the sender. These messages are used when a link is brought up.
- By comparing its own values with those of the sender, the receiver can determine who has the most recent values.
- Either partner can request link state information from the other one by using **LINK STATE REQUEST** messages.
- The result of this algorithm is that each pair of adjacent routers checks to see who has the most recent data, and new information is spread throughout the area this way.
- All these messages are sent directly in IP packets.

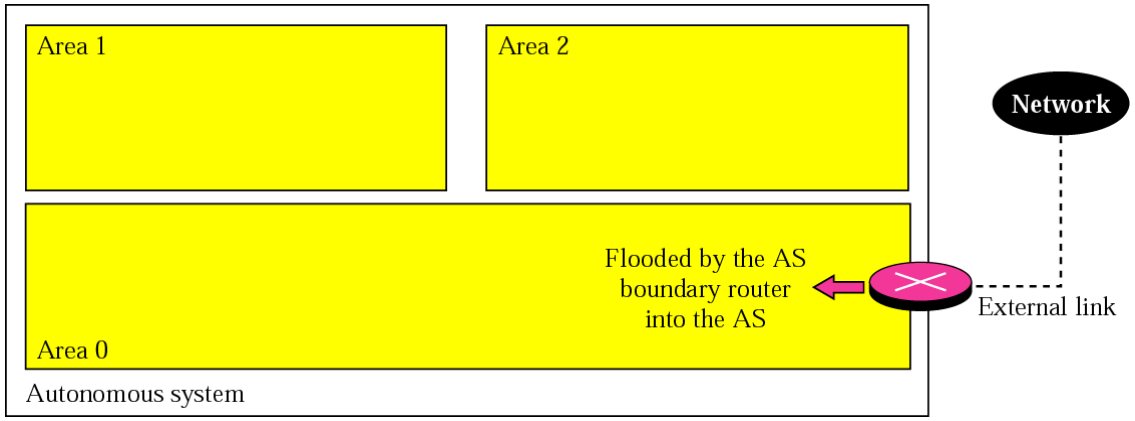
OSPF: Link State Advertisement



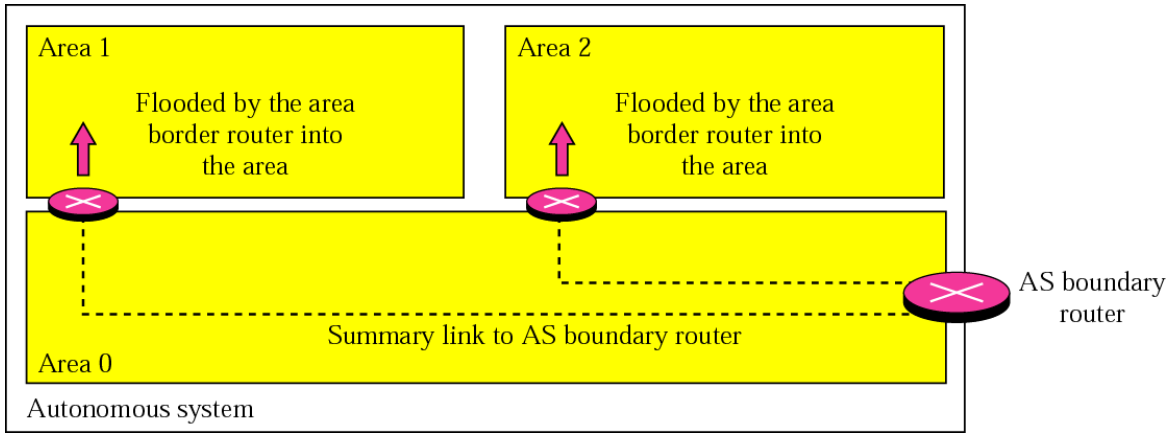
Summary link to Network



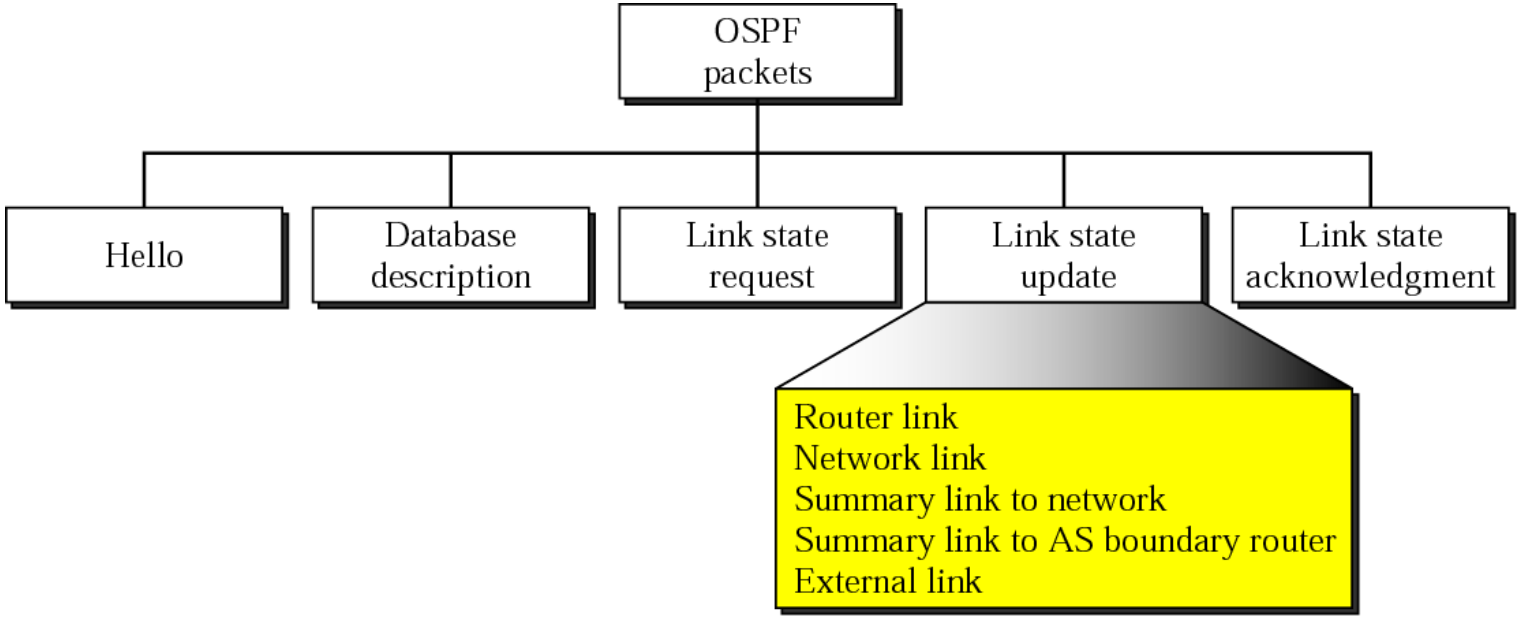
External Link



Summary link to AS boundary router



Types of OSPF Packets and Header Format



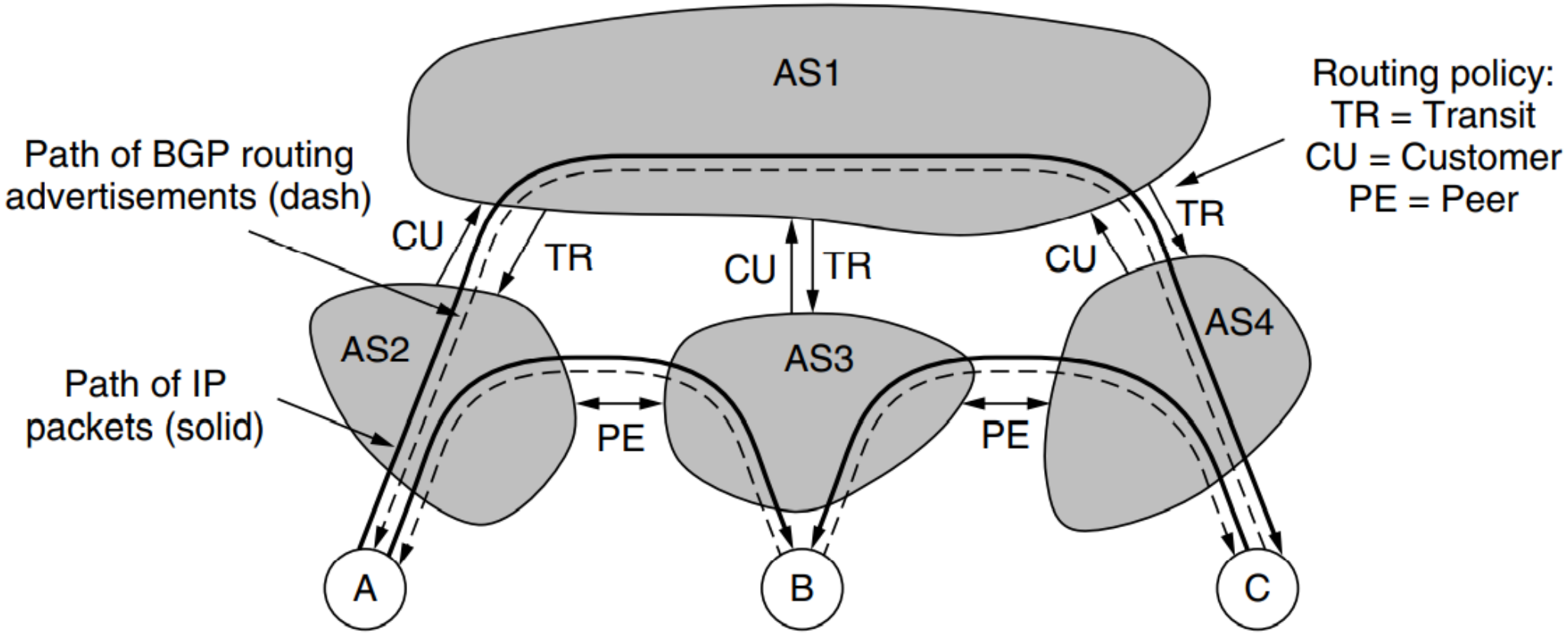
Version	Type	Message length
Source router IP address		
Checksum		Authentication type
Authentication		

Message type	Description
Hello	Used to discover who the neighbors are
Link state update	Provides the sender's costs to its neighbors
Link state ack	Acknowledges link state update
Database description	Announces which updates the sender has
Link state request	Requests information from the partner

Border Gateway Protocol (BGP)

