1.5. Internet Routing

PROFESSORS:

Rubén Santiago Montero

Eduardo Huedo Cuesta

OTHER AUTHORS:

Rafael Moreno Vozmediano

Juan Carlos Fabero Jiménez

Introduction: The Routing Problem

- In a packet switching network, routing consists in finding a route, from source to destination, through the intermediate switching nodes or routers
- **Alternative routes**
  - It is necessary to decide which is the best (i.e. shortest or least cost) route
  - The shortest route minimizes a routing metric
- **Routing metrics**
  - **Hop count** has into account the number of routers and/or intermediate networks a packet must traverse to reach the destination
  - **Geographic distance** has into account the distance (in Km) a packet has to travel to reach the destination
  - **Average delay** has into account the delay of transmission lines. Since it is proportional to distance, this metric is similar to the previous one
  - **Bandwidth/speed** has into account the transmission speed of communication lines where the packet circulates
  - **Traffic level** has into account the usage level of the communication lines, to try to use those with less saturation level
  - Linear combination of several metrics

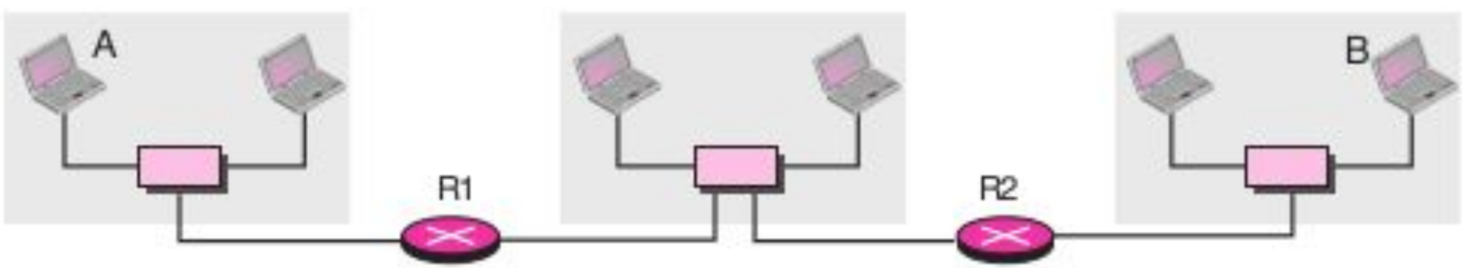
Packet Forwarding

- When a router receives a packet, it forwards it through the appropriate link to reach the destination
- Link selection is performed according to:
    - **Routing tables:** Using the destination address field in the IP packet (connectionless)
        - Based on the next hop
        - Table entries (routes) per host, network or by default
        - Network destinations can be classful or classless
    - **Labels:** Each IP datagram is labelled and then switched according to that label (connection oriented)
        - Reduces complexity in routing table
        - Same circuit/route used (in-order delivery, predictable delay...)
        - Flow Label field in IPv6 header
        - MPLS (MultiProtocol Label Switching)

Routing Tables

Next-Hop Routing

- Based on Bellman’s **principle of optimality**: if the shortest route between two routers, A and B, is through C, then the shortest route between C and B is through the same route
  - To route the packet throughout the shortest route, we only need the address of the next immediate router



A

Destination	Route
Host B	R1, R2, Host B

R1

Destination	Route
Host B	R2, Host B

R2

Destination	Route
Host B	Host B

a. Routing tables based on route

A

Destination	Next Hop
Host B	R1

R1

Destination	Next Hop
Host B	R2

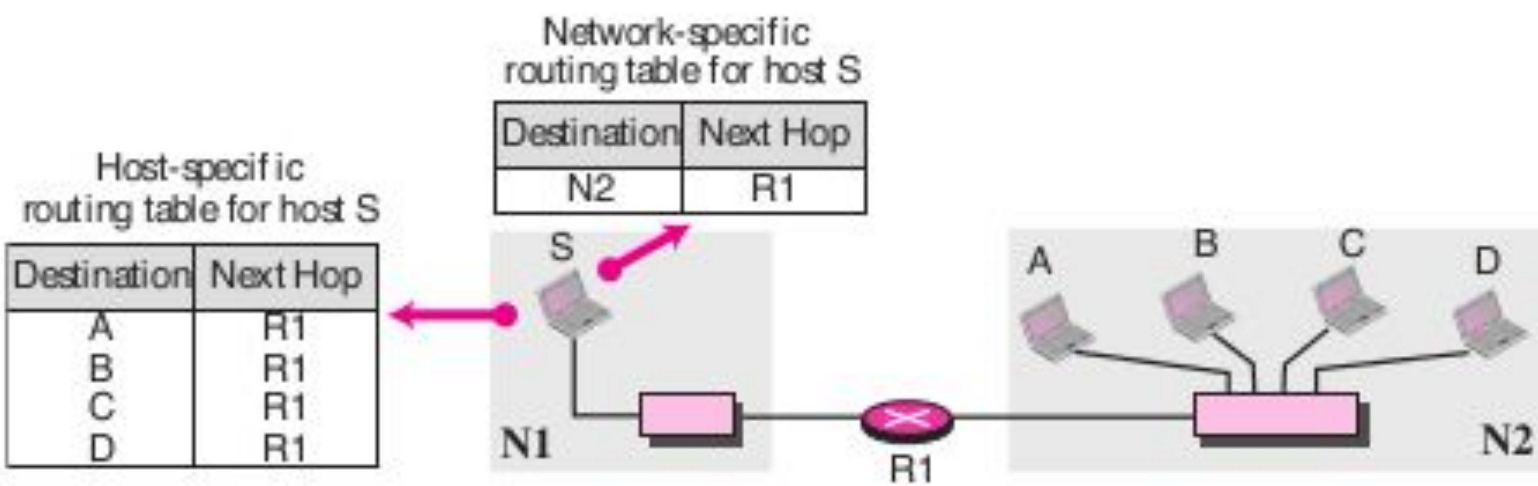
R2

Destination	Next Hop
Host B	---

b. Routing tables based on next hop

Routing Tables

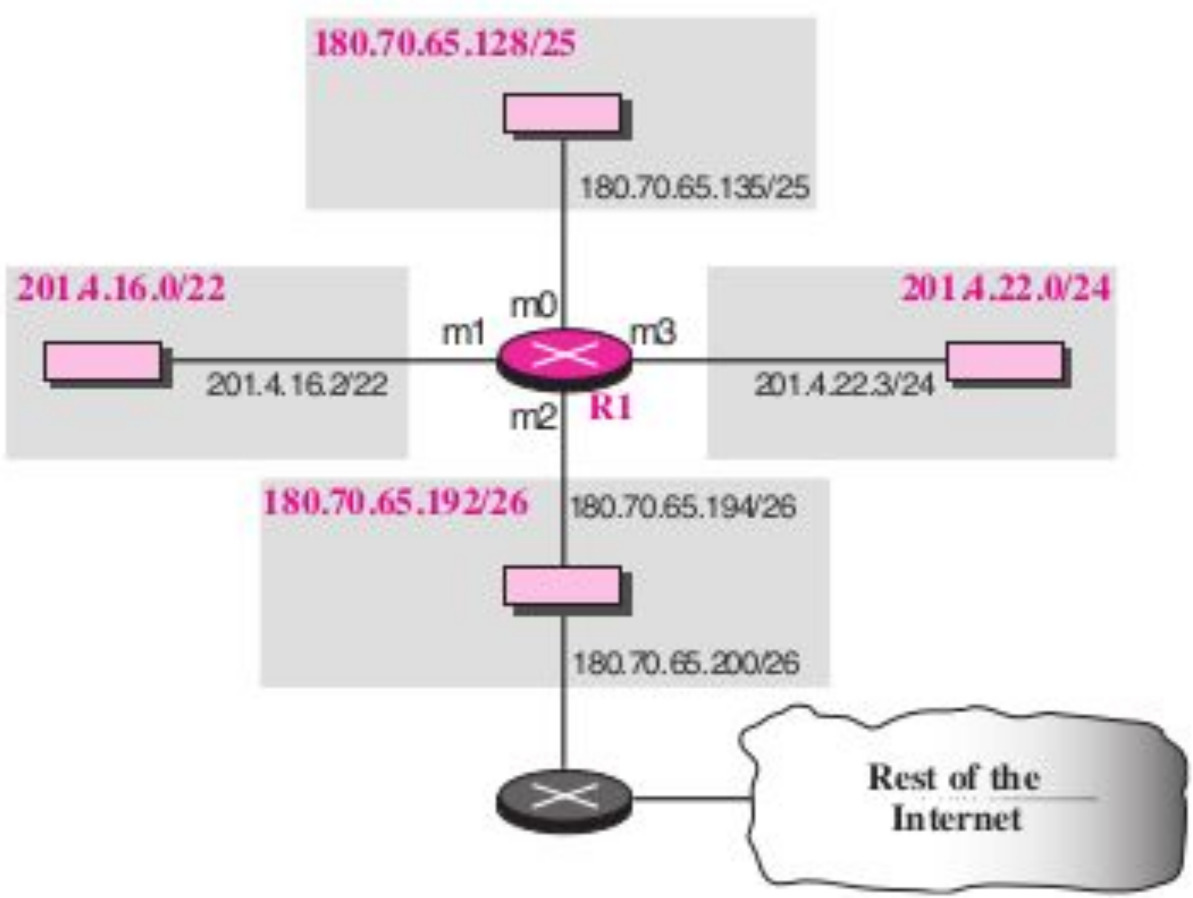
- In general, a routing table stores information about:
  - Destination
  - Network mask or prefix length (CIDR)
  - Interface (for direct delivery) and/or next hop (for indirect delivery)
  - Metric associated to the route
- Destination could be:
  - A host (not viable for Internet routing)
  - A network: For classless networks, prefix lengths are needed
  - Default: Route for packets not matching any destination



