

IP Routing

Introduction (Static, Default, Dynamic),
RIP (Distance Vector), OSPF (Link State),
Introduction to Internet Routing (BGP, CIDR)

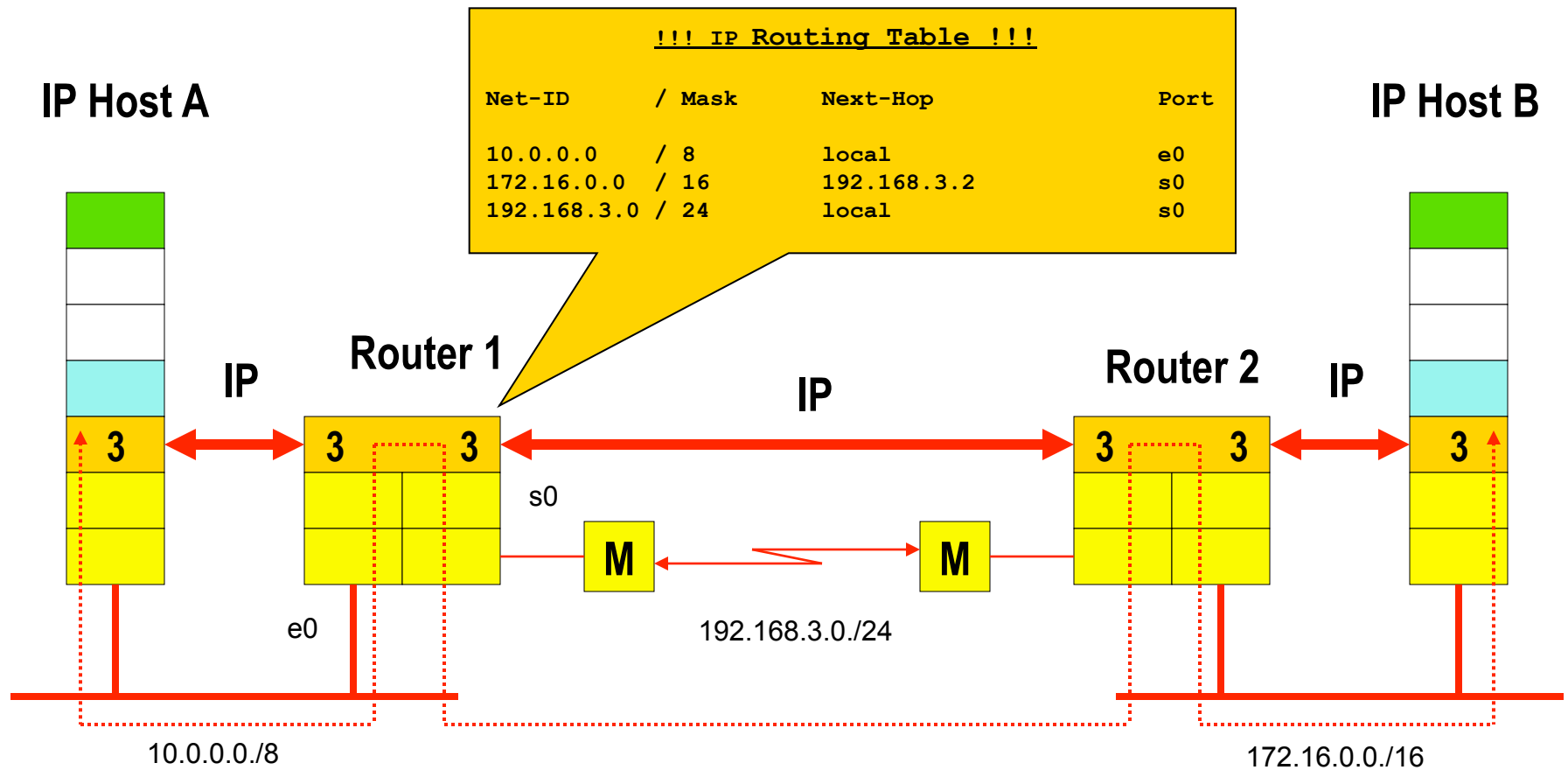
Agenda

- **Introduction to IP Routing**
 - Basics
 - Static Routing
 - Default Route
 - Dynamic Routing
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

IP, IP Routing Protocol, IP Routing Table

Layer 3 Protocol = IP

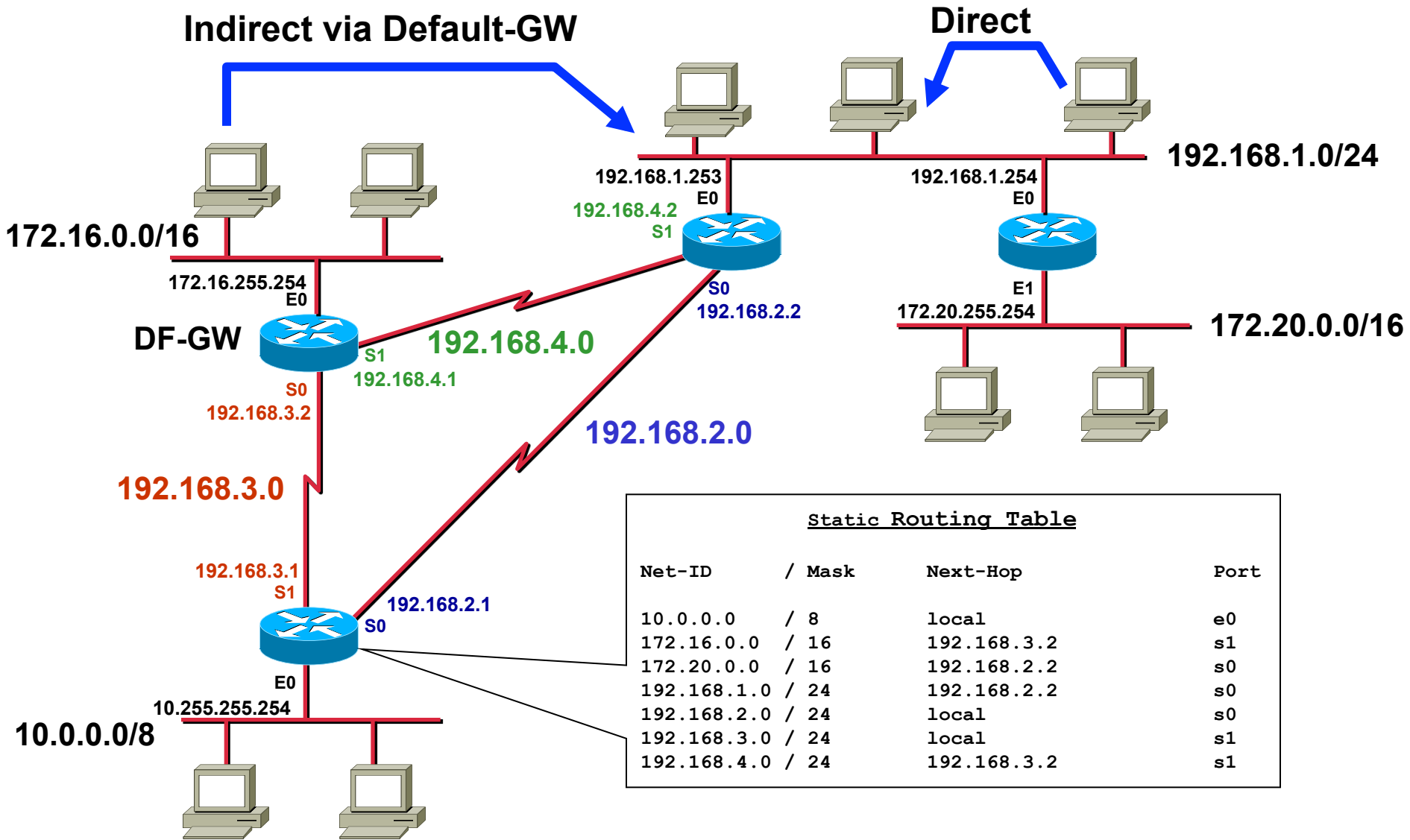
Layer 3 Routing Protocols = RIP, OSPF, EIGRP, BGP



What is Routing?