- *The* de facto standard, the current version is 4, known as BGP4
- The initial protocol (called EGP) was designed for specialized topology, such as a tree topology.
- BGP replaces EGP – generalizes the topology structure of the Internet.
- BGP assumes that the Internet is an arbitrary interconnected set of ASs.
  - **Local Traffic:** Originates at or terminates on nodes within an AS
  - **Transit Traffic:** Traffic that passes through an AS

Traffic through an AS



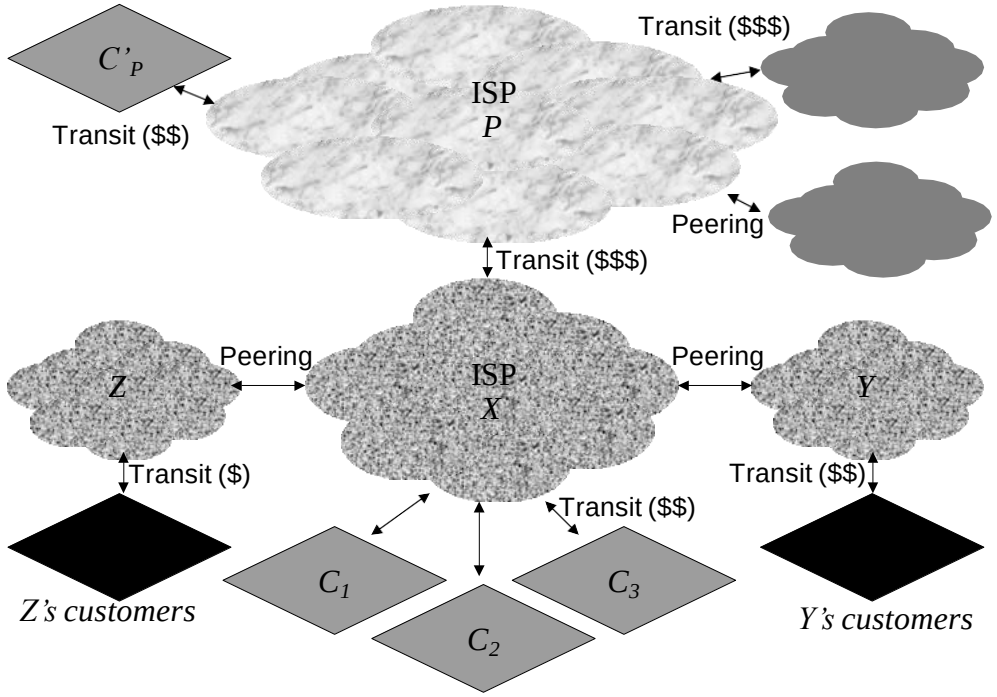
Routing policies between four ASes

Border Gateway Protocol (BGP)

- There are four Ases that are connected. The connection is often made with a link at IXPs (Internet eXchange Points), facilities to which many ISPs have a link for the purpose of connecting with other ISPs.
- AS2, AS3, and AS4 are customers of AS1. They buy transit service from it.
- When source A sends to destination C, the packets travel from AS2 to AS1 and finally to AS4.
- The routing advertisements travel in the opposite direction to the packets.
- AS4 advertises C as a destination to its transit provider, AS1, to let sources reach C via AS1.
- Later, AS1 advertises a route to C to its other customers, including AS2, to let the customers know that they can send traffic to C via AS1.