Routing Tables

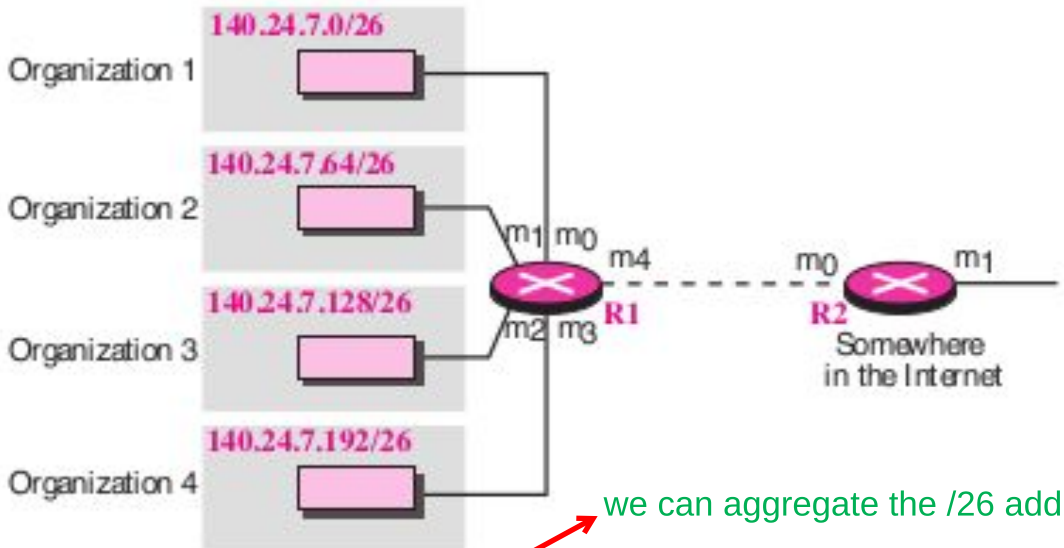
Example: Given the following network topology:

- Determine the routing table for router R1
- Describe the processing of two packets with destination address `201.4.22.35` and `18.24.32.78`, respectively



Routing Tables

- Scalable routing in the Internet depends on controlling the size of routing tables
  - Classful routing is not viable, due to the high number of networks (i.e. table entries) in the Internet
- Internet routing is based on CIDR and hierarchical routing:
  - CIDR allows address aggregation or route summarization
  - Hierarchical routing limits the information exchanged



we can aggregate the /26 addresses into the /24 one

Mask	Network address	Next-hop address	Interface
/26	140.24.7.0	----------	m0
/26	140.24.7.64	----------	m1
/26	140.24.7.128	----------	m2
/26	140.24.7.192	----------	m3
/0	0.0.0.0	default router	m4

Routing table for R1

Mask	Network address	Next-hop address	Interface
/24	140.24.7.0	----------	m0
/0	0.0.0.0	default router	m1

Routing table for R2

Routing Techniques

Local routing

- It doesn’t take into account network topology, using only local information
- Common techniques are:
  - Random routing
  - Isolated routing
  - Flooding

Static routing

- It takes into account network topology
- Routing tables are manually created and they don’t adapt to network changes

Dynamic routing

- Routing tables are automatically created, by means of periodic exchange of information between routers
- Allows automatically adapting routing to changes in network topology
- Common techniques are:
  - Distance vector routing (e.g. RIP)
  - Link state routing (e.g. OSPF)
  - Path vector routing (e.g. BGP)

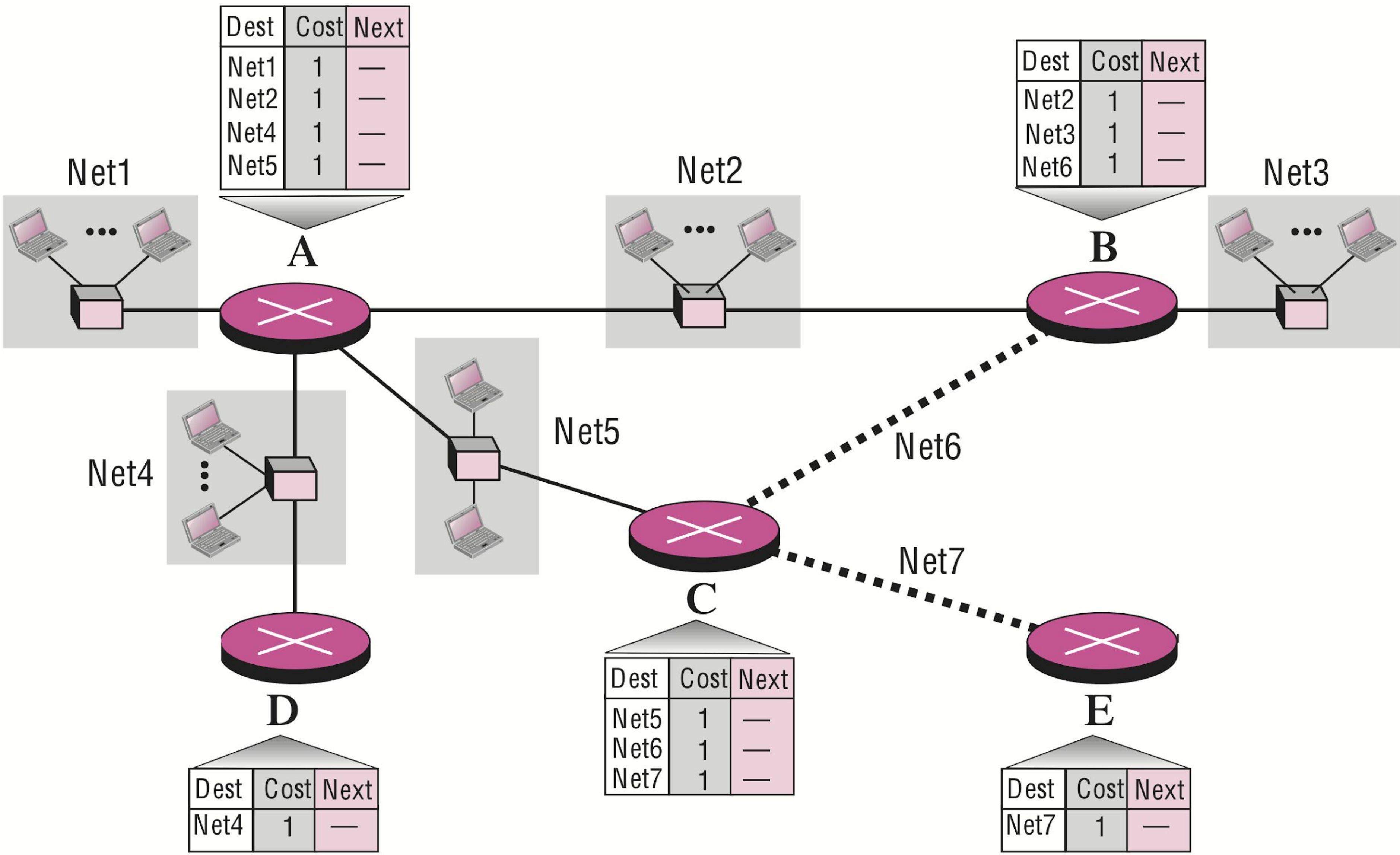
Distance Vector Routing

Fundamentals

- Each router maintains a routing table with an entry for each possible destination
- Each table entry contains:
  - Destination (usually a network, or a host)
  - Next node or router to reach that destination
  - Distance or metric to destination
- To create the routing table, routers periodically exchange information (distance vector, with destination and distance) with their neighbors
  - Total distance to each destination is the distance announced by the router plus the distance to the router distance usually measured in number of hops
  - If total distance is lower than the distance in the current route, the route is replaced
- The iterative process of distance vector exchange ideally converges to the optimal routes
  - Distance vector routing is also called Bellman-Ford’s algorithm
- The cost or distance metric is usually the number of hops
- Example: RIP (Routing Information Protocol)