- *Finding / choosing a path to a destination address*
- **Direct delivery** performed by IP host
 - Destination network = local network
- **Indirect delivery** performed by router
 - Destination network \neq local network
 - Datagram is forwarded to **default gateway**
 - Passed on by the router based on routing table
- **Routing table**
 - Database of known destinations
 - Signposts leading to next hop

Direct versus Indirect Delivery Default Gateway / Routing Table



IP Routing Paradigm

- **Destination Based Routing**

- Source address is not taken into account for the forward decision

- **Hop by Hop Routing**

- IP datagrams follow the path (signpost) given by the current state of routing table entries

- **Least Cost Routing**

- Typically only the best path is considered for forwarding of IP datagrams
- Alternate paths will not be used in order to reach a given destination
 - Note: Some methods allow load balancing if paths are equal

Static versus Dynamic Routing

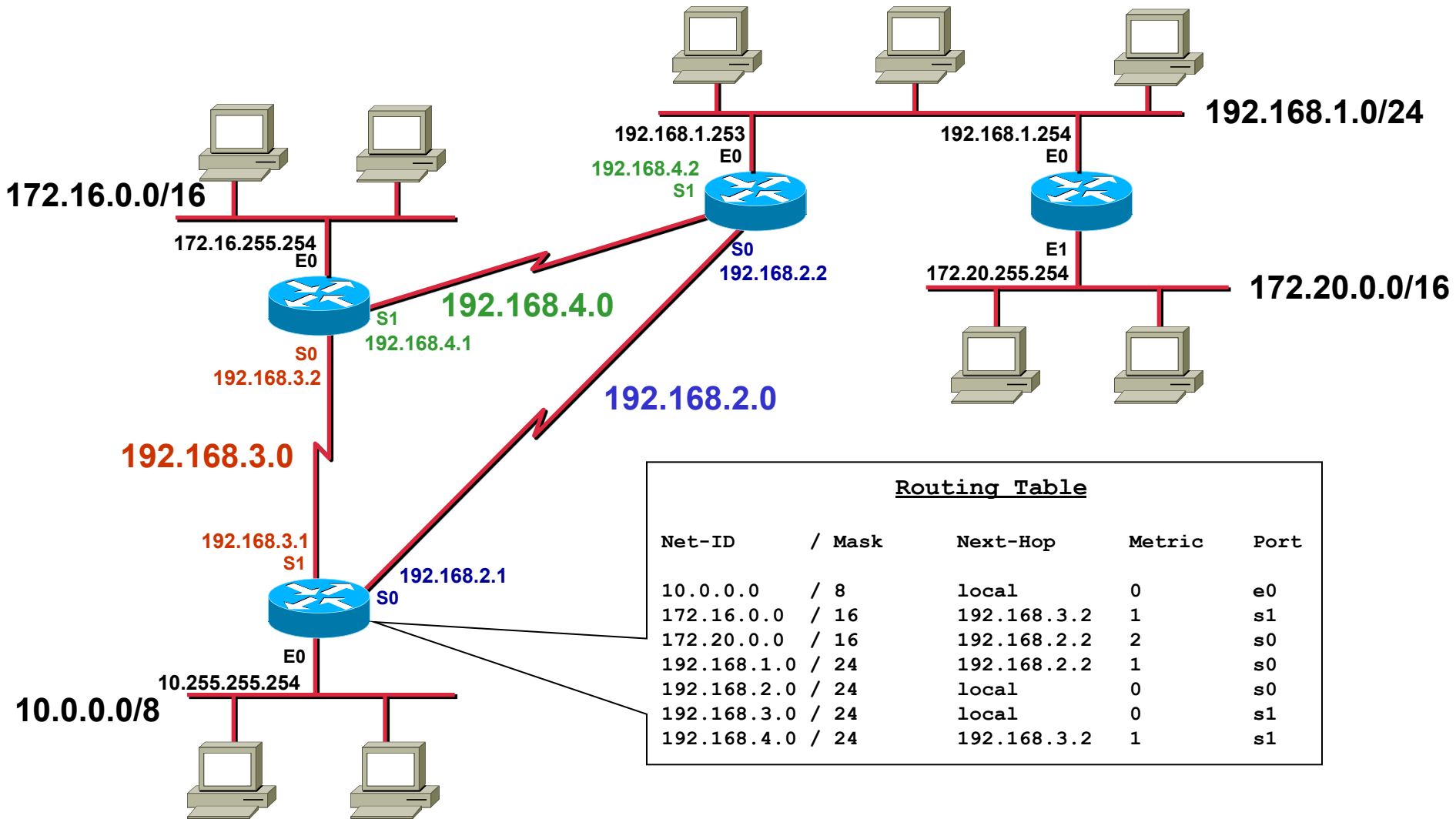
- **Static**

- Routing tables are preconfigured by network administrator
- Non-responsive to topology changes
- Can be labor intensive to set up and modify in complex networks
- No overhead concerning CPU time and traffic

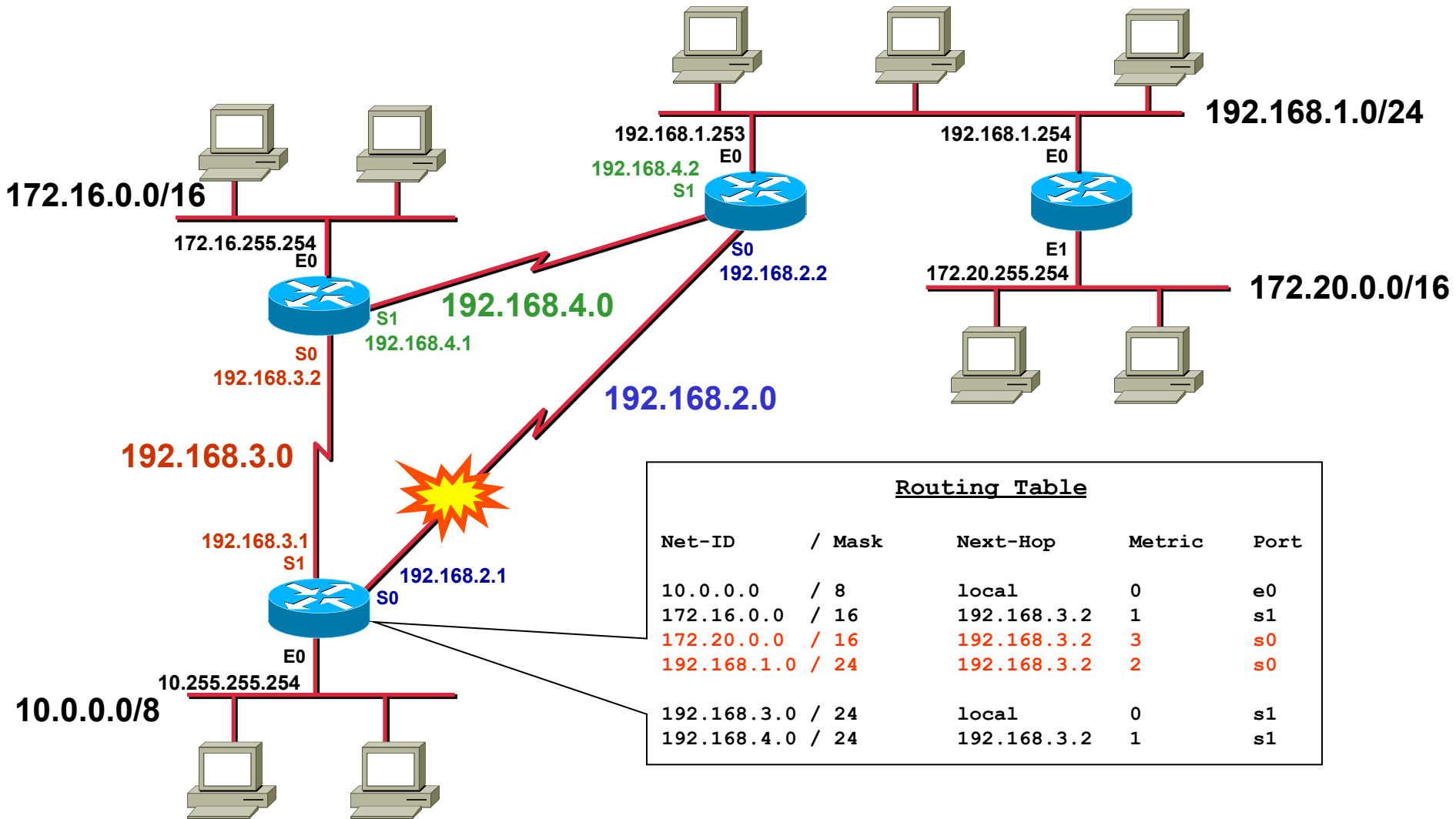
- **Dynamic**

- Routing tables are dynamically updated with information received from other routers
- Responsive to topology changes
- Low maintenance labor cost
- Communication between routers is done by routing protocols using routing messages for their communication
- Routing messages need a certain percentage of bandwidth
- Dynamic routing need a certain percentage of CPU time of the router
- That means overhead