Transit vs. Peering

- **Transit:** the provider charges its customers for Internet access, in return for forwarding packets on behalf of customers to destinations
- **Peering:** two ASes (typically ISPs) provide mutual access to a subset of each other's routing tables, although a business deal, does not involve financial settlement in general.
  - Peering between Tier-1 ISPs ensures that they have explicit default-free routes to all Internet destinations (prefixes)
  - Two As's sometimes set up a transit-free link between each other to forward packets for their direct customers and thus avoid paying transit costs to their respective providers



- A route advertisement from B to A for a destination prefix is an agreement by B that it will forward packets sent via A destined for any destination in the prefix.
- The importing/exporting routes is defined by the decisions targeted towards making or saving money

Border Gateway Protocol (BGP)

- To implement peering, two ASes send routing advertisements to each other for the addresses that reside in their networks.

- Doing so makes it possible for AS2 to send AS3 packets from A destined to B and vice versa.

- Peering is not transitive.

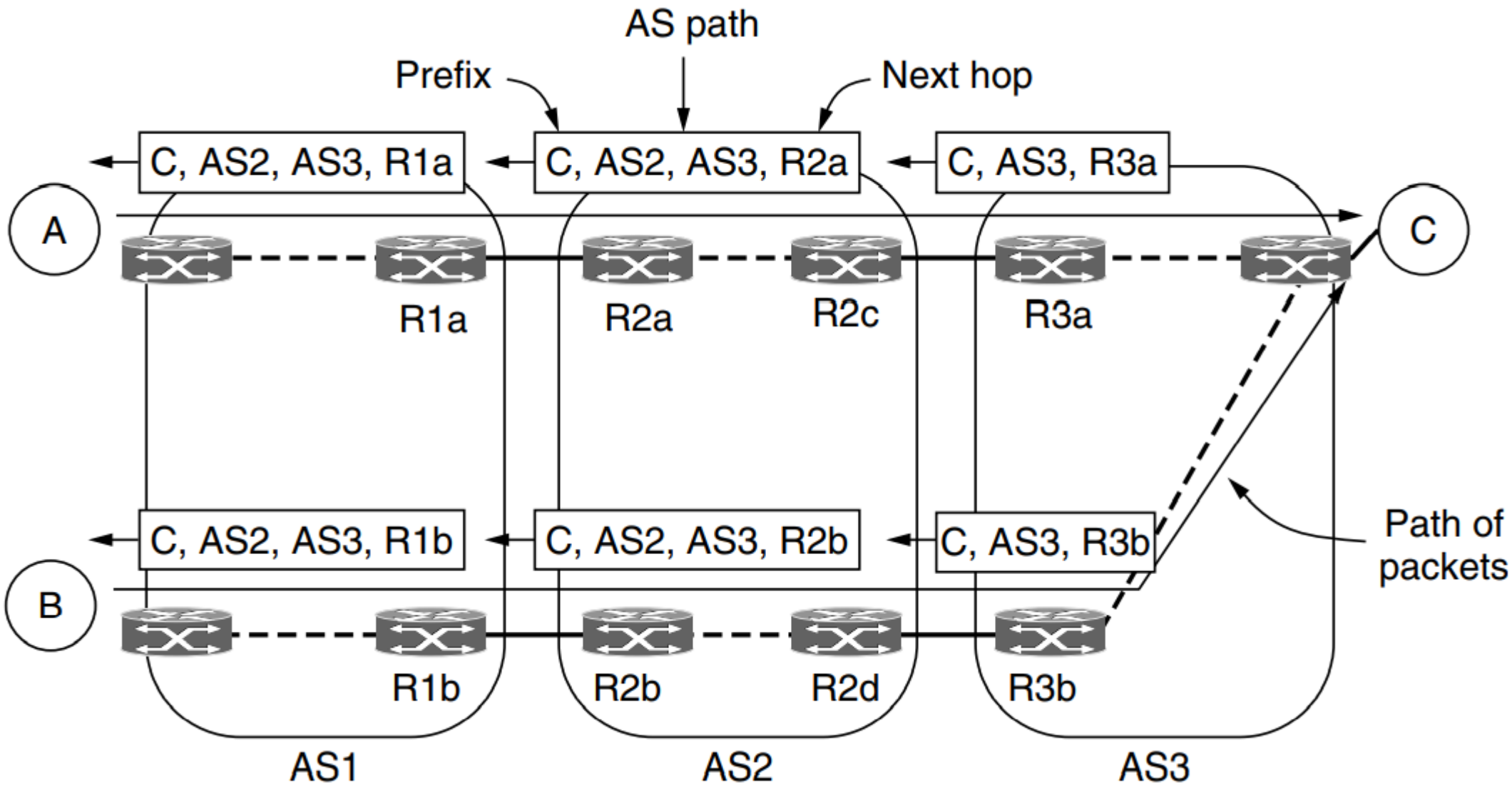
Stub-Network and Multi-homing

- Stub-Network: Network that is connected to the rest of the Internet by only one link. Ex: A, B, C
- Multi-homing: Networks that are connected to multiple ISPs, mainly to improve reliability, since if the path through one ISP fails, the company can use the path via the other ISP.
- In this case, the company network is likely to run an interdomain routing protocol (e.g., BGP) to tell other ASes which addresses should be reached via which ISP links.

BGP Routing

- BGP is a form of distance vector protocol,
  - Policy, instead of minimum distance, is used to pick which routes to use.
  - Instead of maintaining just the cost of the route to each destination, each BGP router keeps track of the path used. This approach is called a **path vector protocol**.
  - The path consists of the next hop router (which may be on the other side of the ISP, not adjacent) and the sequence of ASes, or AS path, that the route has followed (given in reverse order).
  - Pairs of BGP routers communicate with each other by establishing TCP connections.
  - Operating this way provides reliable communication and also hides all the details of the network being passed through.