Distance Vector Routing

Example: Initially, routers only know direct routes

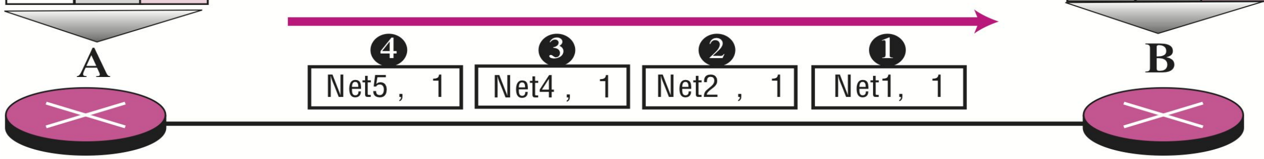


Distance Vector Routing

Example: Exchange process

Dest	Cost	Next
Net1	1	—
Net2	1	—
Net4	1	—
Net5	1	—

Dest	Cost	Next
Net2	1	—
Net3	1	—
Net6	1	—



Routing Table B

Dest	Cost	Next
Net1	2	A
Net2	1	—
Net3	1	—
Net6	1	—

After receiving record 1

Routing Table B

Dest	Cost	Next
Net1	2	A
Net2	1	—
Net3	1	—
Net6	1	—

After receiving record 2

Routing Table B

Dest	Cost	Next
Net1	2	A
Net2	1	—
Net3	1	—
Net4	2	A
Net6	1	—

After receiving record 3

Routing Table B

Dest	Cost	Next
Net1	2	A
Net2	1	—
Net3	1	—
Net4	2	A
Net5	2	A
Net6	1	—

After receiving record 4

After the exchange of all tables

A

Dest	Cost	Next
Net1	1	—
Net2	1	—
Net3	2	B
Net4	1	—
Net5	1	—
Net6	2	C
Net7	2	C

B

Dest	Cost	Next
Net1	2	A
Net2	1	—
Net3	1	—
Net4	2	A
Net5	2	A
Net6	1	—
Net7	2	C

C

Dest	Cost	Next
Net1	2	A
Net2	2	A
Net3	2	B
Net4	2	A
Net5	1	—
Net6	1	—
Net7	1	—

D

Dest	Cost	Next
Net1	2	A
Net2	2	A
Net3	3	A
Net4	1	—
Net5	1	A
Net6	3	A
Net7	3	A

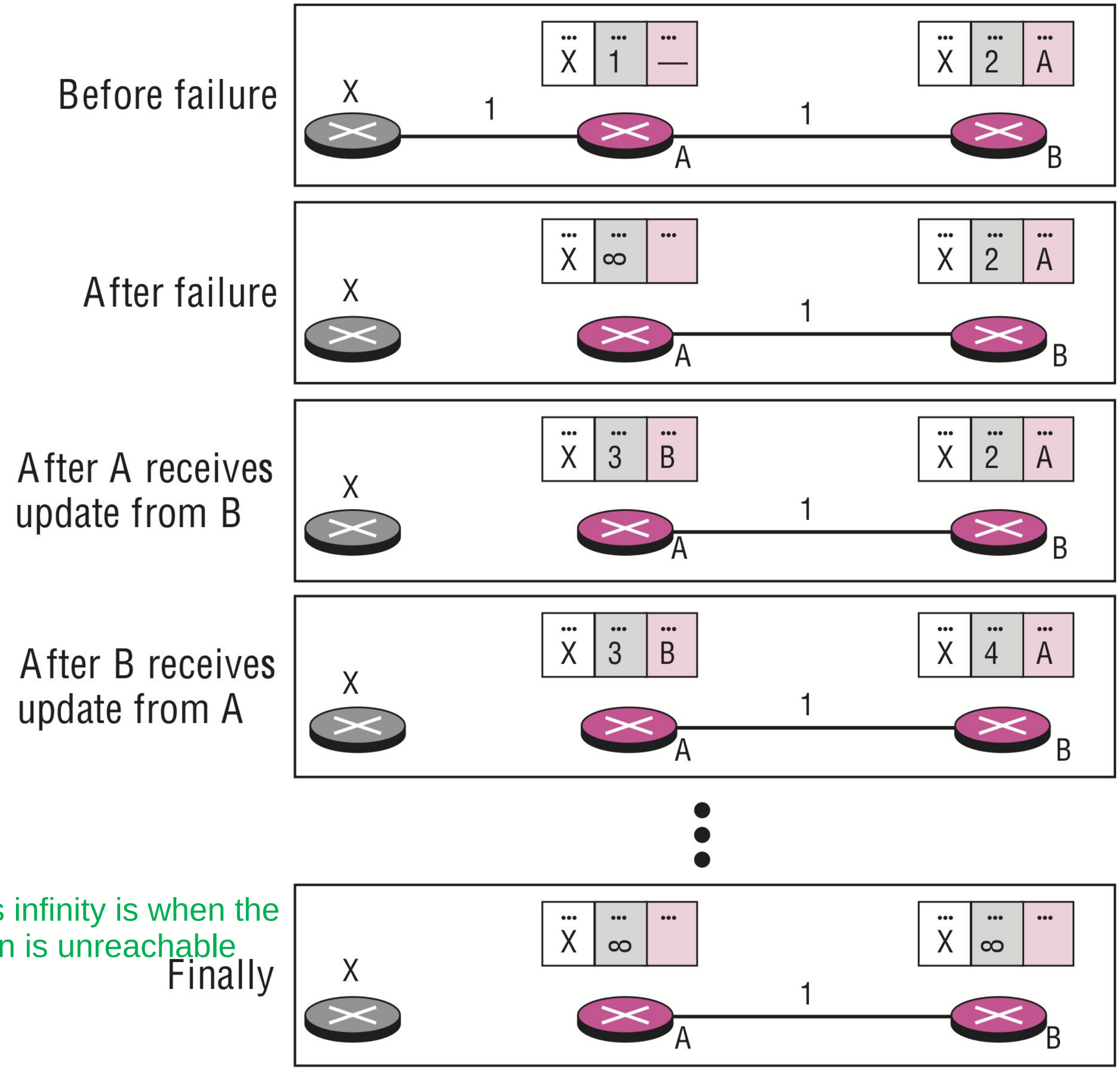
E

Dest	Cost	Next
Net1	3	C
Net2	3	C
Net3	3	C
Net4	3	C
Net5	2	C
Net6	2	C
Net7	1	—

Distance Vector Routing

Convergence problems. Counting to infinity

- Changes in network topology must be propagated to all routers
- When a link increases its distance, this change will propagate slowly
- The needed updates to communicate a broken link may not converge



when the cost reaches infinity is when the rouers think destination is unreachable

Distance Vector Routing

Counting to infinity. Solutions

- **Small infinity**
  - Infinity is set to a small number
  - For example, in RIP a distance of 16 hops is considered infinite (unreachable), therefore there is a limit of 15 hops in a route
- **Split horizon**
  - Routes learned through a given link are not advertised through that link
  - Example: Node B will not send information about destination X to node A
- **Split horizon with poisoned reverse**
  - Routes learned through a given link are advertised through that link, but with an infinite distance
  - Example: Node B will announce to node A that destiny X is at infinite distance, to indicate that the route was learned from A
- **Triggered updates**
  - When a router detects a change in its route table, immediately disseminates this information to its neighbors
  - This way, topology changes are quickly propagated to all routers

Distance Vector Routing

Convergence problems. Loops

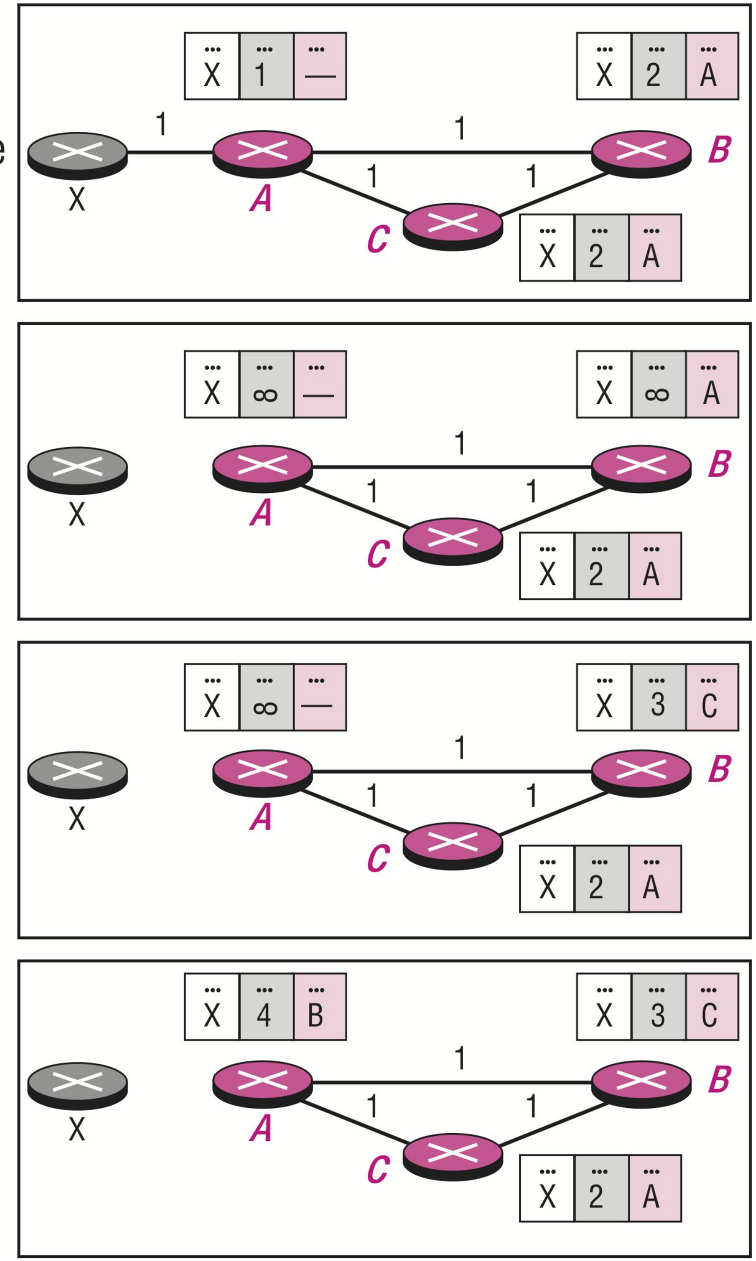
- In networks with loops the algorithm may not converge
- In this case, split horizon techniques don’t solve the problem
- Triggered updates speed up the convergence