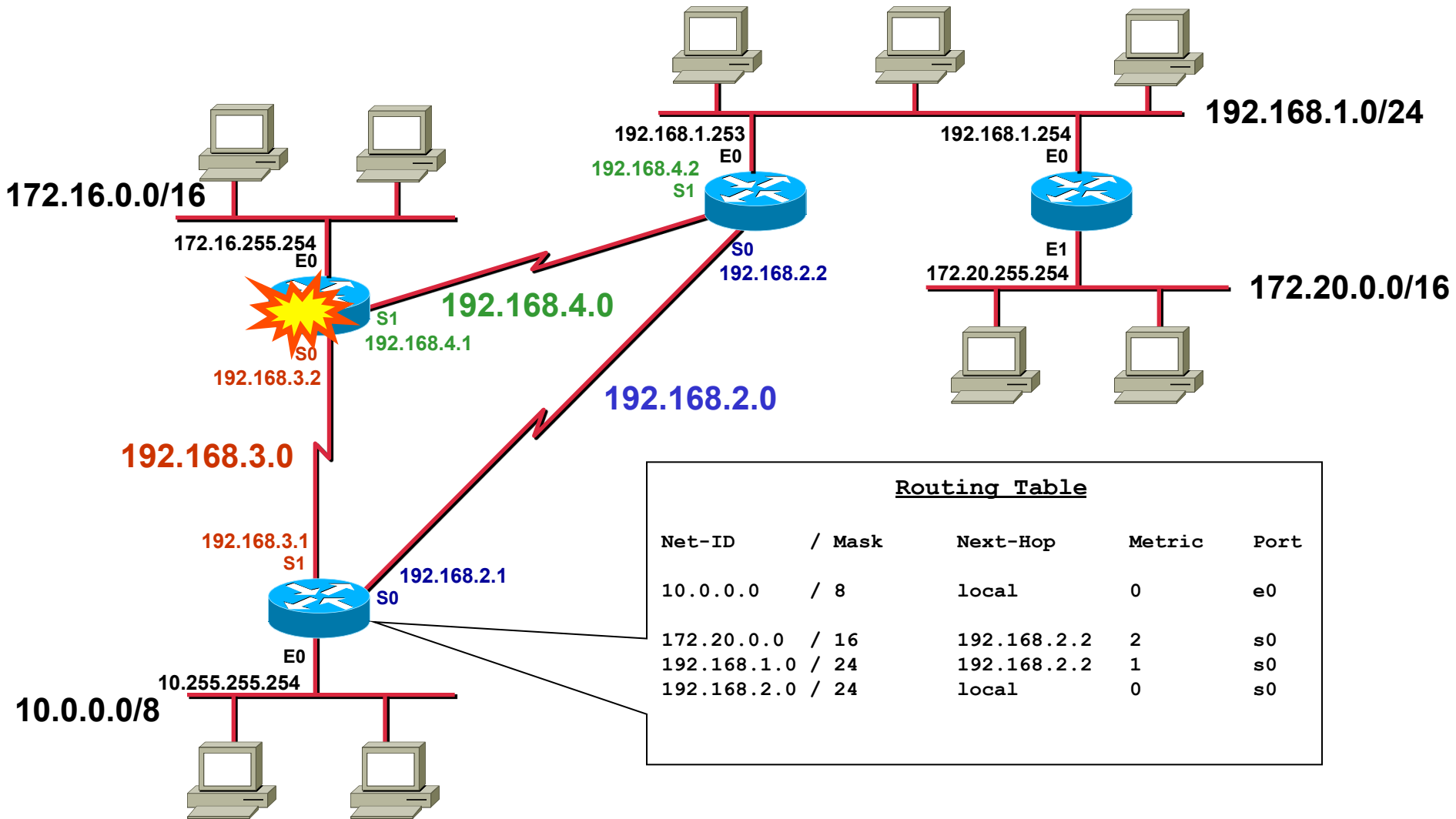
Routing Table - Dynamic Routing (1)



Routing Table - Dynamic Routing (2)



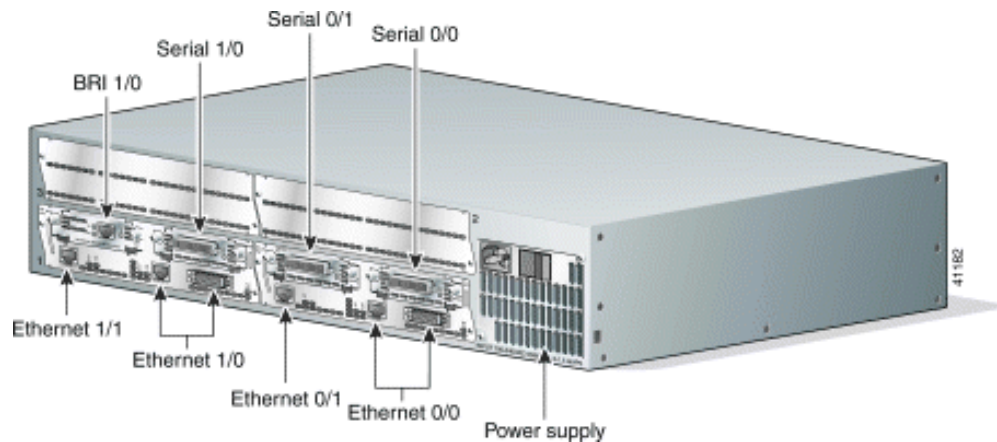
Routing Table - Dynamic Routing (3)



IP Router

- Initially Unix workstations with several network interface cards
- Today specialized hardware

Cisco 3600 Router



Agenda

- **Introduction to IP Routing**
 - Basics
 - Static Routing
 - Default Route
 - Dynamic Routing
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

Reasons for Static Routing

- **Very low bandwidth links**
- **Link is the only path to a stub network**
- **Dialup links and backup links**
 - X.25 SVC, ISDN, Frame Relay SVC, ATM SVC
- **Administrator needs full control over the link**
 - E.g. for security reasons
 - E.g. in hub and spoke topologies avoiding any-to-any communication
- **Router has very limited resources and cannot run a routing protocol**
- **Cisco syntax:**

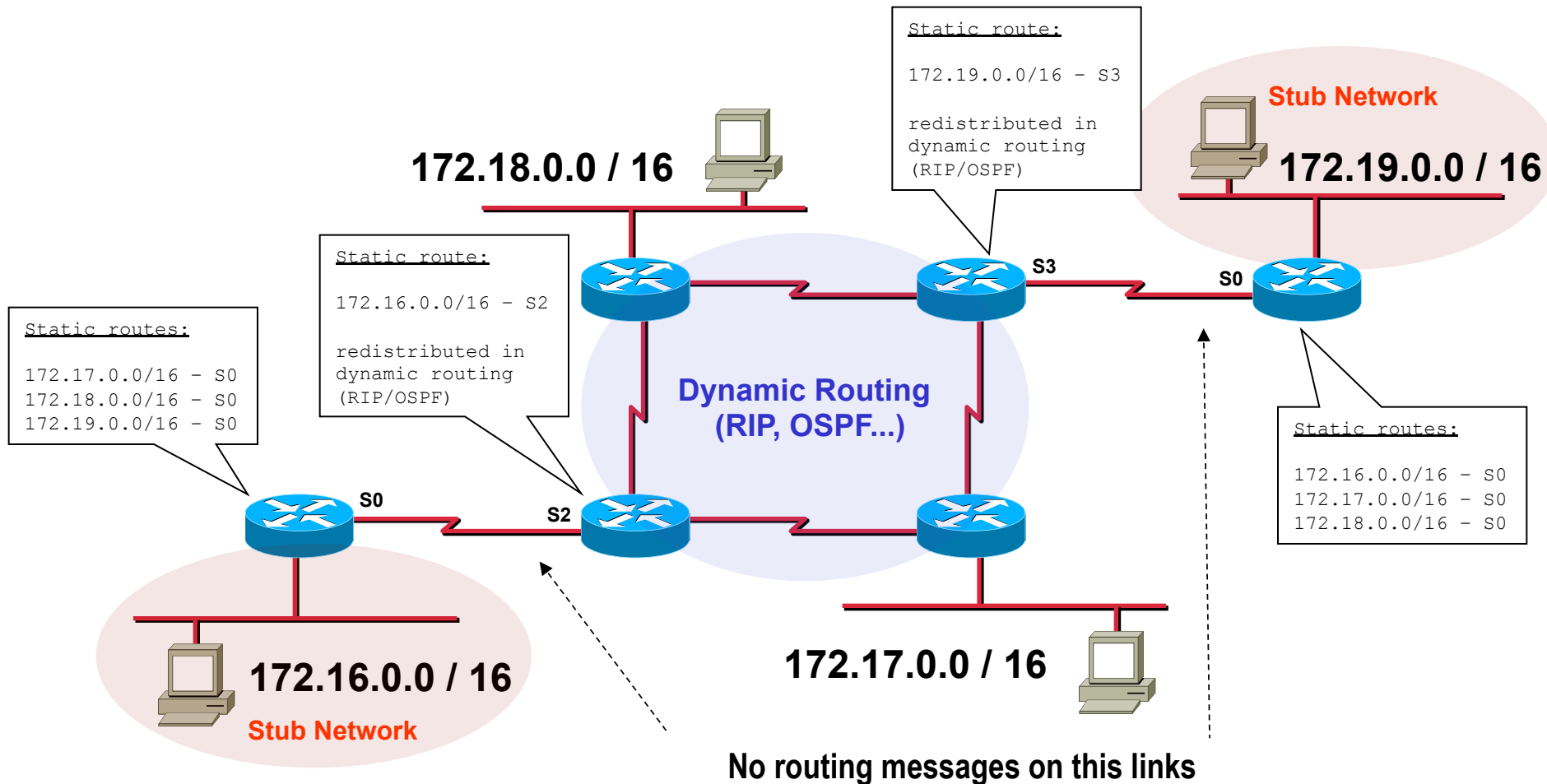
```
ip route prefix mask {ip-address | interface-type interface-number} [distance] [tag tag] [permanent]
```

Tag value that can be used as a “match” value for controlling redistribution via route maps

Specifies that the route will not be removed, even if the interface shuts down

Static Routing (1)