BGP Routing

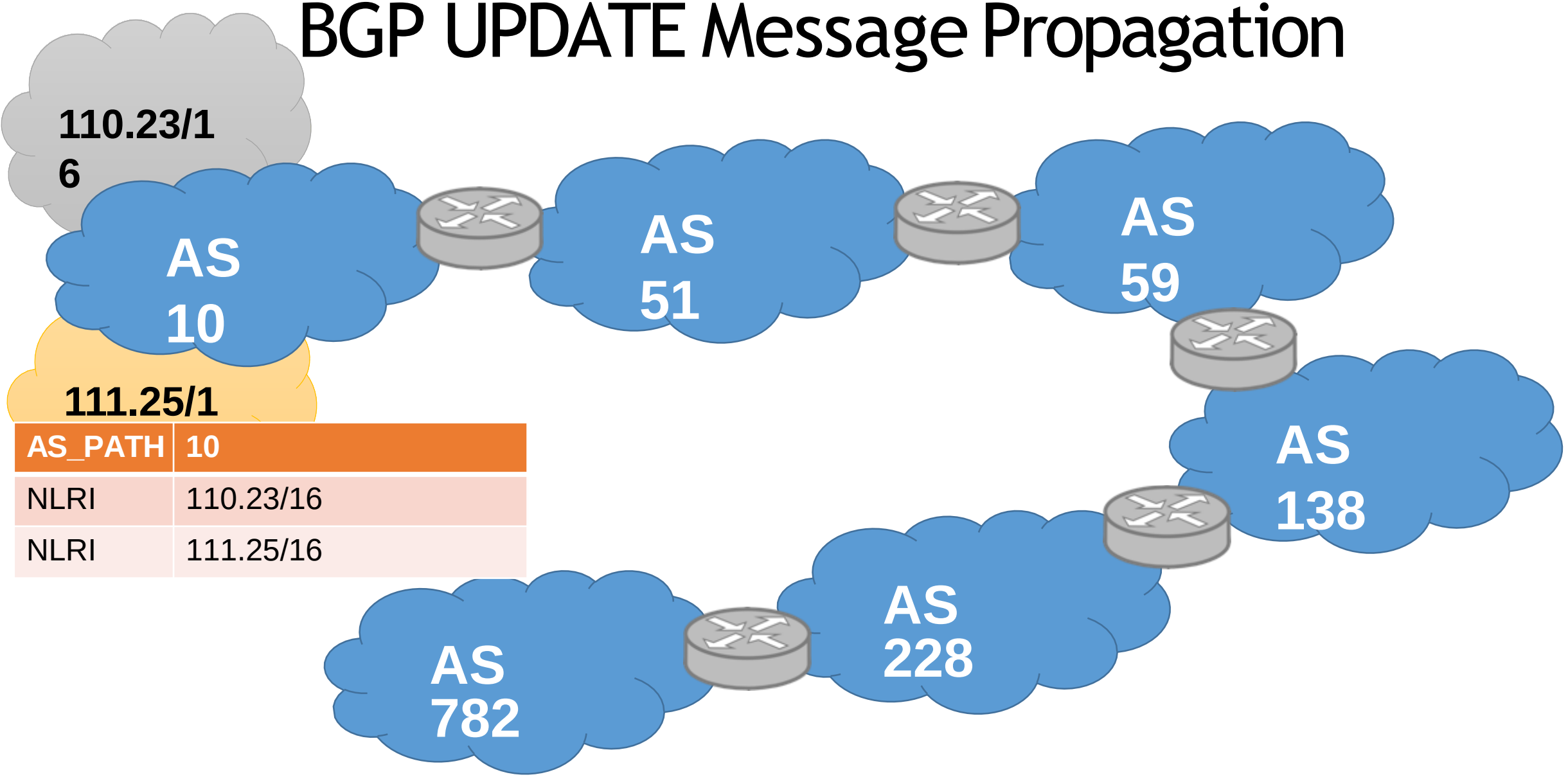

AS path
Prefix
Next hop
C, AS2, AS3, R1a
C, AS2, AS3, R2a
C, AS3, R3a
A
C
R1a
R2a
R2c
R3a
C, AS2, AS3, R1b
C, AS2, AS3, R2b
C, AS3, R3b
Path of packets
B
R1b
R2b
R2d
R3b
AS1
AS2
AS3

BGP Routing

- The rule is that each router that sends a route outside of the AS prepends its own AS number to the route.
- When a router receives a route, it checks to see if its own AS number is already in the AS path. If it is, a loop has been detected and the advertisement is discarded.

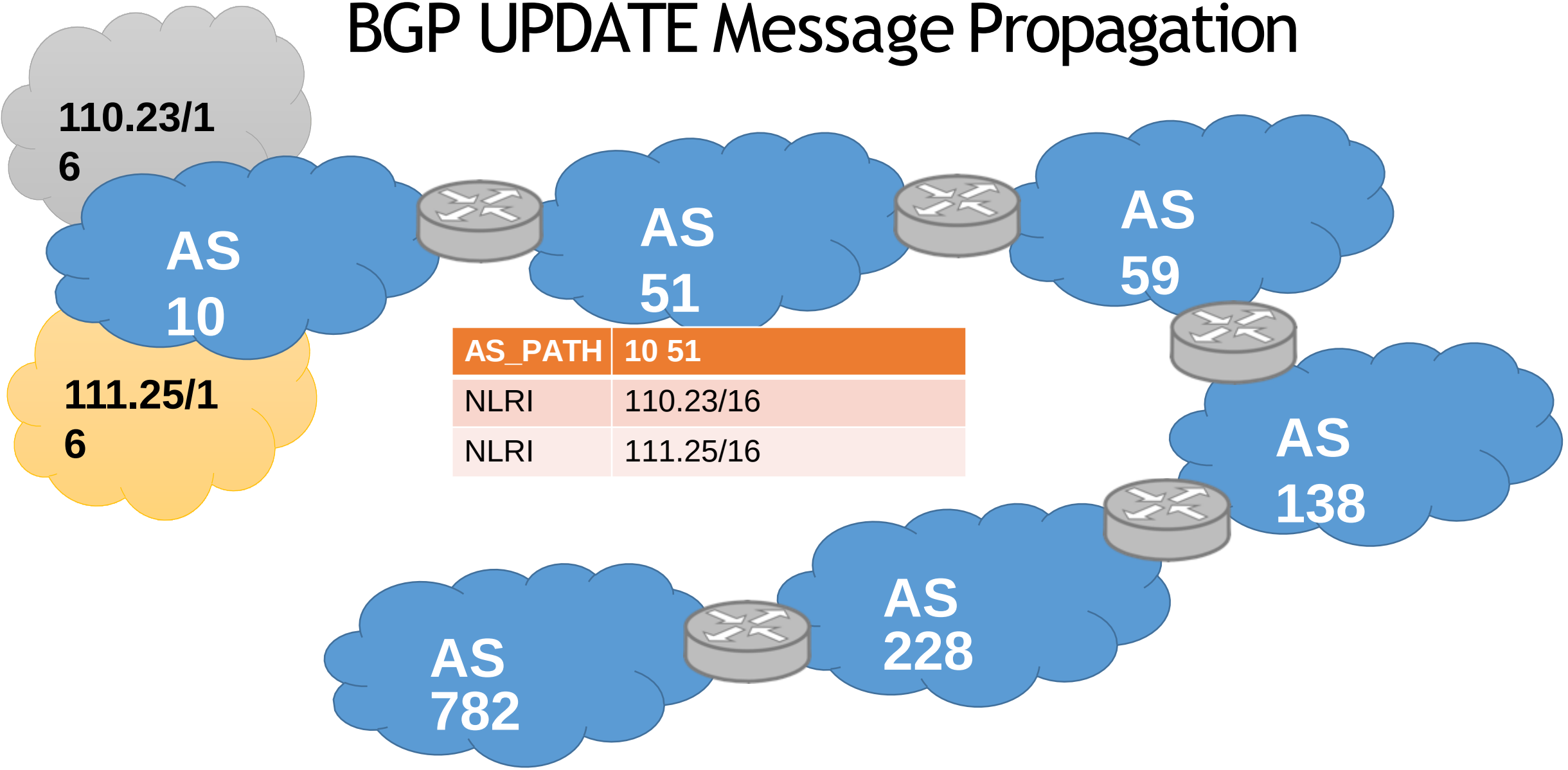
- iBGP: The task of propagating BGP routes from one side of the ISP to the other is handled by a intra-domain variant of BGP protocol called iBGP (internal BGP) to distinguish it from the regular use of BGP as eBGP (external BGP).