Before failure

After A sends the route to B and C, but the packet to C is lost

After C sends the route to B

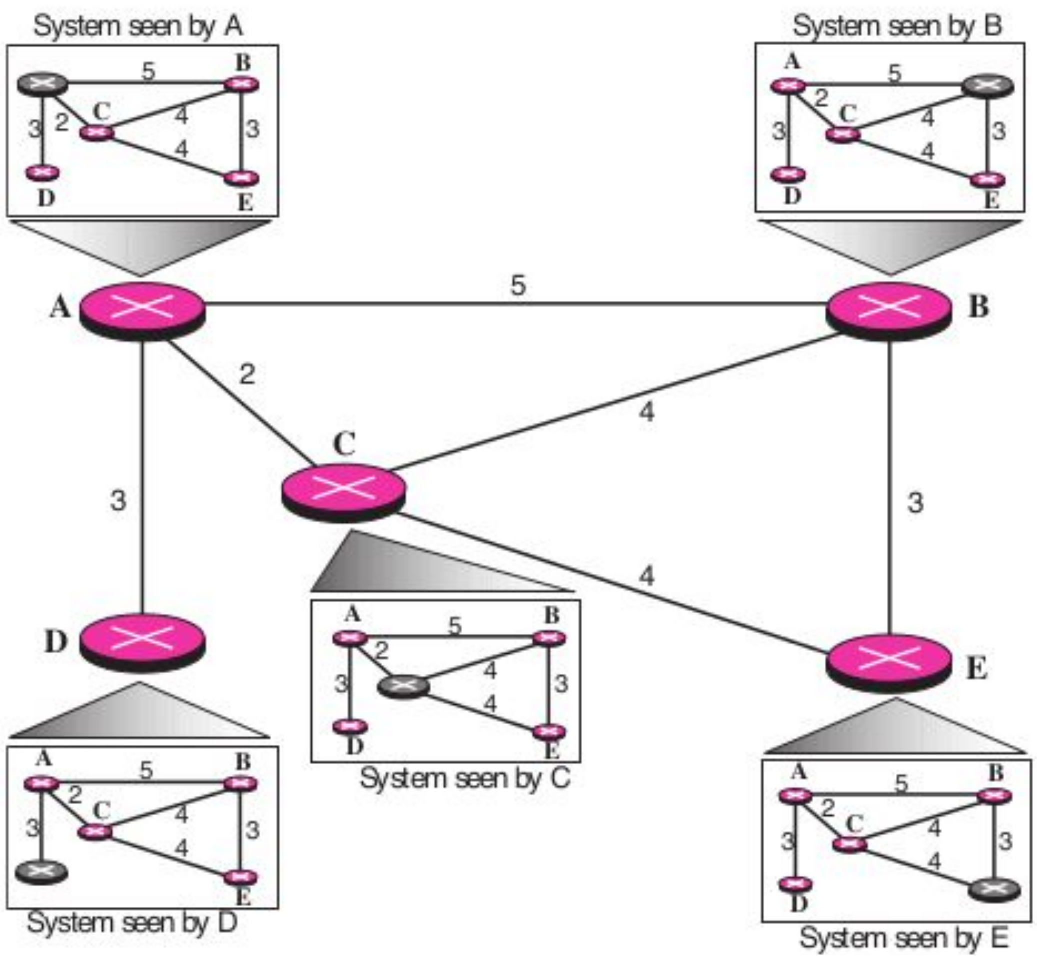
After B sends the route to A



Link State Routing

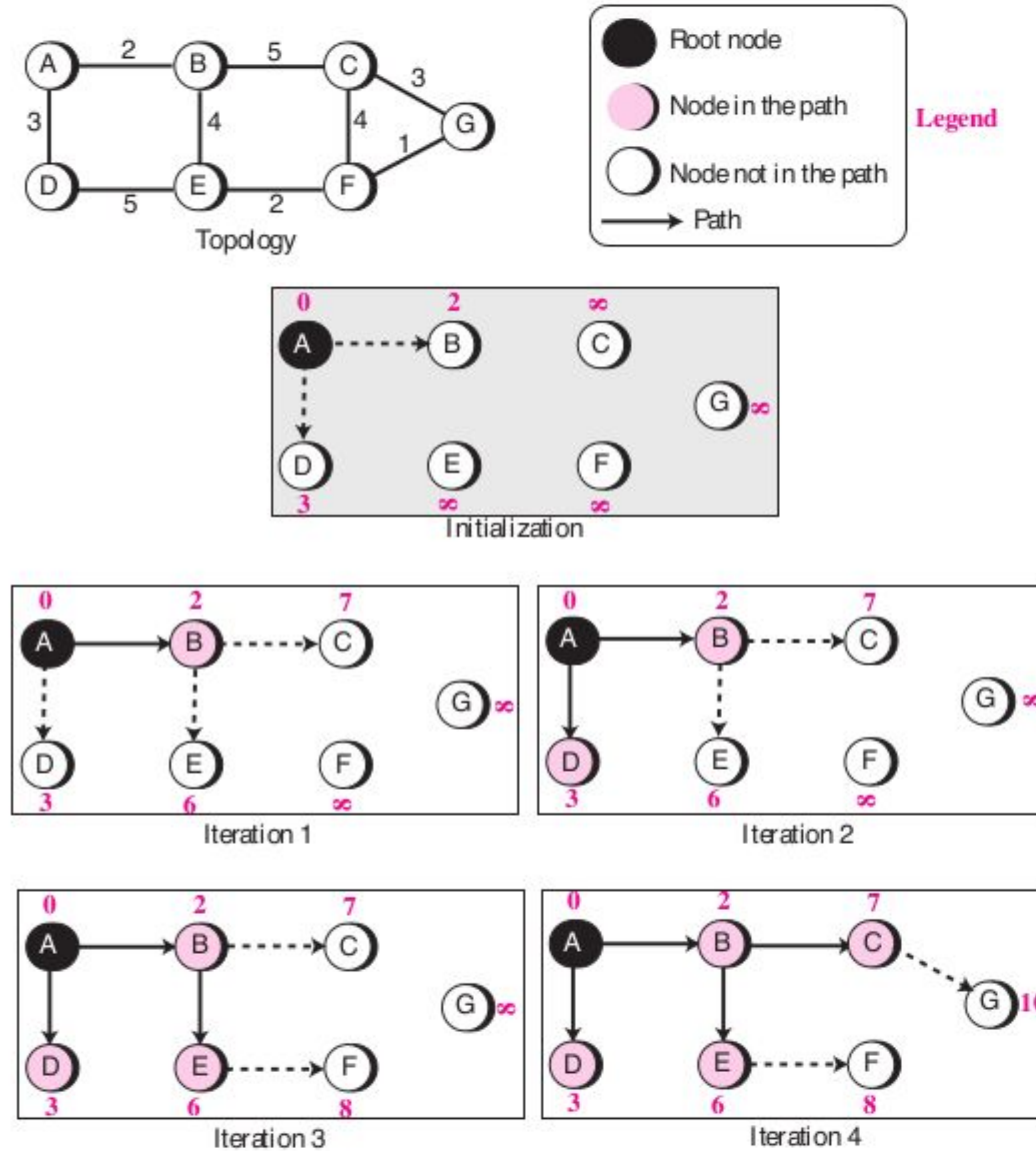
Fundamentals

- Each router maintains a *link state database* with information about the exact network topology
- To create this database:
  - Each router identifies its neighboring routers and their distance (link state)
  - Each router announces this information to all routers in the network (flooding)
- Using the whole information about the network (graph), each node creates a route map (tree) from its point of view using the Dijsktra's algorithm
- Example: OSPF (Open Shortest Path First)



Link State Routing

Example: Routes for node A



Internet Routing

- Internet is organized in **Autonomous Systems (AS)**
  - An AS is a collection of networks and routers managed and administered by the same authority
  - Each AS is identified by an AS Number (ASN)
  - There are more than 54,000 ASes
- Internal routers of the AS
  - They interconnect networks within their AS
  - They know internal routes of their AS, but don't know the route to other ASes
  - They use routing protocols named IGP (Interior Gateway Protocol)
- External (or border) routers of the AS
  - They interconnect to other ASes
  - They know routes to other ASes, but don't know their internal routes in detail
  - They use routing protocols named EGP (Exterior Gateway Protocol)

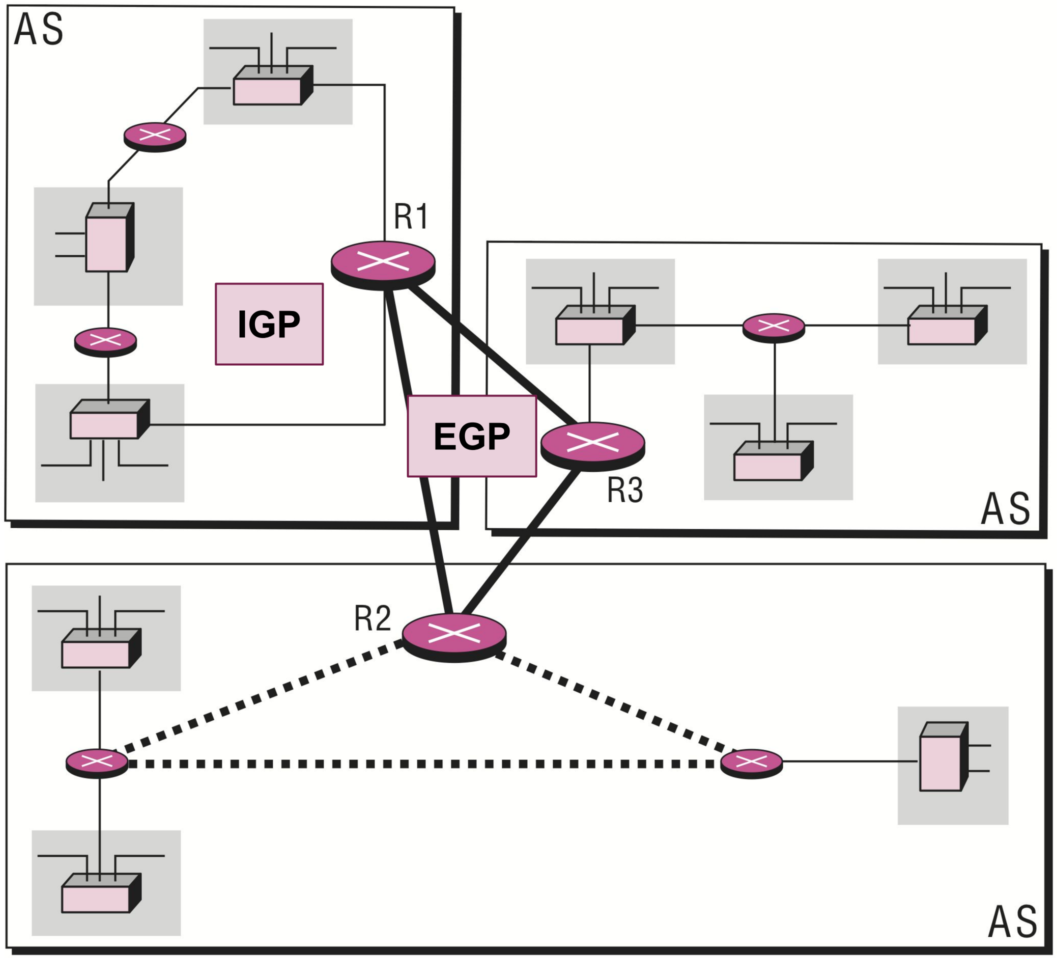
Internet Routing

Interior Gateway Protocols (IGP): Used by internal routers for the routing within an AS