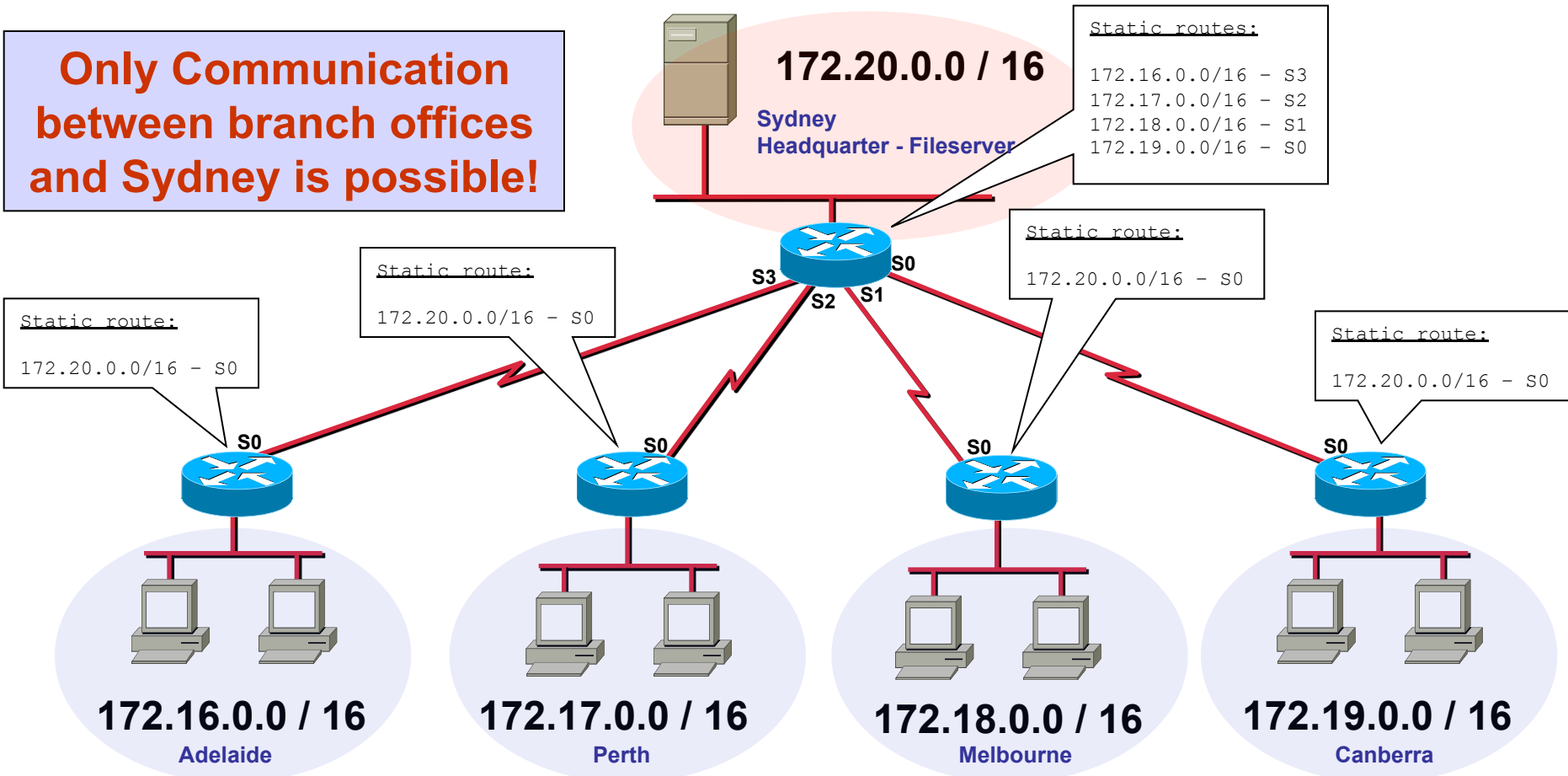
- Static routes to and from stub networks



Static Routing (2)

- Static routes in "Hub and Spoke" topologies

Only Communication between branch offices and Sydney is possible!



Agenda

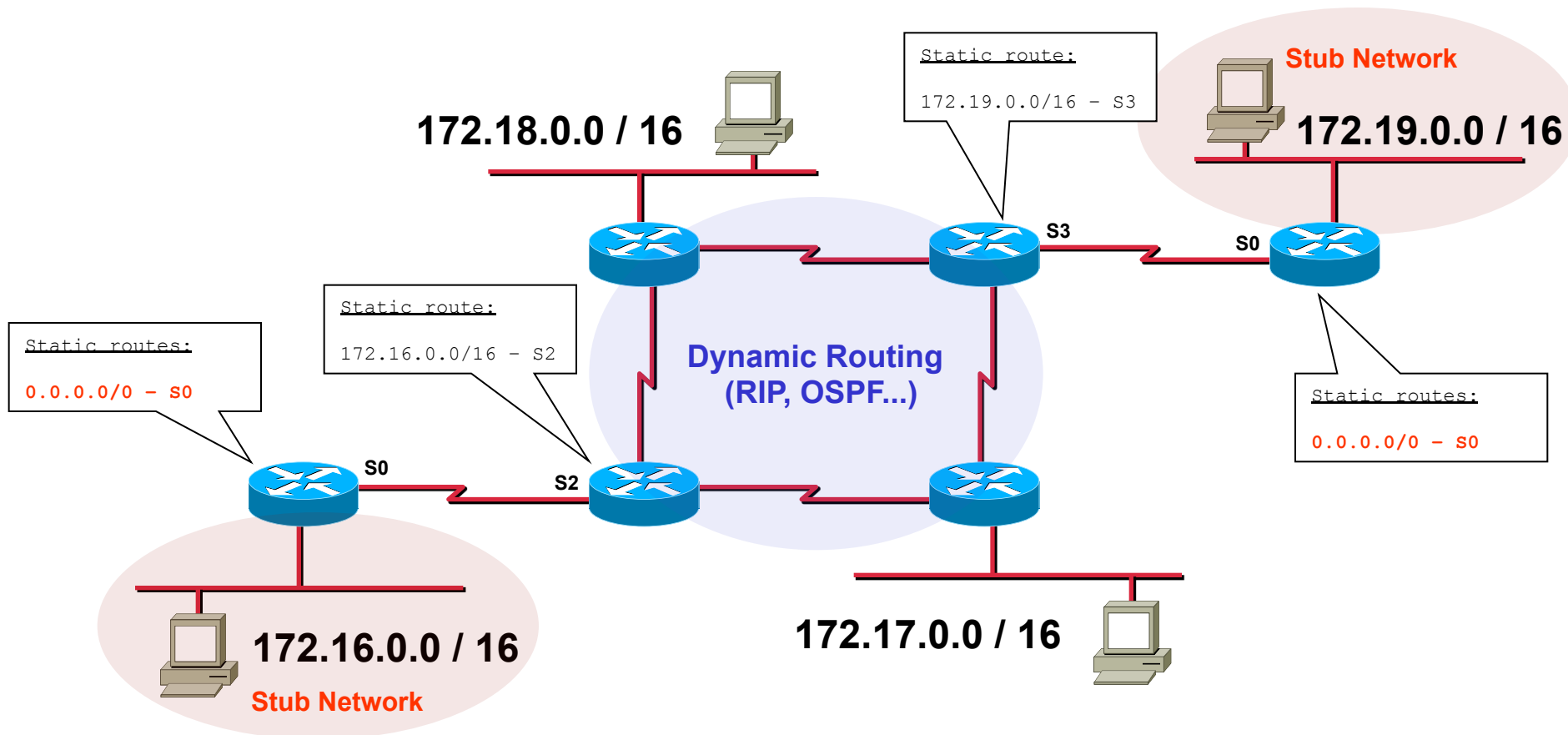
- **Introduction to IP Routing**
 - Basics
 - Static Routing
 - Default Route
 - Dynamic Routing
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

Default Route (DR)

- **Special static route in a router**
 - Traffic to unknown destinations are forwarded into the direction specified by the default route
 - Pointing to "**Gateway of Last Resort**"
- **In routing tables and in certain routing updates**
 - The default route is marked "0.0.0.0 0.0.0.0"
- **Hopefully, next router knows more about destination networks**
 - DR implies that another router might know more networks
- **Advantage: Smaller routing tables!**
 - DR permits routers to carry less than full routing tables