BGP UPDATE Message Propagation



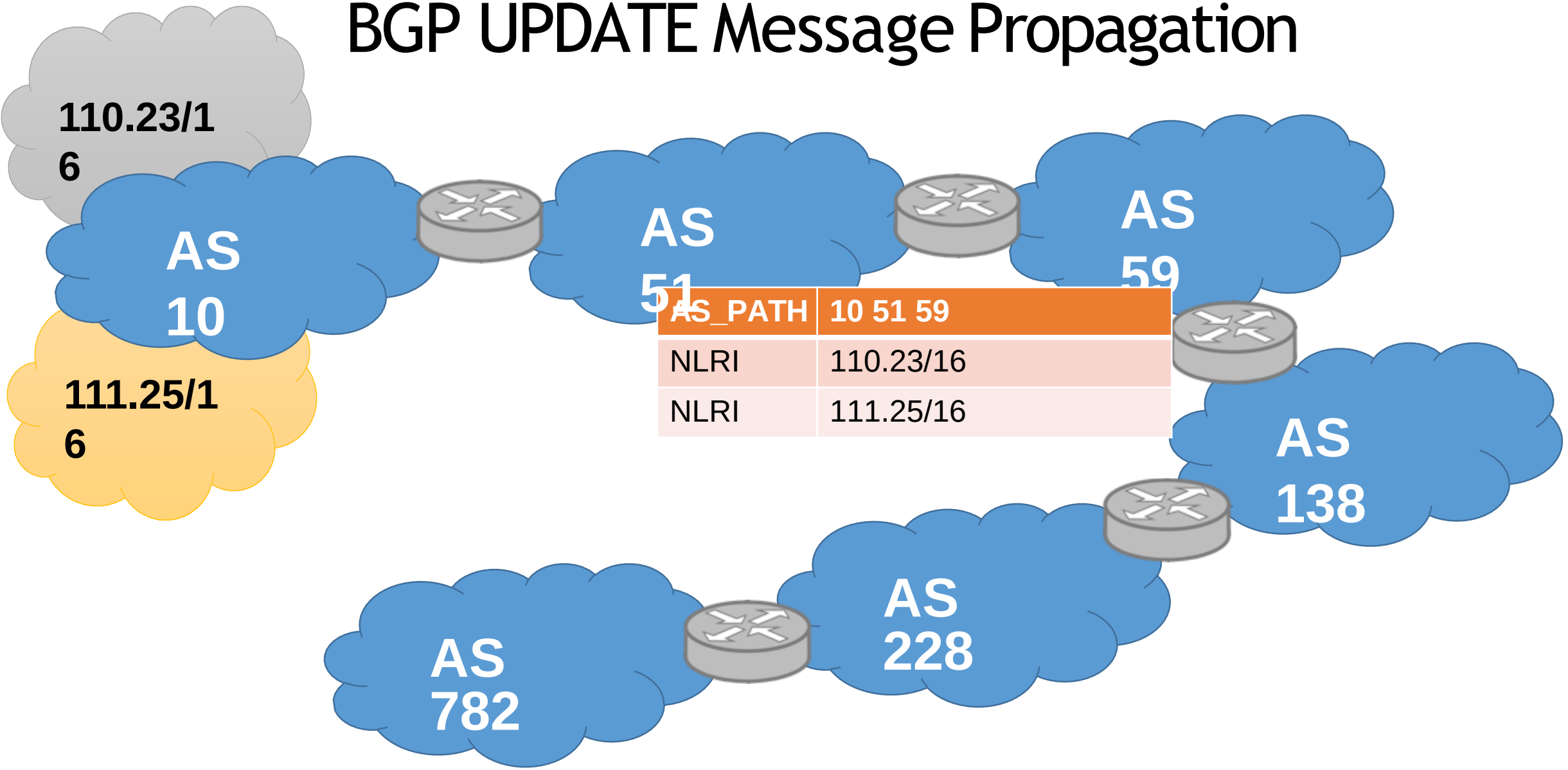
AS_PATH	10
NLRI	110.23/16
NLRI	111.25/16