- RIP: Routing Information Protocol
- OSPF: Open Shortest Path First
- IGRP: Interior Gateway Routing Protocol (from Cisco)

Exterior Gateway Protocols (EGP): Used by border routers for the routing between different ASes

- EGP: Exterior Gateway Protocol (obsolete)
- BGP: Border Gateway Protocol



Path Vector Routing

Interior Gateway Protocols (IGP)

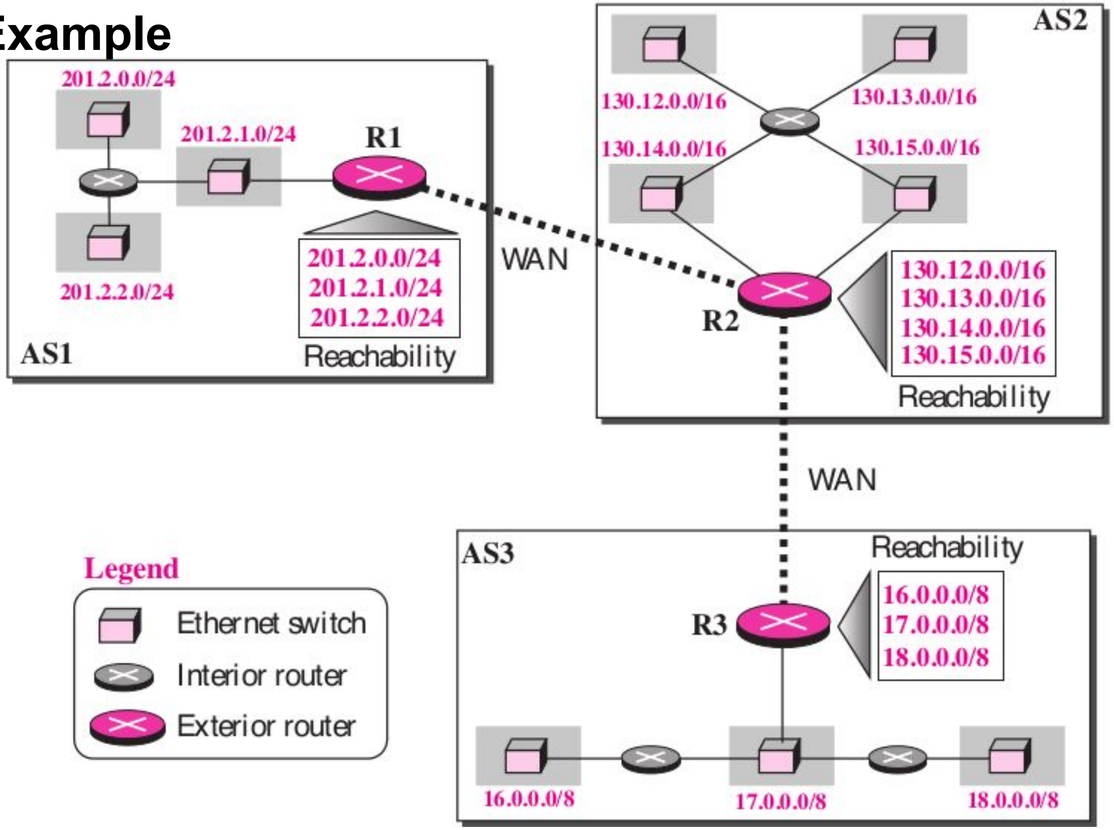
- Distance vector protocols show instabilities with a few hops between networks and convergence problems
- Link state protocols converge fast, but require the exchange of a great volume of information
- None of these algorithms can be applied to inter-AS routing

Path Vector Routing

- Based on distance vector routing, it tries to solve convergence problems for inter-AS routing (EGP)
- Starting from the destinations reachable in the AS (*reachability*), by means of an exchange process, each router obtains:
  - The list of reachable destinations (networks)
  - The full *path* to each destination, as a list of ASes that must be traversed
- CIDR is used to aggregate network addresses in routing tables
- Easy loop prevention, by dropping paths of which the own AS is part
- *Policy routing* is implemented by checking if a given AS is part of the path

Path Vector Routing

Example



After all tables are exchanged

R1

Network	Path
201.2.0.0/22	AS1 (This AS)
130.12.0.0/18	AS1, AS2
16.0.0.0/6	AS1, AS2, AS3

Path-Vector Routing Table

R2

Network	Path
201.2.0.0/22	AS2, AS1
130.12.0.0/18	AS2 (This AS)
16.0.0.0/6	AS2, AS3

Path-Vector Routing Table

R3

Network	Path
201.2.0.0/22	AS3, AS2, AS1
130.12.0.0/18	AS3, AS2
16.0.0.0/6	AS3 (This AS)

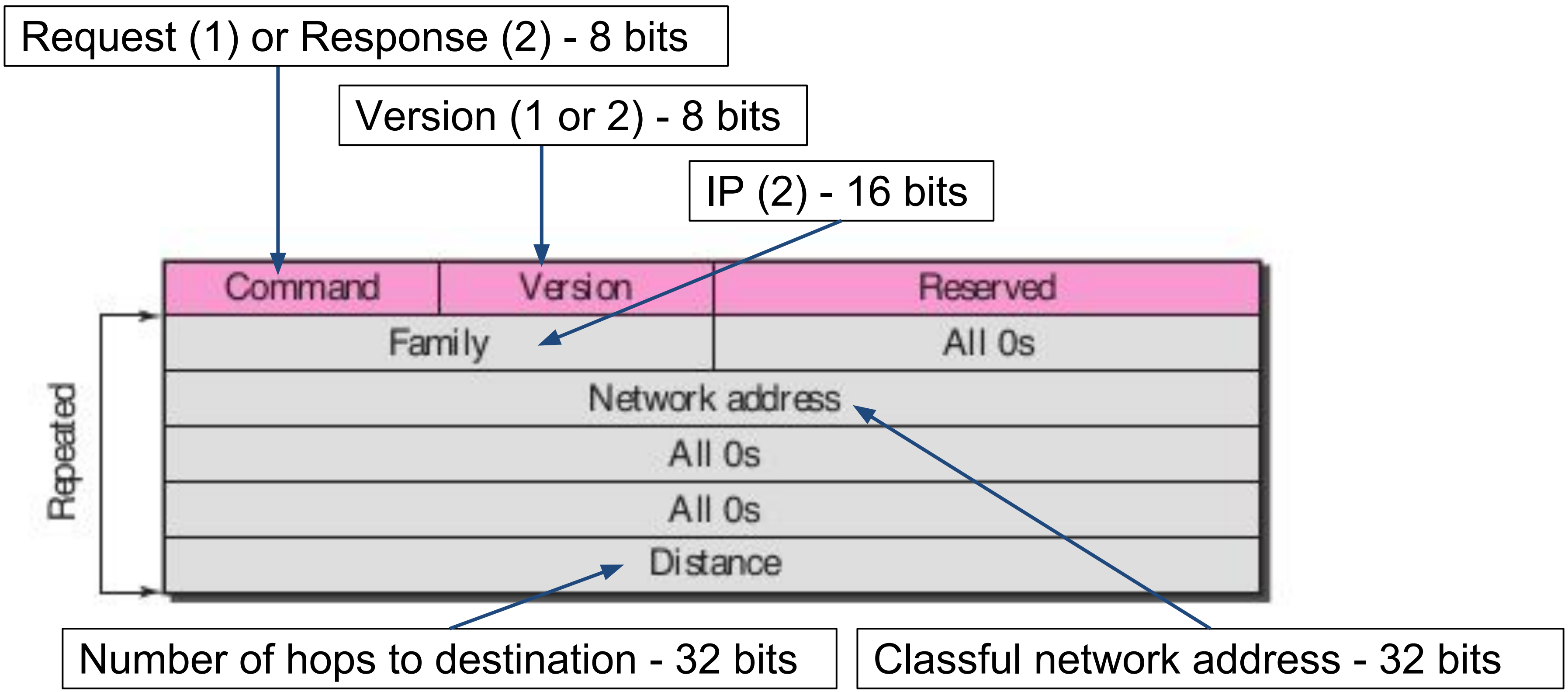
Path-Vector Routing Table

Routing Information Protocol (RIP)

Routing Information Protocol (RIP)

- Interior routing protocol (IGP) based on distance vector (Bellman-Ford's algorithm)
- Versions and RFCs:
  - RIP version 1 → RFC 1058 (1993)
  - RIP version 2 → RFC 2453 (1998)
  - RIPng (for IPv6) → RFC 2080 (1997)
- Distance vector includes:
  - The list of destination (networks) that are reachable from each router
  - The distance to each destination
- Messages are encapsulated in UDP datagrams addressed to port 520
- The distance metric used is the number of hops
- Infinity is set to 16 hops
- RIP can use the following techniques:
  - Split horizon
  - Split horizon with poisoned reverse
  - Triggered updates