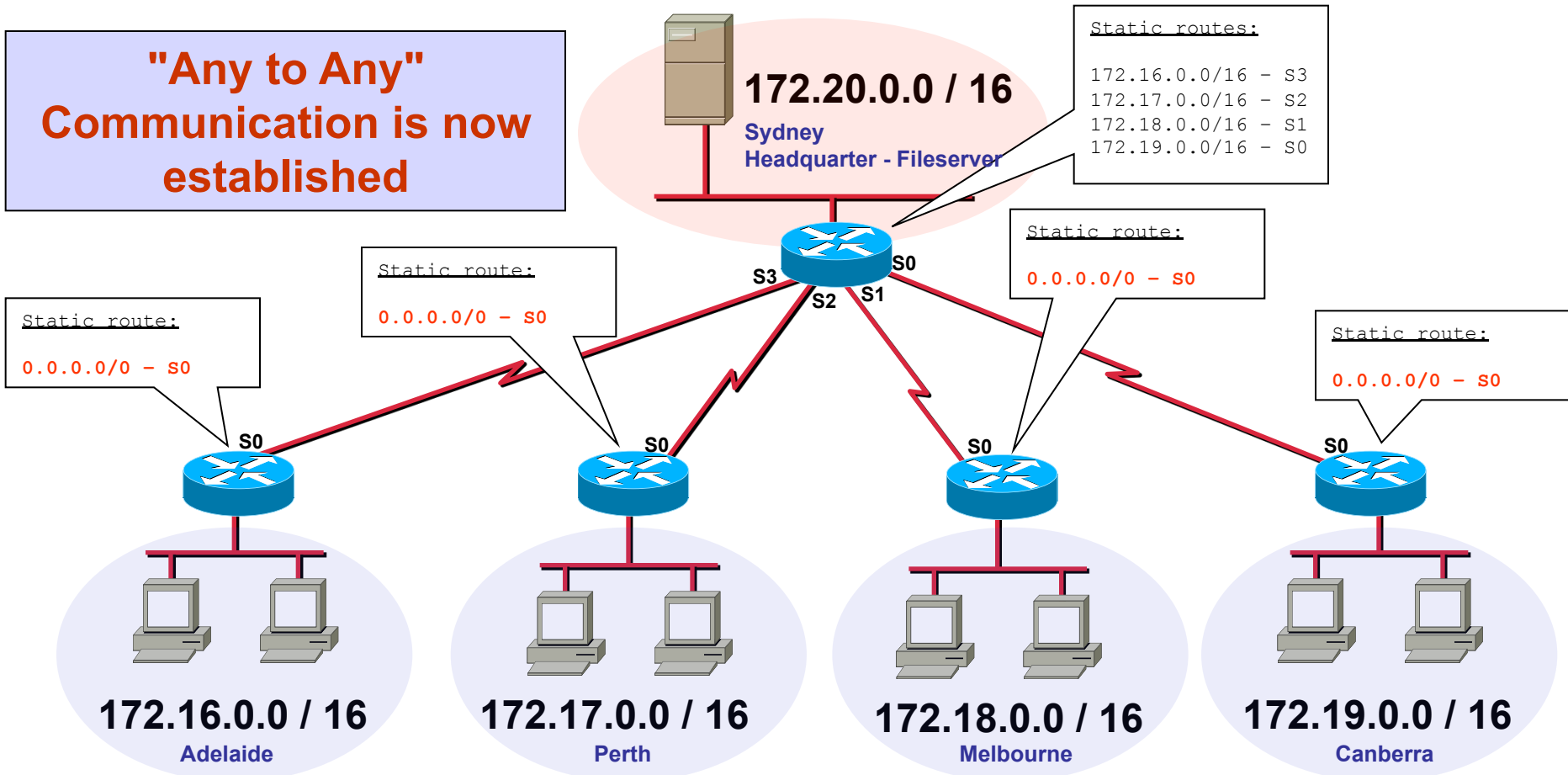
Default Routing (1)

- **Default Routes from stub networks**



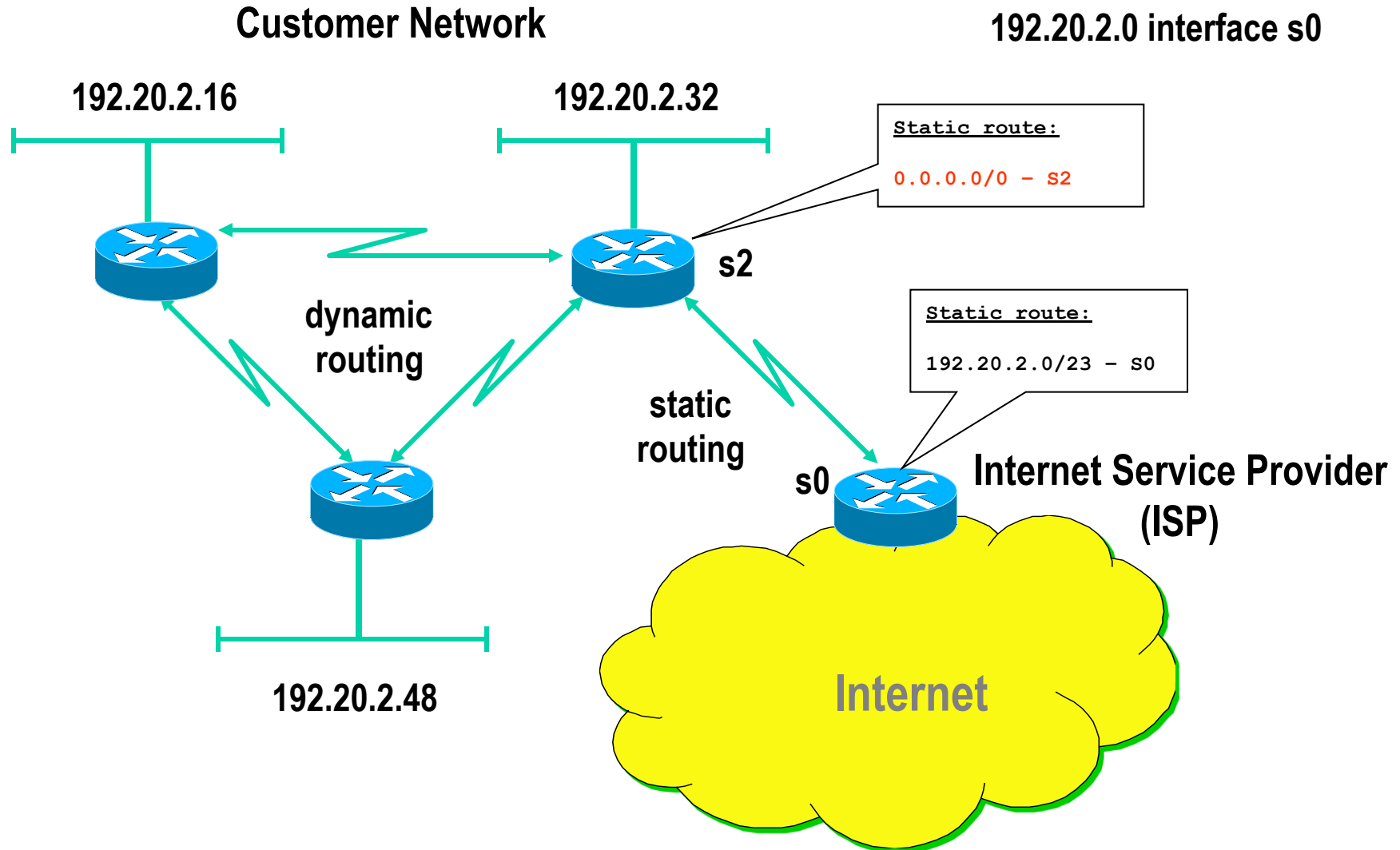
Default Routing (2)

- Default routes in "Hub and Spoke" topologies



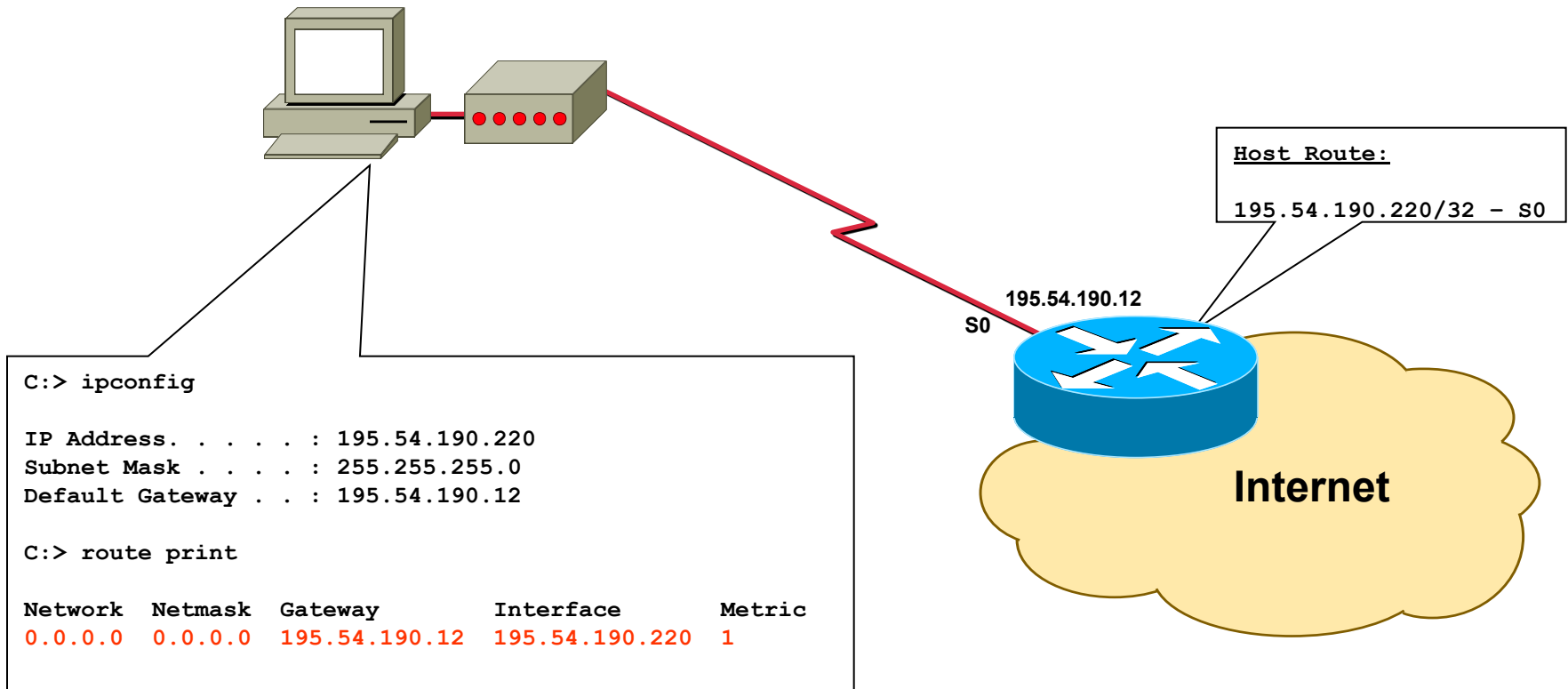
Default Routing (3) - Internet Access

Static Route Definition:
192.20.2.0 interface s0



Default Routing (4)