BGP UPDATE Message Propagation

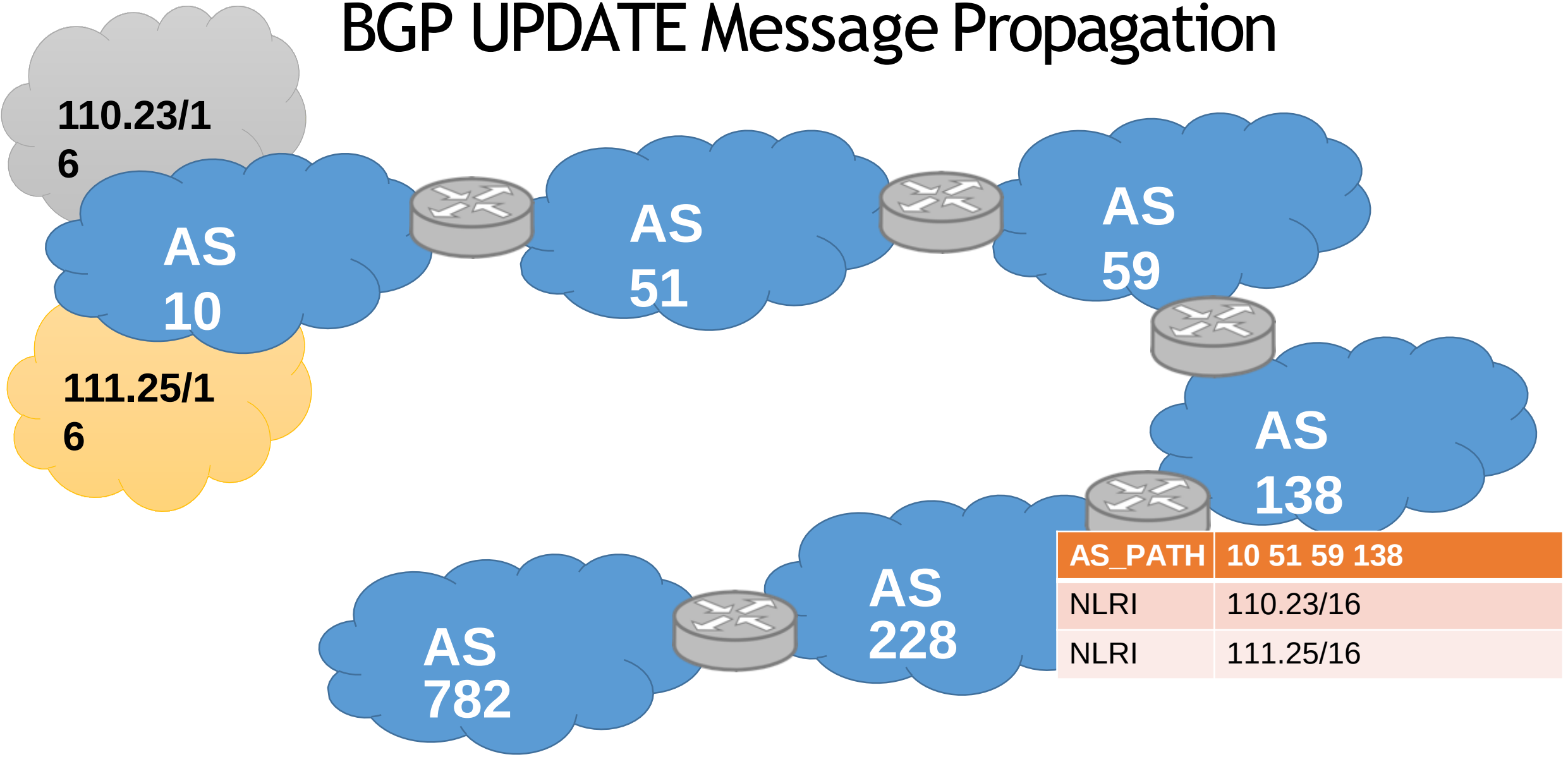


AS_PATH	10 51
NLRI	110.23/16
NLRI	111.25/16

BGP UPDATE Message Propagation

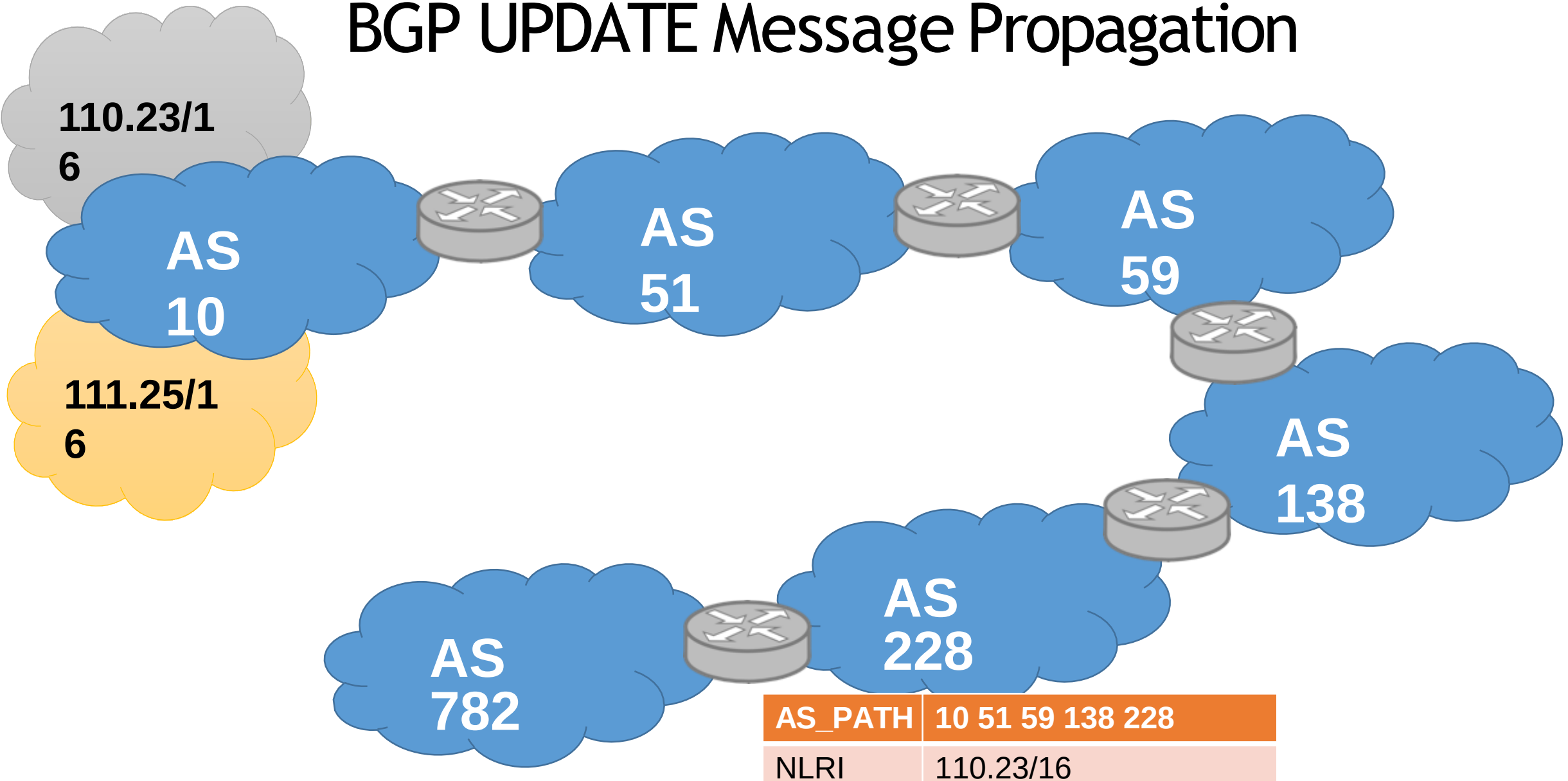


BGP UPDATE Message Propagation



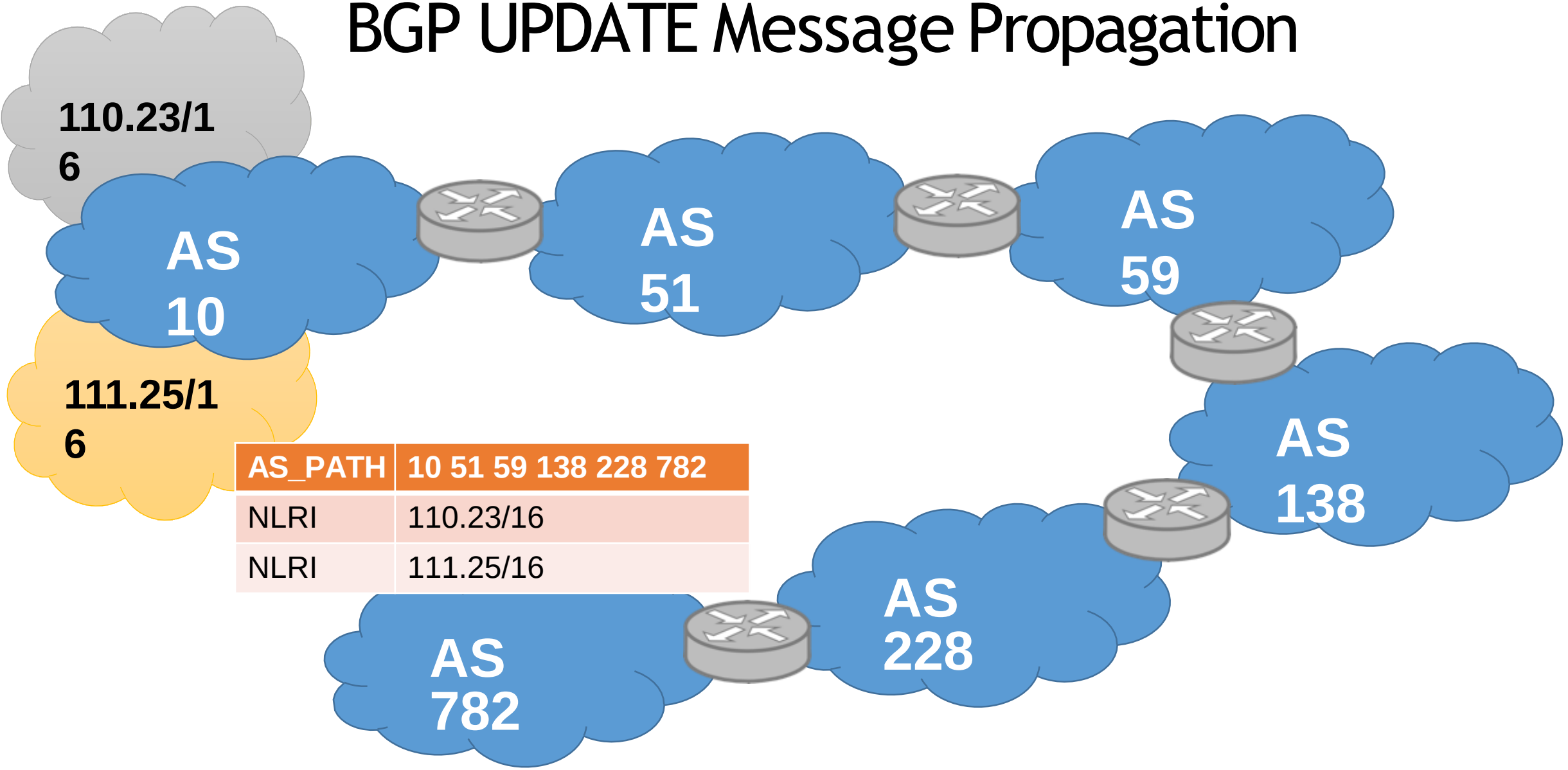
AS_PATH	10 51 59 138
NLRI	110.23/16
NLRI	111.25/16