RIP-1: Message Format



Request messages

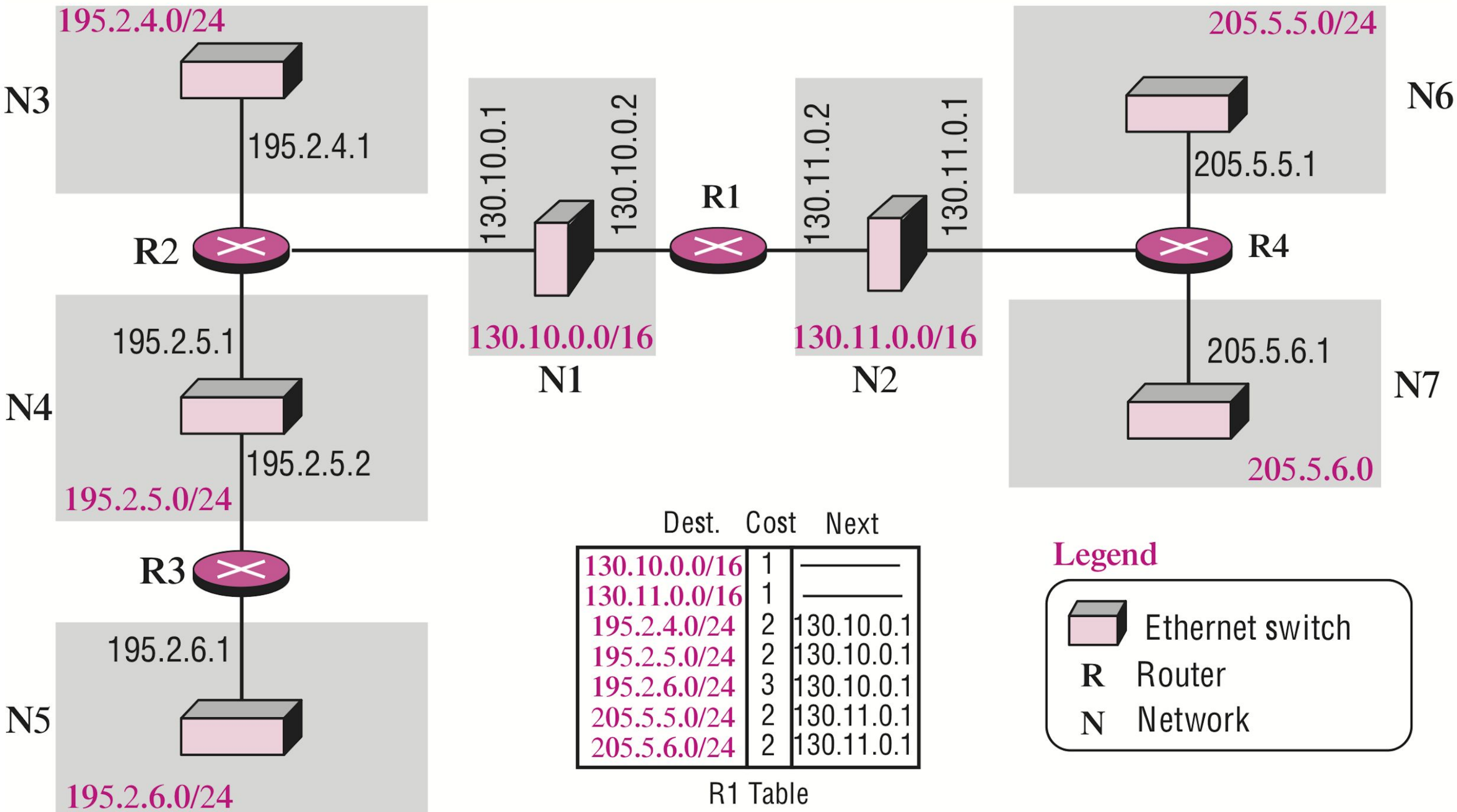
- Sent by a router when it connects to the network → Network Address = `0.0.0.0`
- Sent when a table entry expires → Network Address = Destination

Response messages

- Periodically broadcasted, with distance vector
- Sent in response to a request
- Triggered update, when the distance to the destination changes

RIP-1: Example

Example: Which RIP (Response) message will R1 send to R2?



Dest.	Cost	Next
130.10.0.0/16	1	———
130.11.0.0/16	1	———
195.2.4.0/24	2	130.10.0.1
195.2.5.0/24	2	130.10.0.1
195.2.6.0/24	3	130.10.0.1
205.5.5.0/24	2	130.11.0.1
205.5.6.0/24	2	130.11.0.1

R1 Table

Dest.	Cost	Next
130.10.0.0/16	1	———
130.11.0.0/16	2	130.10.0.2
195.2.4.0/24	1	———
195.2.5.0/24	1	———
195.2.6.0/24	2	195.2.5.2
205.5.5.0/24	3	130.10.0.2
205.5.6.0/24	3	130.10.0.2

R2 Table

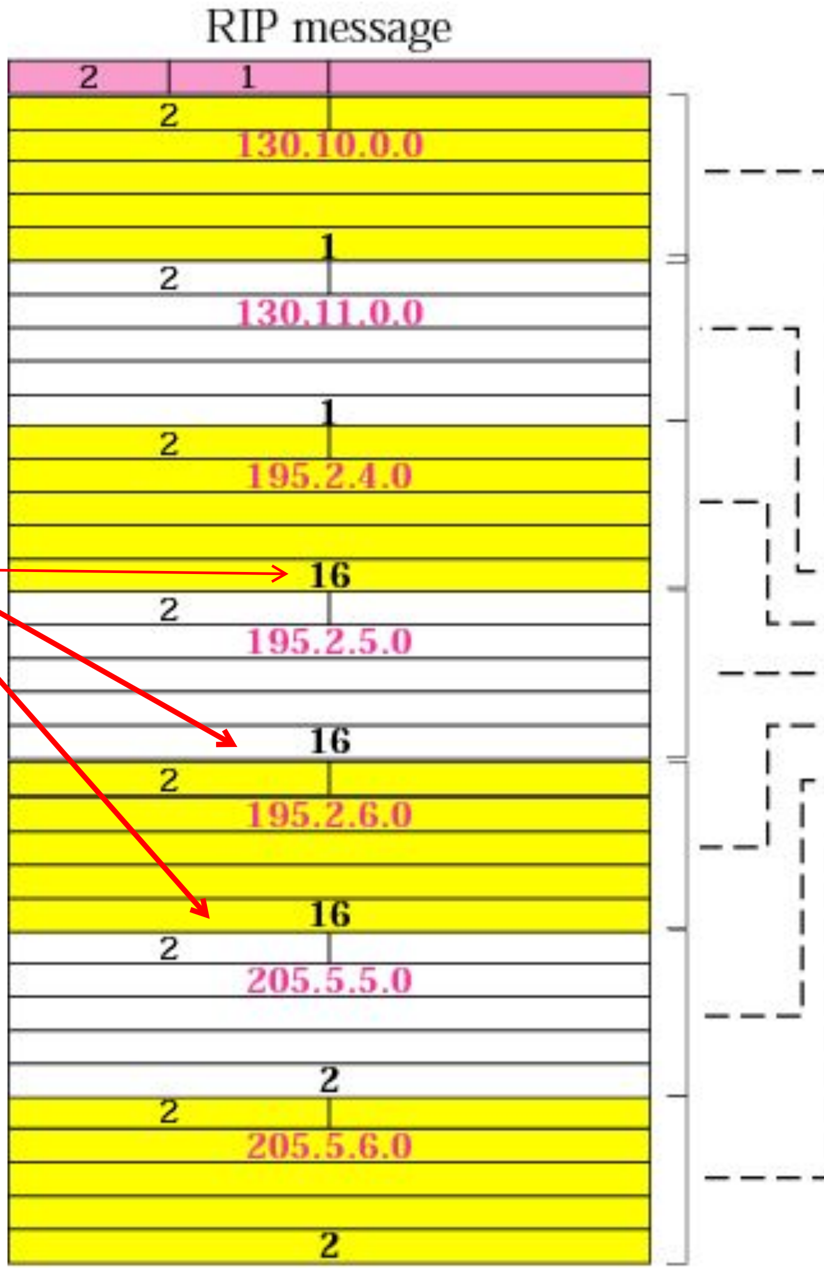
Dest.	Cost	Next
130.10.0.0/16	2	195.2.5.1
130.11.0.0/16	3	195.2.5.1
195.2.4.0/24	2	195.2.5.1
195.2.5.0/24	1	———
195.2.6.0/24	1	———
205.5.5.0/24	4	195.2.5.1
205.5.6.0/24	4	195.2.5.1

R3 Table

Dest.	Cost	Next
130.10.0.0/16	2	130.11.0.2
130.11.0.0/16	1	———
195.2.4.0/24	3	130.11.0.2
195.2.5.0/24	3	130.11.0.2
195.2.6.0/24	4	130.11.0.2
205.5.5.0/24	1	———
205.5.6.0/24	1	———

R4 Table

RIP-1: Example



in the table is at distance 2, but here is at 16 (infinity) which means that they are unreachable destination because they were learned through router 2, so R1 announces those destinations with infinity distance --> they are poisoned

Dest.	Hop
130.10.0.0	1
130.11.0.0	1
195.2.4.0	16
195.2.5.0	16
195.2.6.0	16
205.5.5.0	2
205.5.6.0	2

Table extracted from message before incrementing

RIP-1: Timers

Update timer (30 s)

- Interval between two unsolicited Response messages to announce the distance vector
- In practice, a random value between 25 and 35 seconds is used

Invalid/Expiration timer (180 s)

- Time that a route can be in the routing table without being updated
- After this time, the hop count of the route will be set to 16, invalidating the route and making the destination unreachable

Flush timer (120 s)

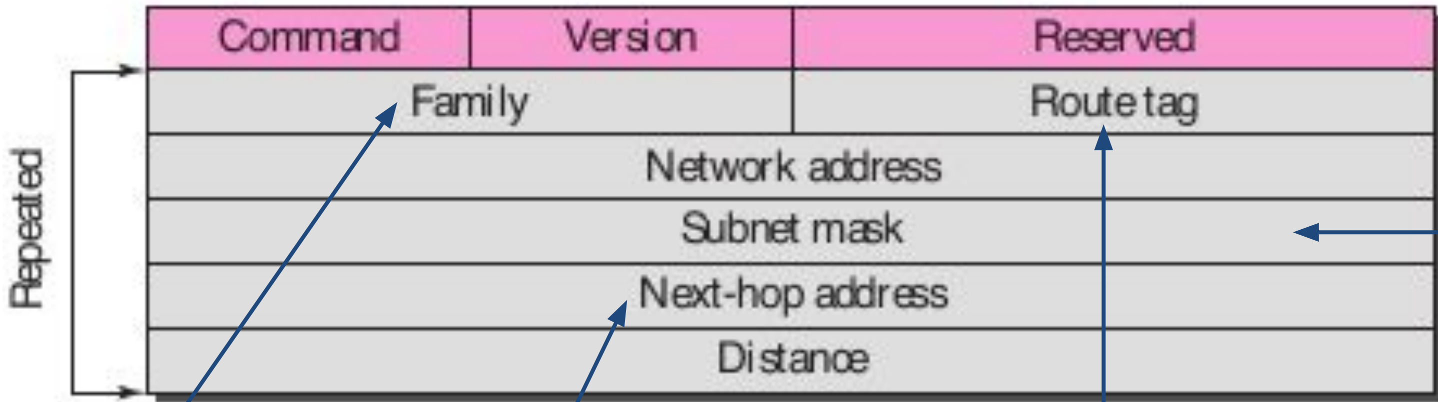
- Time that an invalidated route keeps being advertised to the neighbors
- After this time, the invalidated route is removed from the routing table