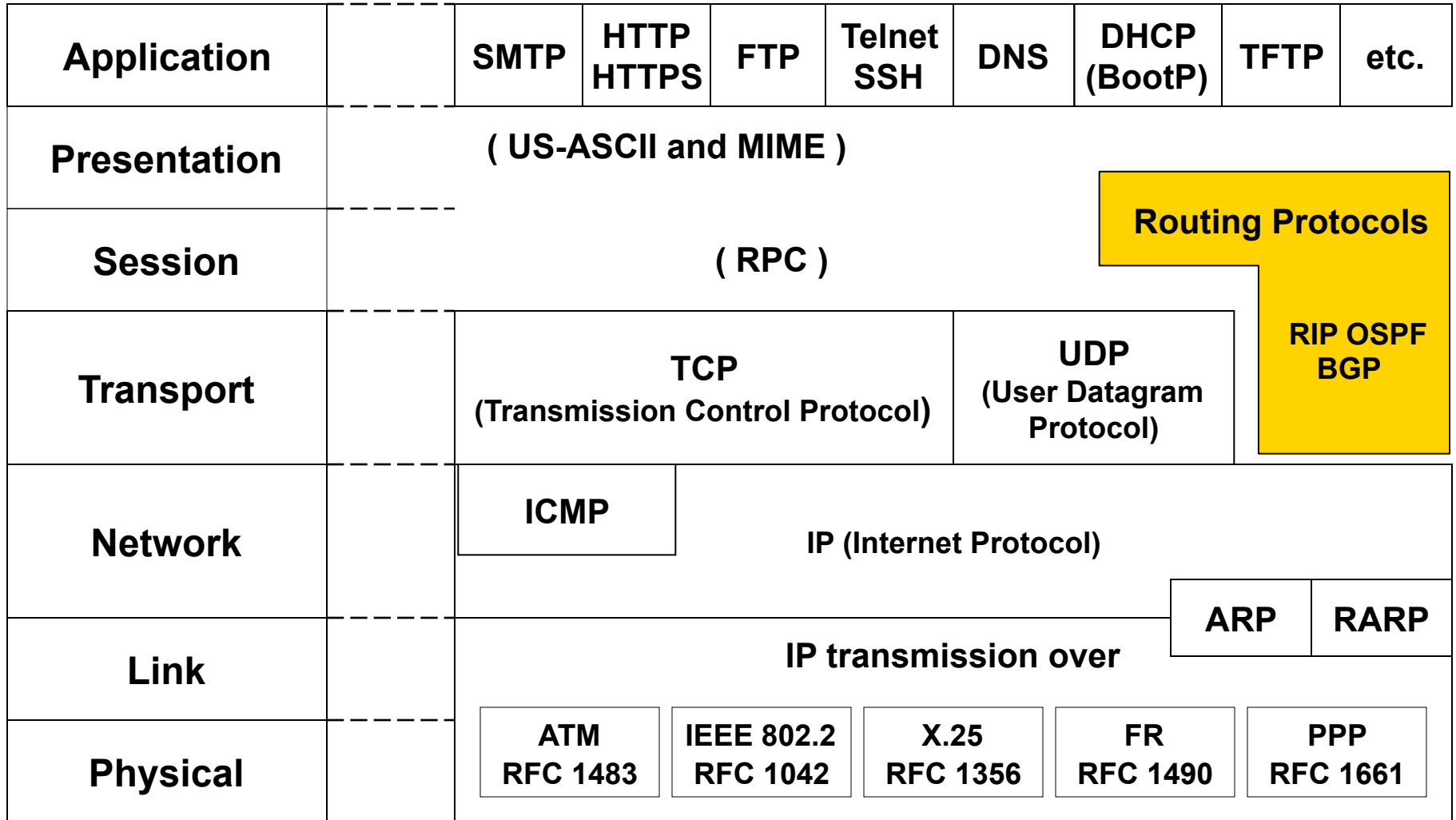
- Default Routes to the Internet



Agenda

- **Introduction to IP Routing**
 - Basics
 - Static Routing
 - Default Route
 - Dynamic Routing
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

TCP/IP Protocol Suite



Dynamic Routing

- **Basic principle**

- Routing tables are dynamically updated with information from other routers exchanged by routing protocols
- Routing protocol
 - Discovers current network topology
 - Determines the best path to every reachable network
 - Stores information about best paths in the routing table
- Metric information is necessary for best path decision
 - In most cases summarization of static preconfigured values along the given path
 - Hops, interface cost, interface bandwidth, interface delay, etc.
- Two basic technologies
 - Distance vector, Link state

Routing Metric

- **Routing protocols typically find out more than one route to the destination**
- **Metric help to decide which path to use**
 - Static values
 - Hop count, distance (RIP)
 - Cost like reciprocal value of bandwidth (OSPF)
 - Bandwidth (EIGRP), Delay (EIGRP), MTU
 - Variable or dynamic values
 - Load (EIGRP)
 - Reliability (EIGRP)
 - Very seldom used
 - Cisco citation:
“If you do not know what you are doing do not even think using or touching them!”

Dynamic Routing

- **Each router can run one or more routing protocols**
- **Routing protocols**
 - Are information sources to create routing table
 - Announce network reachability information
 - By doing this a router declares that traffic destined to a certain network can be sent to him
 - Network reachability information flows in the opposite direction to the traffic destined to a network
- **Routing protocols differ in**
 - Convergence time, loop avoidance, maximum network size, reliability and complexity

Routing Protocol Comparison

Routing Protocol	Complexity	Max. Size	Convergence Time	Reliability	Protocol Traffic
RIP	very simple	16 Hops	High (minutes)	Not absolutely loop-safe	High
RIPv2	very simple	16 Hops	High (minutes)	Not absolutely loop-safe	High
IGRP	simple	X	High (minutes)	Medium	High
EIGRP	complex	X	Fast (seconds)	High	Medium
OSPF	very complex	Thousands of Routers	Fast (seconds)	High	Low
IS-IS	complex	Thousands of Routers	Fast (seconds)	High	Low
BGP-4	very complex	more than 100,000 networks	Middle	Very High	Low

Administrative Distance

Longest Match Routing Rule

- Several routing protocols independently find out different routes to same destination
 - Which one to choose?
- "Administrative Distance" is a **trustiness-value** associated to each routing protocol
 - The lower the better
 - Can be changed
- **Note:**
 - If a destination network (seen in an IP datagram) matches more than one entry in the routing table
 - Then "Longest Match Routing Rule" is used and the best match will be taken
 - Best means the highest amount of bits from left to right in a given IP address are identical to the routing entry

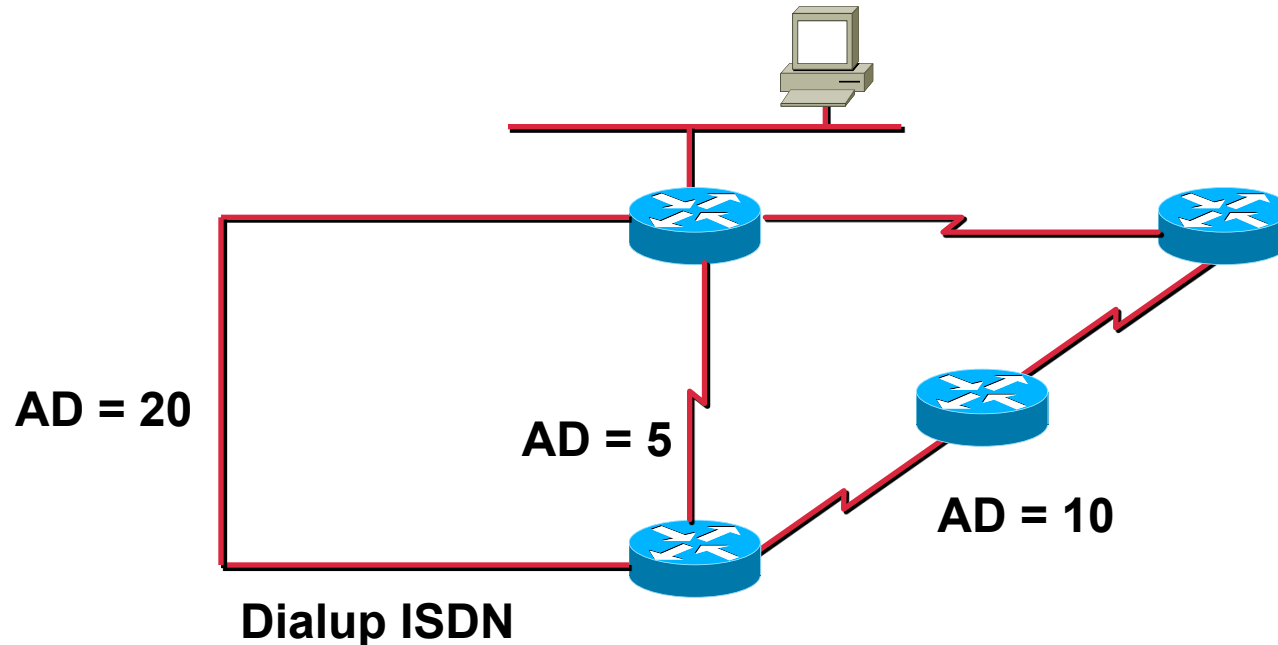
Administrative Distances Chart

FYI

Unknown	255
I-BGP	200
E-EIGRP	170
EGP	140
RIP	120
IS-IS	115
OSPF	110
IGRP	100
I-EIGRP	90
E-BGP	20
EIGRP Summary Route	5
Static route to next hop	1
Static route through interface	0
Directly Connected	0