BGP UPDATE Message Propagation



AS_PATH	10 51 59 138 228
NLRI	110.23/16
NLRI	111.25/16

BGP UPDATE Message Propagation



AS_PATH	10 51 59 138 228 782
NLRI	110.23/16
NLRI	111.25/16

Limitation of traditional IP routing

- Header analysis performed at each hop
- Increased demand on routers
- Utilizes the best available path
- Some congested links and some underutilized links!
  - Degradation of throughput
  - Long delays
  - More losses
- No QoS
  - No service differentiation
  - Not possible with connectionless protocols

Need for MPLS

- Rapid growth of Internet
- New *latency dependent* applications
- Quality of Service (QoS)
  - Less time at the routers
- Traffic Engineering
  - Flexibility in routing packets
- Connection-oriented forwarding techniques with connectionless IP
  - Utilizes the IP header information to maintain interoperability with IP based networks
  - Decides on the path of a packet before sending it

Multi Protocol Label Switching (MPLS)

- Multi Protocol – supports protocols even other than IP
  - Supports IPv4, IPv6, IPX, AppleTalk at the network layer
  - Supports Ethernet, Token Ring, FDDI, ATM, Frame Relay, PPP at the link layer
- Label – short fixed length identifier to determine a route
  - Labels are added to the top of the IP packet
  - Labels are assigned when the packet enters the MPLS domain
- Switching – forwarding a packet
  - Packets are forwarded based on the label value
  - NOT on the basis of IP header information

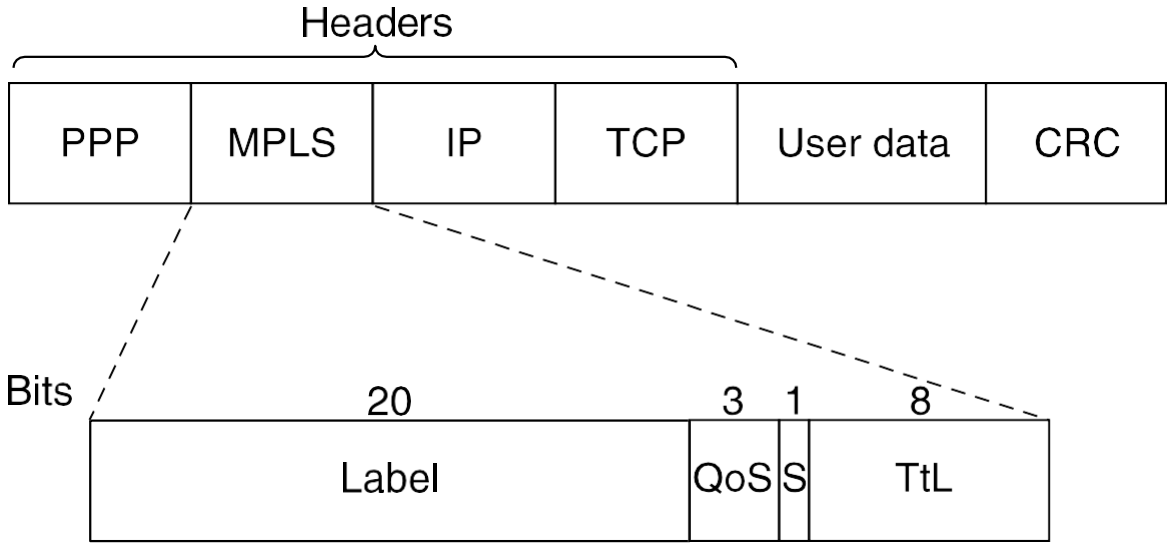
MPLS Background

- Integration of layer 2 and layer 3
  - Simplified connection-oriented forwarding of layer 2
  - Flexibility and scalability of layer 3 routing
- MPLS does not replace IP; it supplements IP
- Traffic can be marked, classified and explicitly routed
- QoS can be achieved through MPLS

IP versus MPLS

- Routing decisions
  - IP routing – based on destination IP address
  - Label switching – based on labels
- Entire IP header analysis
  - IP routing – performed at each hop of the packets path in the network
  - Label switching – performed only at the ingress router
- Support for unicast and multicast data
  - IP routing – requires special multicast routing and forwarding algorithms
  - Label switching – requires only one forwarding algorithm

MPLS Header Format



- Label: 20-bit label value
- QoS: indicates the class of service
- S: bottom of stack indicator
  - 1 for the bottom label, 0 otherwise
- TTL: time to live