RIP-1: Limitations

- It can generate a **high amount of broadcast traffic**, due to periodic broadcast of distance vectors (Response messages)
- It **does not allow alternative distance metrics** to hop count
- Once the tables are calculated, **alternative paths are not allowed** to balance the network load
- When the network grows, it takes a **long time to propagate changes** to all the nodes in the network
- **Infinity** is set to 16 hops, and big networks may need more hops
- It provides **no support for CIDR**
- Route information is **not authenticated**

RIP-2

- RIP version 2 is a routing protocol similar to RIP-1 that overcomes some limitations:
  - Support for network masks (classless addressing)
  - Support for multicast (`224.0.0.1`)
  - Support for authentication



`0xFFFF`: Authentication (first entry)
`0x0002`: IP routing (route entry)

Additional information about the route (in route entries)
- AS-number: to separate internal and external routes

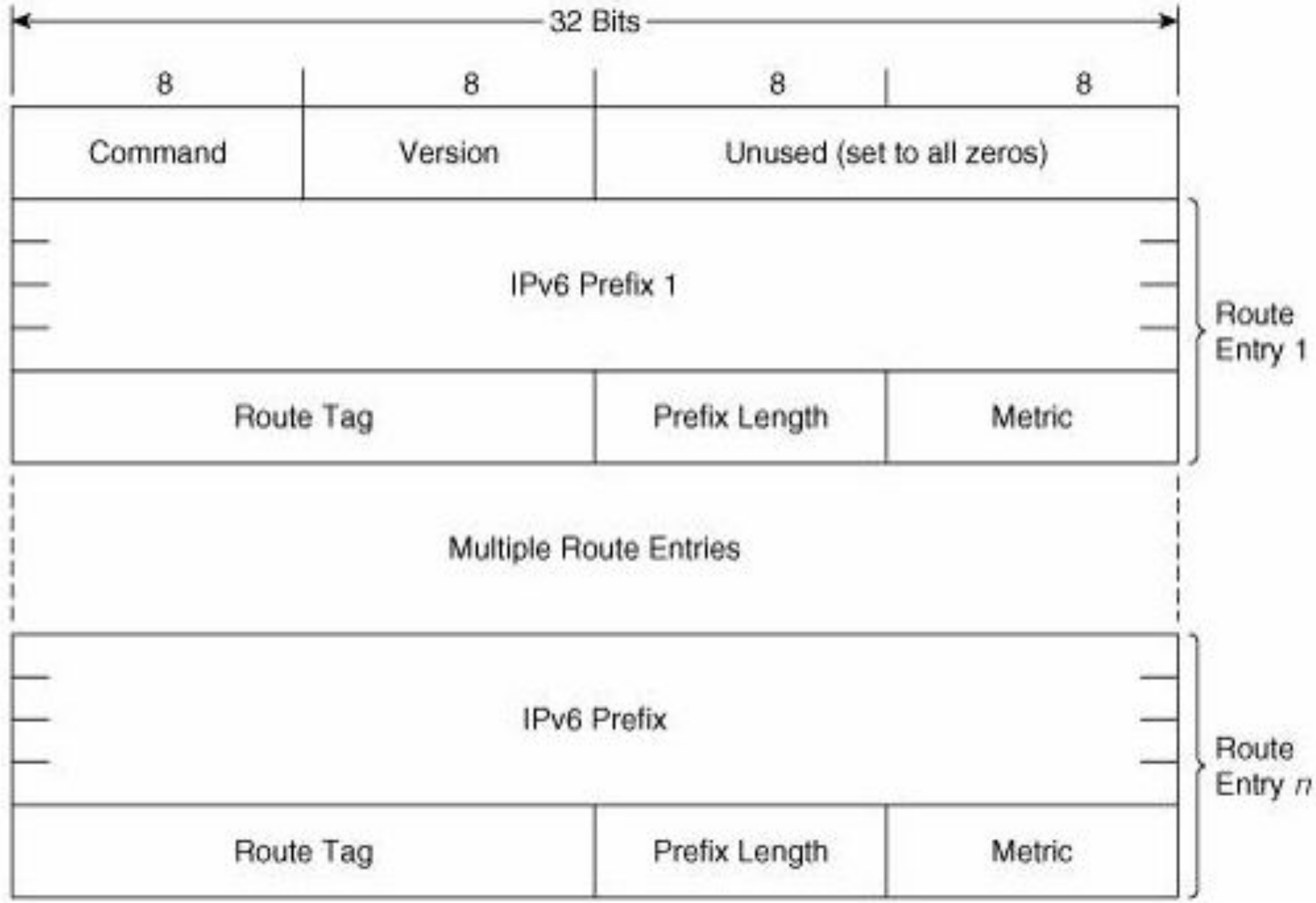
Authentication algorithm (in authentication entry):
- `0` (none), `2` (plain password), `3` (keyed message digest)

It suggests the address of the next hop to avoid non optimal paths (e.g. non-RIP router). Usually, it is `0.0.0.0` (next hop is the address of the sender of the message)

RIPng: RIP for IPv6

- RIPng (RIP new generation) is the adaptation of RIP-2 for IPv6
- Differences with RIP-2:
  - RIPng messages are encapsulated in UDP datagrams addressed to port 521 and sent to the IPv6 multicast address `FF02::9`
  - Distance vector in Response messages announces IPv6 network prefixes, instead of IPv4 network address
  - Path information contained in a distance vector does not include the Next Hop field (that would nearly double the size of each entry)
    - A specific Next Hop entry (with `0xFF` in the Metric field) can be included, affecting the following entries until a new Next Hop entry appears
  - It doesn’t use authentication information as in RIP-2
    - The encryption and authentication mechanisms available in IPv6 are used

RIPng: Message Format



Route Entry:

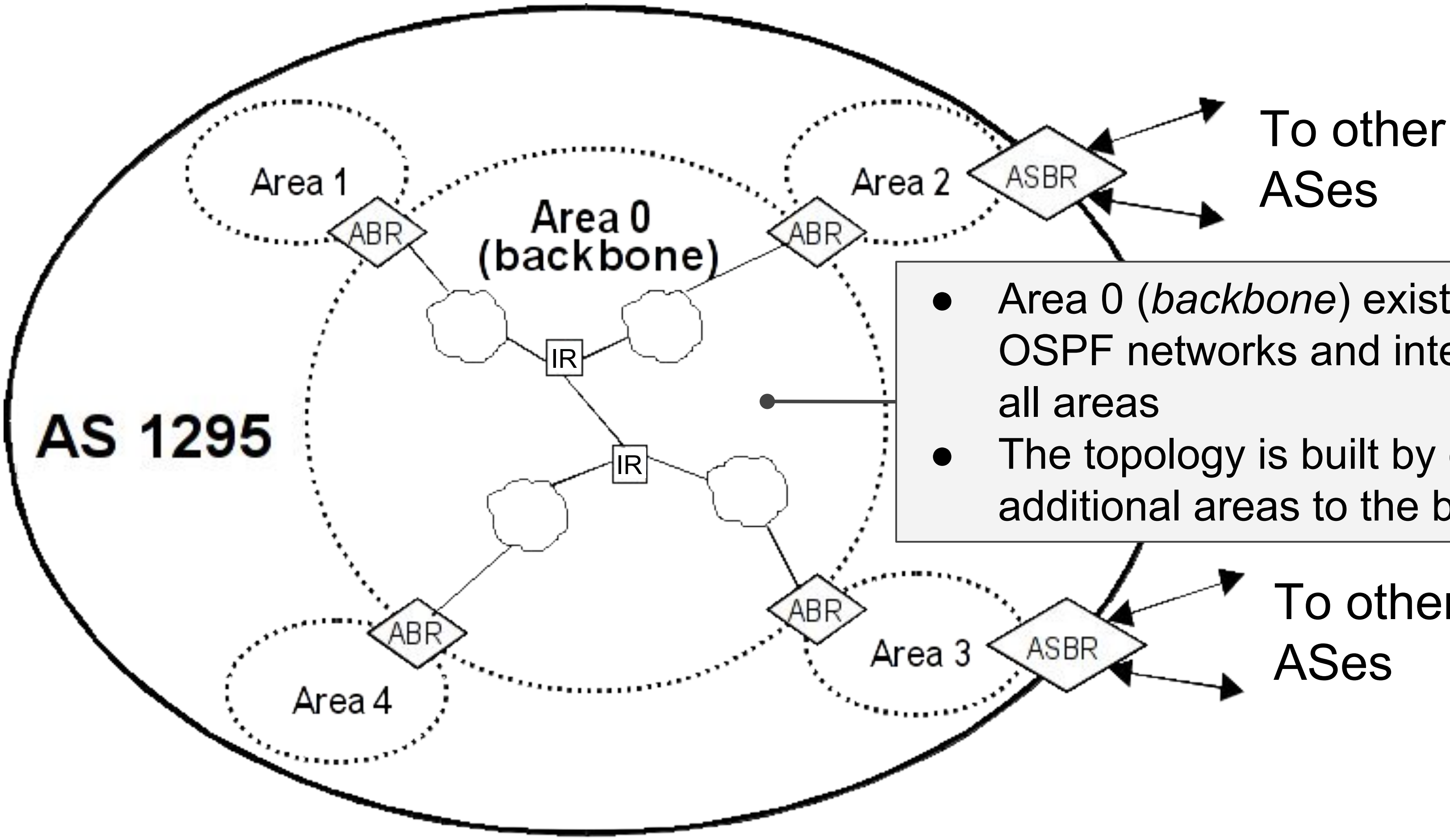
- IPv6 prefix (128 bits): IPv6 network prefix of the destination network announced
- Prefix length (8 bits): Length of the network prefix announced
- Route Tag and Metric: as in RIP-2