AD with Static Routes

- Each static route can be given a different administrative distance
- This way fall-back routes can be configured



Classification of Routing Protocols

- **Depending on age:**
 - Classful (no subnet masks)
 - Routing updates carries IP net-ID only
 - Classless (VLSM/CIDR supported)
 - Routing updates carries IP net-ID and subnet mask
 - Very often prefix/length notation is used !!!
- **Depending on scope:**
 - IGP (inside an Autonomous System)
 - EGP (between Autonomous Systems)
- **Depending on algorithm:**
 - Distance Vector (Signpost principle)
 - Link State (Roadmap principle)
 - Hybrid (mixture of distance vector and link state)

Routing Table Example

Output of Cisco CLI command "show ip route":

C ... Directly Connected

R ... Learnt from RIP

S ... Static Route

S*... Default Route

administrative
distance

RIP metric:
hop count

last seen in RIP
update message 5
seconds ago

Gateway of last resort is 175.18.1.2 to network 0.0.0.0

10.0.0.0 255.255.0.0 is subnetted, 4 subnets

C 10.1.0.0 is directly connected, Ethernet1

R 10.2.0.0 [120/1] via 10.4.0.1, 00:00:05, Ethernet0

R 10.3.0.0 [120/5] via 10.4.0.1, 00:00:05, Ethernet0

C 10.4.0.0 is directly connected, Ethernet0

R 192.168.12.0 [120/3] via 10.1.0.5, 00:00:08, Ethernet1

S 194.30.222.0 [1/0] via 10.4.0.1

S 194.30.223.0 [1/0] via 10.1.0.5

C 175.18.1.0 255.255.255.0 is directly connected, Serial0

S* 0.0.0.0 0.0.0.0 [1/0] via 175.18.1.2

network 0.0.0.0 subnet mask 0.0.0.0
means all destination addresses matches this entry

Distance Vector Protocols (1)

- After powering-up each router only knows about directly attached networks
- **Routing table** is sent periodically to all neighbor-routers
- Received updates are examined, changes are adopted in own routing table
 - Changes announced by next periodic routing update
- **Metric information is based on hops (distance between hops)**
 - Hop count metric is a special case for the more generic distance value between two routers
 - Hop count means distance = 1 between any two neighboring routers
- **"Bellman-Ford" algorithm**

Distance Vector Protocols (2)

- **Limited view of topology**
 - Next hop is always originating router
 - Topology behind next hop unknown
 - **Signpost principle**
- **Loops can occur!**
- **Additional mechanisms needed**
 - Maximum hop count
 - Split horizon (with poison reverse)
 - Triggered update
 - Hold down
 - Route Poisoning

Distance Vector Protocols (3)

- **Examples**

- RIP, RIPv2 (Routing Information Protocol)
- IGRP (Cisco, Interior Gateway Routing Protocol)
- IPX RIP (Novell)
- AppleTalk RTMP (Routing Table Maintenance Protocol)

Link State Protocols (1)

- **Each two neighbored routers establish adjacency**
- **Routers learn real topology information**
 - Through "Link State Advertisements (LSAs)"
 - Stored in database (**Roadmap principle**)
- **Routers have a global view of network topology**
 - Exact knowledge about all routers, links and their costs (metric) of a network
- **Updates only upon topology changes**
 - Propagated by *flooding* of LSAs (very fast convergence)

Link State Protocols (2)

- **Routing table entries are calculated by applying the **Shortest Path First (SPF)** algorithm on the database**
 - Loop-safe
 - Only the lowest cost path is stored in routing table
 - But alternative paths are immediately known
 - Could be CPU and memory greedy
 - Mainly a concern in the past
- **Large networks can be split into **areas****

Link State Protocols (3)

- **With the lack of topology changes**
 - Local hello messages are used to supervise local links (to test reachability of immediate-neighboring routers)
 - Therefore less routing overhead concerning link bandwidth than periodic updates of distance vector protocols
- **But more network load is caused by such a routing protocol**
 - During connection of former separated parts of a network
 - During topology database synchronization

Link State Protocols (4)

- **Examples**

- OSPF (Open Shortest Path First)
- Integrated IS-IS (IP world)
 - note: Integrated IS-IS takes another approach to handle large networks (topic outside the scope of this course)
- IS-IS (OSI world)
- PNNI (in the ATM world)
- APPN (IBM world),
- NLSP (Novell world)

Summary

- **Routing is the "art" of finding the best way to a given destination**
- **Can be static or dynamic**
 - Static means: YOU are defining the way packets are going
 - Dynamic means: A routing protocol is "trying" to find the best way to a given destination
- **In today's routers the route with the longest match is used**
- **Routing protocols either implement the principle *Distance Vector* or *Link State***

Agenda

- **Introduction to IP Routing**
- **RIP**
 - Introduction
 - Split Horizon
 - Count-To-Infinity
 - Max-Hop-Count
 - Poison Reverse
 - Hold Down
 - Some Details and Summary
 - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

RIPv1 - Routing Information Protocol

- **Interior Gateway Protocol (IGP)**
 - Due to inherent administrative overhead traffic, RIP suits best only for smaller networks
 - Routing decisions are based upon hop count measure
- **Distance-Vector Routing Protocol**
 - Bellman Ford Algorithm
 - RFC 1058 released in 1988
- **Classful**
 - No subnet masks carried
- **RIPv1 was initially released as part of BSD 4.2 UNIX**
 - Hence RIP got wide-spread availability
- **RIPv1 is specified in RFC 1058**
 - RFC category „historic“

RIP Basics

- **Signpost principle**

- Own routing table is sent periodically (every 30 seconds)

- **What is a signpost made of ?**

- Destination network
- Hop Count (metric, "distance")
- Next Hop ("vector", given implicitly by sender's address!)

- **Receiver of update extracts new information**

- New is information about a network either not known so far or an already known network with a better metric
- Already known routes with worse metric are ignored
- Adapts the routing table and again sent periodically its routing table

"Routing By Rumor"



- **Good news propagate quickly**
 - 30 seconds per network
- **Bad news are ignored**
 - Except when sent by routers from which these routes had been learned initially
 - But better news from ANY router will be preferred
- **A network disappears from the routing table**
 - If not refreshed within 180 seconds by some routing updates
- **Hence unreachability of networks is propagated very slowly**
 - At least 180 seconds

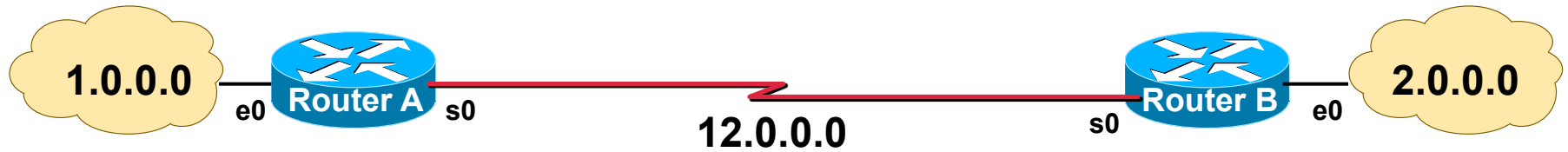
RIPv1 Message Format

0	8	16	31
Command		Version	must be zero
Address Family Identifier for Net1		must be zero	
IP address of Net 1			
must be zero			
must be zero			
Distance to Net 1 = Metric			
Address Family Identifier for Net 2		must be zero	
IP address of Net 2			
must be zero			
must be zero			
Distance to Net 2 = Metric			
Address Family Identifier for Net 3		must be zero	
.....			