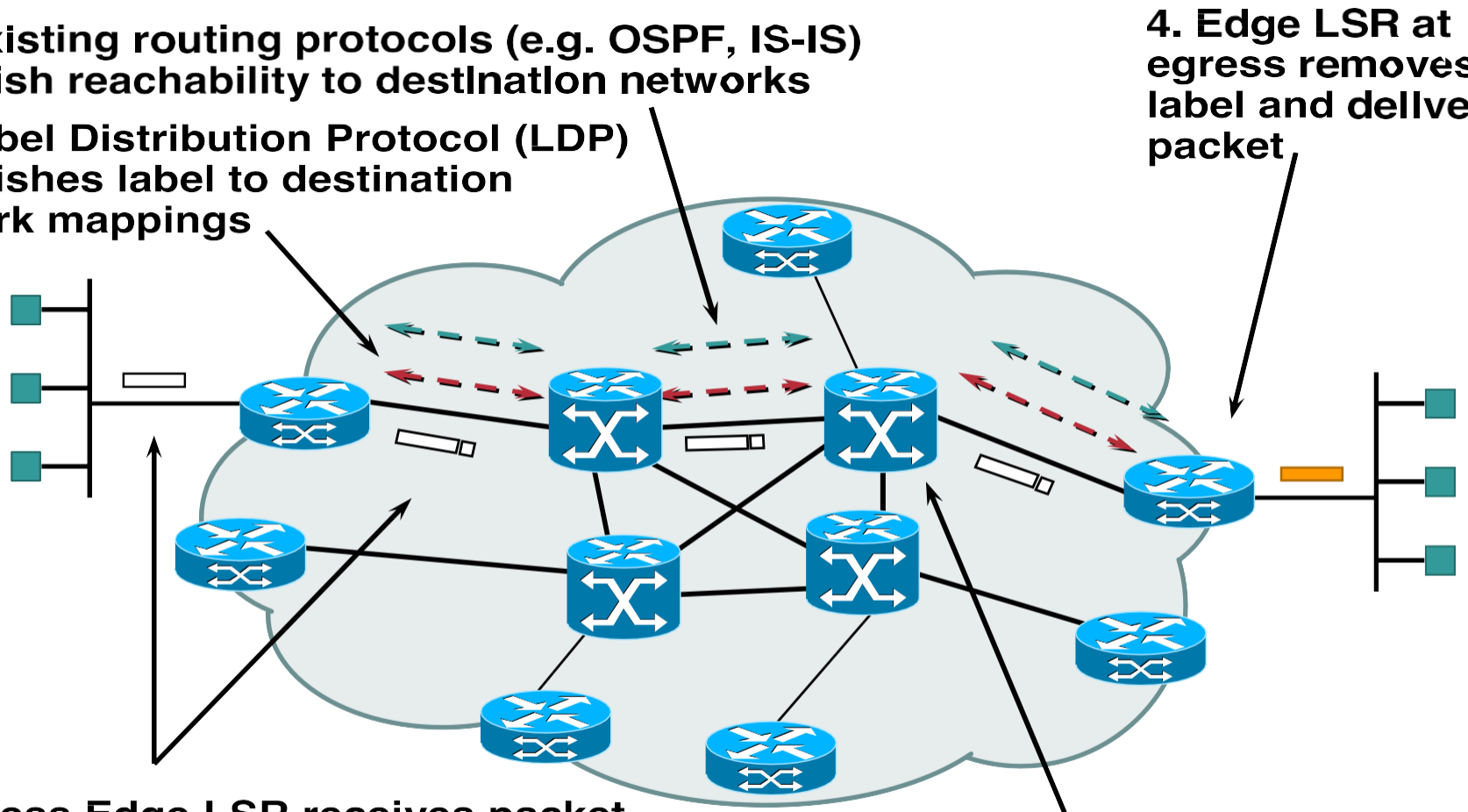
Forwarding Equivalence Class (FEC)

- A group of packets that require the same forwarding treatment across the same path
- Packets are grouped based on any of the following
  - Address prefix
  - Host address
  - Quality of Service (QoS)
- FEC is encoded as the label

Labels

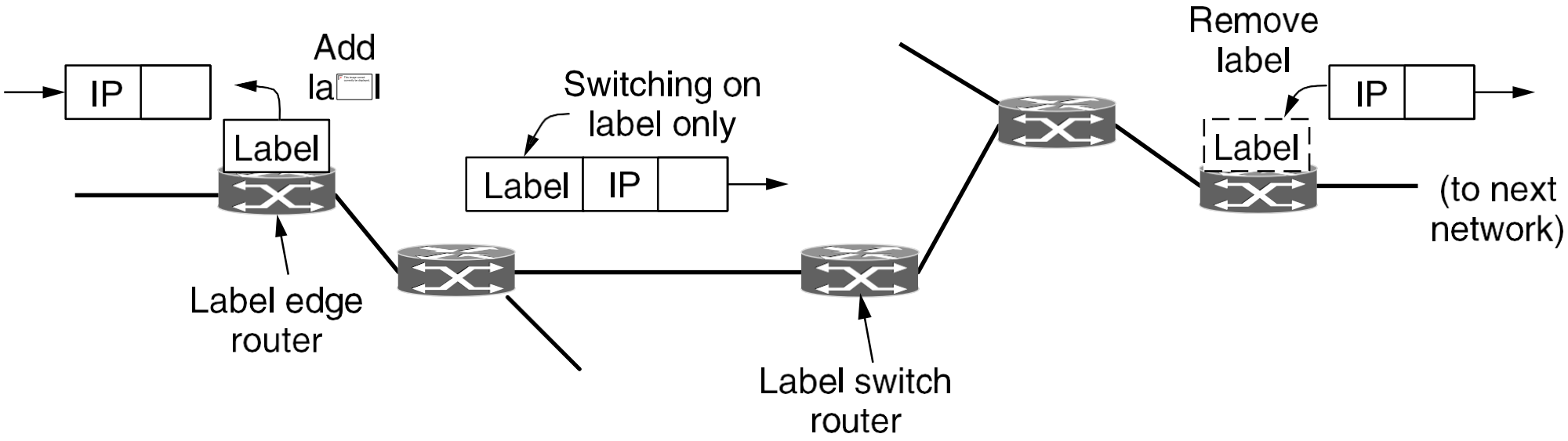
- A short, fixed length identifier (20 bits)
- Sent with each packet
- Local between two routers
- Can have different labels if entering from different routers
- One label for one FEC
- Decided by the downstream router (the receiver)
  - LSR (Label Switching Router) binds a label to an FEC
  - It then informs the upstream LSR (the sender router) of the binding

MPLS Operation



Source: cisco.com

MPLS Forwarding



Label Stacking

- A packet may carry multiple labels, organized as a last-in-first-out stack
- A label may be added to/removed from the stack at any LSR
- Processing always done on the top label
- Allow the aggregation of LSPs into a single LSP for a portion of the route, creating a tunnel
  - At the beginning of the tunnel, the LSR assigns the same label to packets from different LSPs by pushing the label onto each packet’s stack
  - At the end of the tunnel, the LSR pops the top label

MPLS versus Virtual CircuitSwitching

Virtual Circuit Switching	MPLS
It is not possible to group several distinct paths with different endpoints onto the same virtual-circuit identifier because there would be no way to distinguish them at the final destination.	Label aggregation is possible in MPLS. Each flow can have its own set of labels and the routers group multiple flows (belongs to the same FEC) and use a single label for them.
Has to set up many label switching paths, one for each of the different labels when many packets with different labels need to follow a common path to some destination.	Can operate at multiple levels at once by adding more than one label to the front of a packet through label stacking.
When a user wants to establish a connection, a setup packet is launched into the network to create the path and make the forwarding table entries.	MPLS does not involve users in the setup phase.