Open Shortest Path First (OSPF)

Open Shortest Path First (OSPF)

- Interior routing protocol (IGP) based on link state
- Developed as an alternative to RIP to alleviate its limitations:
  - Load balancing between equivalent routes
  - Logical network partitioning to reduce the amount of exchanged information
  - Faster convergence, by immediately propagating changes in routes
  - Support for VLSM (Variable-Length Subnet Mask) and CIDR
  - Support for authentication
- Uses its own encapsulation protocol (89) and multicast addresses:
  - `224.0.0.5` or `FF02::5` - OSPF routers in a network
  - `224.0.0.6` or `FF02::6` - OSPF designated routers in a network
- Versions and RFCs:
  - OSPF version 2 → RFC 2328 (1998)
  - OSPF version 3 (for IPv6) → RFC 5340 (2008)

OSPF: Areas



- Area 0 (*backbone*) exists in all OSPF networks and interconnects all areas
- The topology is built by connecting additional areas to the backbone

Area: Logical group of routers and networks, with a 32-bit Area ID

- Routers maintain information about its area only
- Areas limit the amount of information about link state to be exchanged

OSPF: Routers and Networks

Routers

- Each router has a unique 32-bit Router ID (RID) in the OSPF network
- Information stored and exchanged depends on router type:
  - **Internal Router (IR)**
    - Located only in one area (all the interfaces are in the same area)
    - Maintain a database with information about their area only
    - Routers in the backbone area are called backbone routers
  - **Area Border Router (ABR)**
    - Connected to two or more areas (one of them must be the backbone)
    - Maintain a database for each area
    - Condense the topological information of their areas for distribution to the backbone, which in turn distributes the information to the other areas
  - **AS Boundary Router (ASBR)**
    - Connected to other ASes, advertise external routing information to OSPF
    - Typically also run an exterior routing protocol (BGP) or use static routes

Networks

- They define the frequency and type of communications between routers
- Types: point-to-point, broadcast multi-access, non broadcast multi-access (NBMA), and point-to-multipoint

OSPF: Neighbor and Adjacency

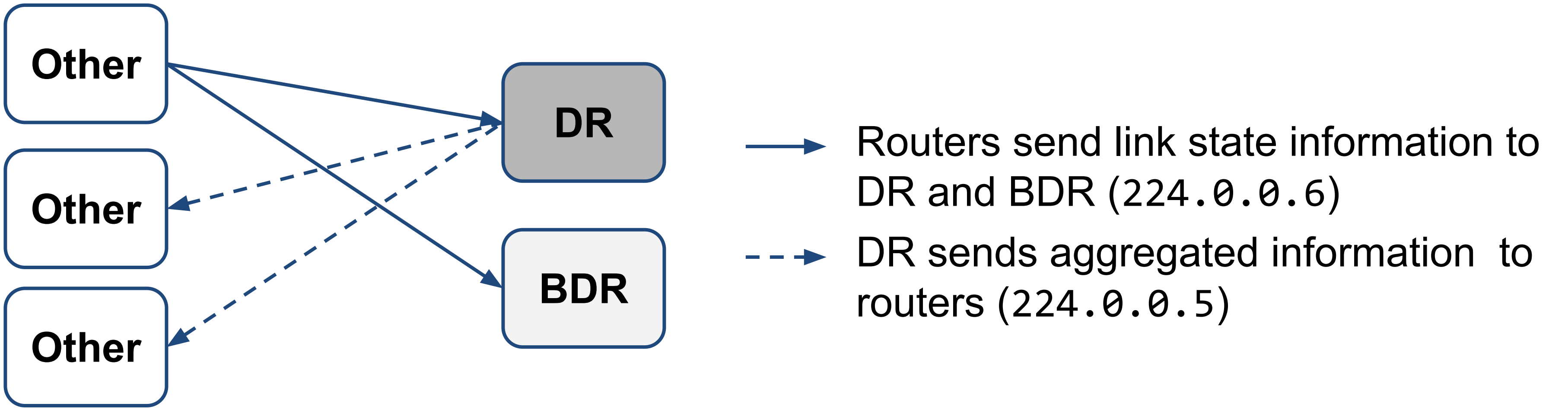
Neighbor Relationship

- Routers sharing a common link, belonging to the same area and using the same authentication mechanism

Adjacency Relationship

- OSPF creates adjacencies between neighboring routers for the purpose of exchanging routing information
    - Link-state databases must be synchronized between pairs of adjacent routers
- Not every two neighboring routers will become adjacent, thus limiting the information exchanged between routers
- Developed according to network type:
    - Multi-access: the Designated Router (DR) and the Backup Designated Router (BDR) become adjacent to all other routers on the network
    - Point-to-point: between the two neighboring routers

OSPF: Neighbor and Adjacency



- The process of information distribution in multi-access networks is an optimization of the flooding strategy
- In case of failure of the DR, the backup (BDR) will take its role
- The DR does not immediately send Link State Update messages (i.e. after an update), to accommodate multiple updates in a single message
- Link State Update messages are acknowledged by Link State Ack messages, providing a reliable update mechanism

OSPF: Operation

Neighbor discovery: OSPF Hello protocol

- Responsible for establishing and maintaining neighbor relationships
- On broadcast and point-to-point networks, each router advertises itself (RID) by periodically multicasting Hello messages, allowing neighbors to be discovered dynamically
- Bidirectional communication between neighbors is ensured
  - Hello messages contain the list of routers (RIDs) whose Hello messages have been seen recently
  - The router must see itself listed in the neighbor's Hello message
- On multi-access networks, the Hello Protocol elects a Designated Router for the network

Designated Router election (DR and BDR)

- DR and BDR are elected using priority information from Hello messages
- Routers with higher priority in a network become DR and BDR, respectively
  - Routers with priority 0 are ineligible to become DR on the attached network
- In case of a tie, the one having the highest RID is chosen

OSPF: Operation

Link-state Database Synchronization

- Adjacent routers must synchronize their link-state databases to establish full adjacency
- **Link-state database exchange**
  - The exchange follows a master-slave approach, where the master initiates the exchange of Database Description messages, which summarize link-state database contents
  - Each message is acknowledged by another Database Description message from the slave
- **Link-state database loading**
  - During the previous exchange process, routers detect obsolete or missing information in their database
  - After that, a copy of those link states is requested to the adjacent router by means of Link State Request messages
  - The requested information is provided in Link State Update messages, which are acknowledged with Link State Ack messages (as for the update mechanism)

OSPF: Operation

Construction of routing tables

- When the router has the information about link state, it creates the shortest-path tree
- The shortest-path tree includes both routers (RID) and networks (IP), as well as associated costs
- According to the shortest-path tree, the routing table is built, containing the destination (network/host), next hop and distance/metric

Border Gateway Protocol (BGP)

Border Gateway Protocol (BGP)

- Inter-AS (or exterior) routing protocol based on path vector
- The primary function of a BGP speaking system is to exchange network reachability information with other BGP systems
  - This information includes the list of ASes that reachability information traverses
  - This information is sufficient for constructing a graph of AS connectivity for this reachability from which routing loops may be pruned
  - Each AS may enforce some policies to accept (import policy) and publish (export policy) the paths received
- Current version is BGP-4, which supports CIDR and route aggregation

BGP: Autonomous Systems (AS)

AS Types

- **Stub:** Only connected to one AS, it is source or destination of network traffic
  - This type of AS doesn't allow *transit* traffic (i.e. from one AS to another AS)
- **Multihomed:** Connected to several AS, but still source or destination of network traffic
  - This type of AS doesn't allow *transit* traffic
- **Transit:** Multihomed AS that allows *transit* traffic

Routing Policies

- Each AS can apply policies to limit the data traffic flow in the network
- Policies, which are not part of BGP, reflect contractual agreements and thus provider-dependent costs
- Depending on the path, each AS can, for example:
  - Be configured as a multihomed AS (not transit), announcing only its directly reachable networks
  - Don't act as transit for some AS
  - Select the path depending on the traffic
  - Avoid a specific AS as transit

BGP: Operation

- Communication between routers is done through TCP port 179
- Routers exchange the route table when they establish the initial connection
- Routers periodically send incremental updates of the initial table
- BGP messages include:
  - **OPEN**: Establishment of the BGP session
    - AS and router identifiers
    - Configuration parameters (hold timer and authentication)
  - **UPDATE**: Incremental update of routing information
    - Each message can include one reachable network in CIDR with its attributes, including the path, and a list of withdrawn networks
  - **NOTIFICATION**: Error or special condition
    - The session is closed and paths become invalid
    - Examples: hold timer expired, error in message, lack of attributes...
  - **KEEPALIVE**: To ensure that the BGP session is active
    - Sent in response to an OPEN message (to acknowledge it) and periodically, to inform about the presence of the router (no TCP keepalive)
    - After a hold time, if no information is received, the session is closed

BGP: Attributes

- UPDATE messages include reachable networks and path attributes
- Allow evaluating alternative paths to the same destination
- Path attributes are generated by each router, and routers can modify the attributes received
- Types:
  - **Well-known**
    - They must be accepted by all BGP implementations
    - They can be *mandatory* or *discretionary*
    - Mandatory attributes must be sent on every update
  - **Optional**
    - They are implementation specific
    - They can be *transitive* or *non-transitive*
    - Transitive attributes must be retransmitted even if not understood
- Examples of well-known and mandatory attributes:
  - **ORIGIN:** Origin of the route information (IGP, EGP or INCOMPLETE). It should not be modified by other BGP router
  - **AS_PATH:** The path as a set or sequence of ASes
  - **NEXT_HOP:** IP address of next hop to reach the destination