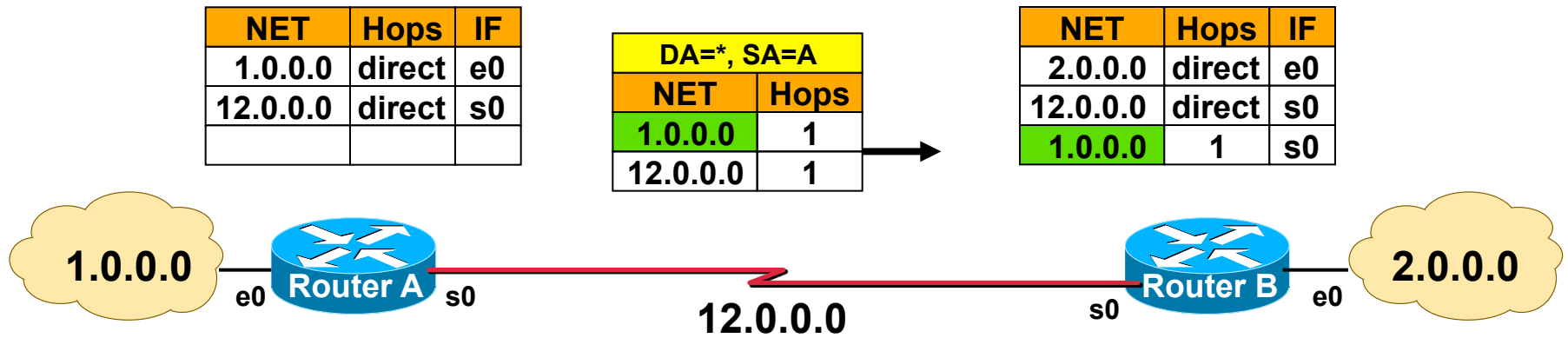
Routing Tables after Power On (1)

NET	Hops	IF
1.0.0.0	direct	e0
12.0.0.0	direct	s0

NET	Hops	IF
2.0.0.0	direct	e0
12.0.0.0	direct	s0



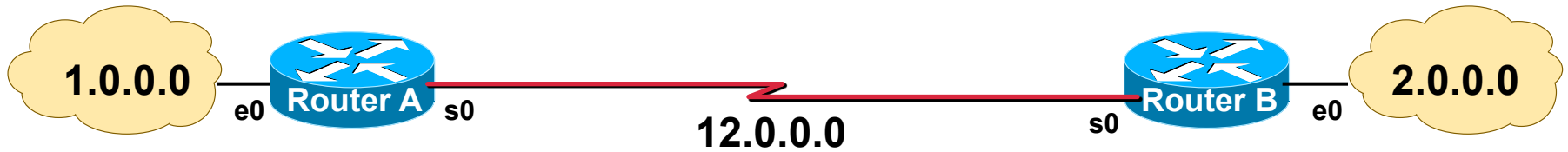
First Update Router A (2)



Update Router B (3)

NET	Hops	IF
1.0.0.0	direct	e0
12.0.0.0	direct	s0
2.0.0.0	1	s0

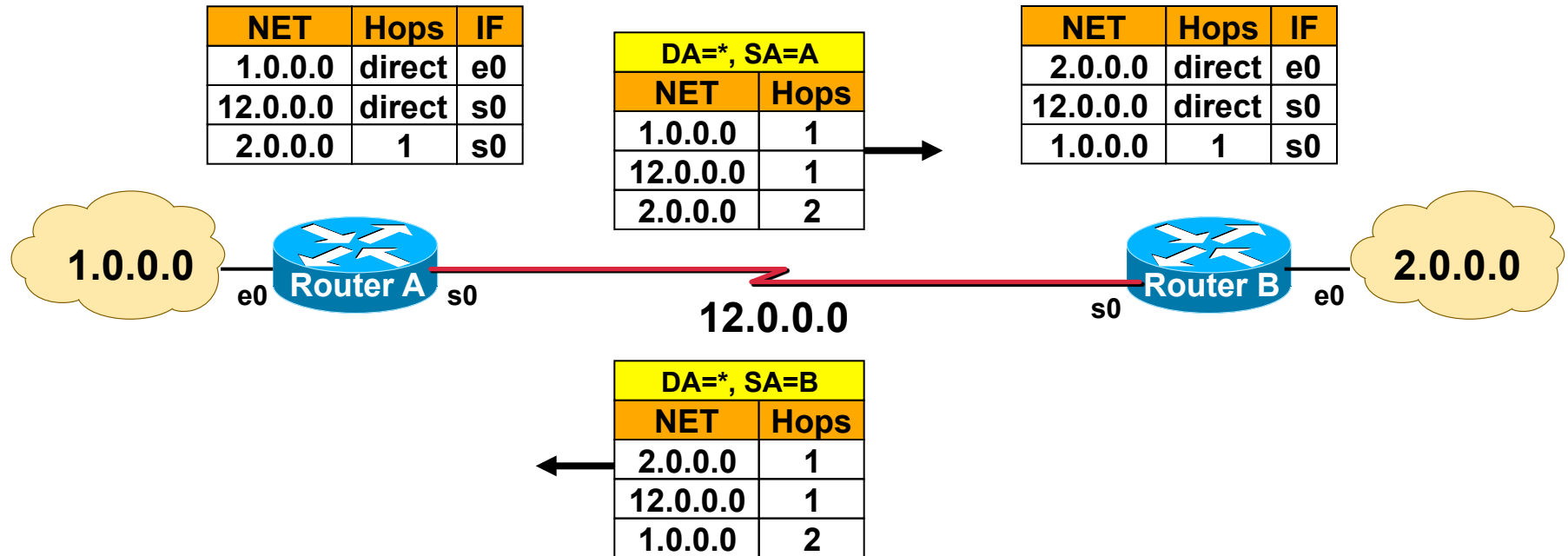
NET	Hops	IF
2.0.0.0	direct	e0
12.0.0.0	direct	s0
1.0.0.0	1	s0



DA=*, SA=B	
NET	Hops
2.0.0.0	1
12.0.0.0	1
1.0.0.0	2



Periodic Updates for Refreshing (4)

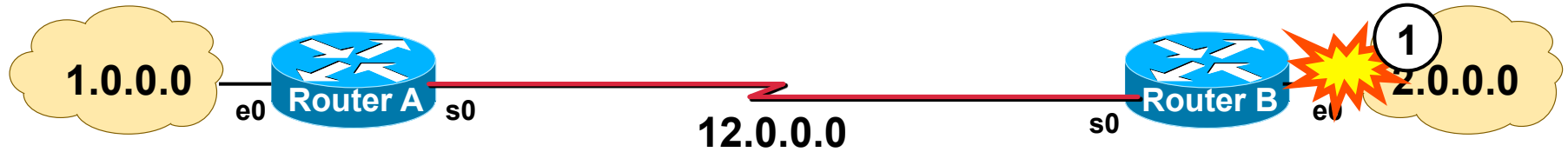


Topology Change (1)

(Without Split Horizon)

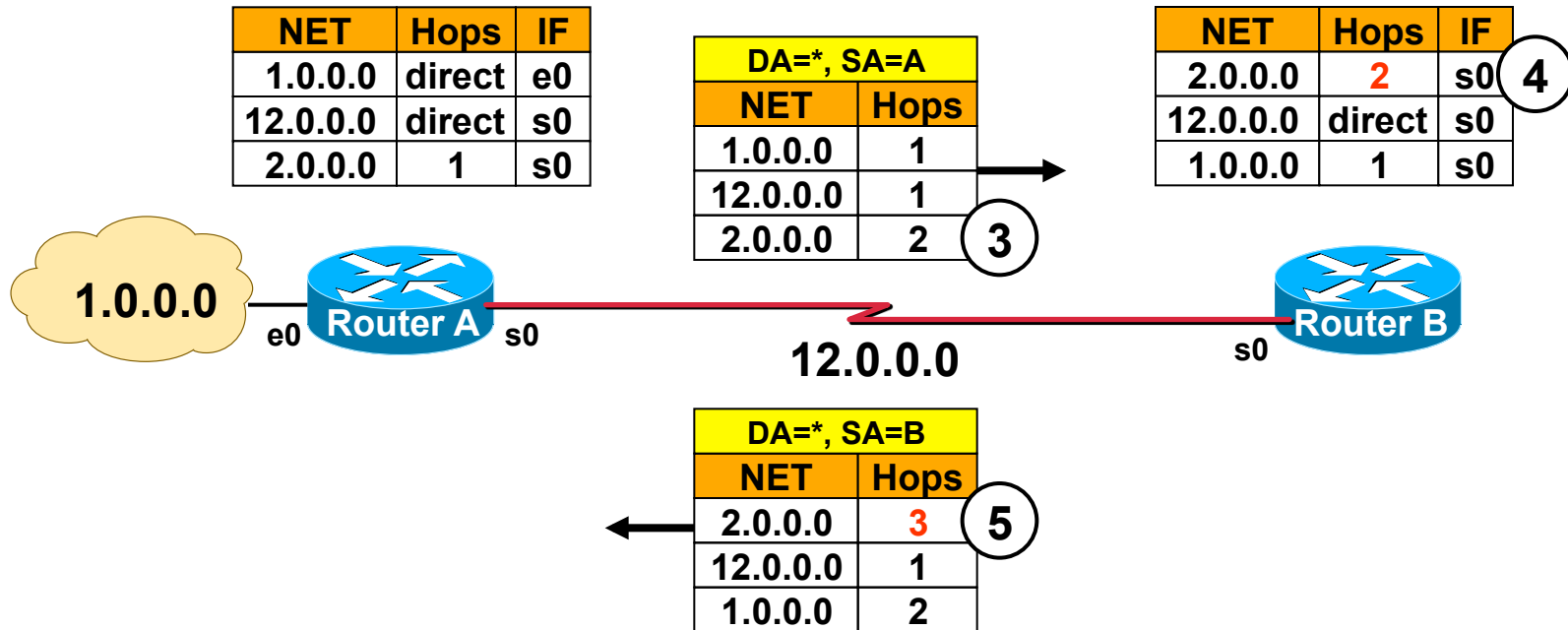
NET	Hops	IF
1.0.0.0	direct	e0
12.0.0.0	direct	s0
2.0.0.0	1	s0

NET	Hops	IF
2.0.0.0	???	??
12.0.0.0	direct	s0
1.0.0.0	1	s0



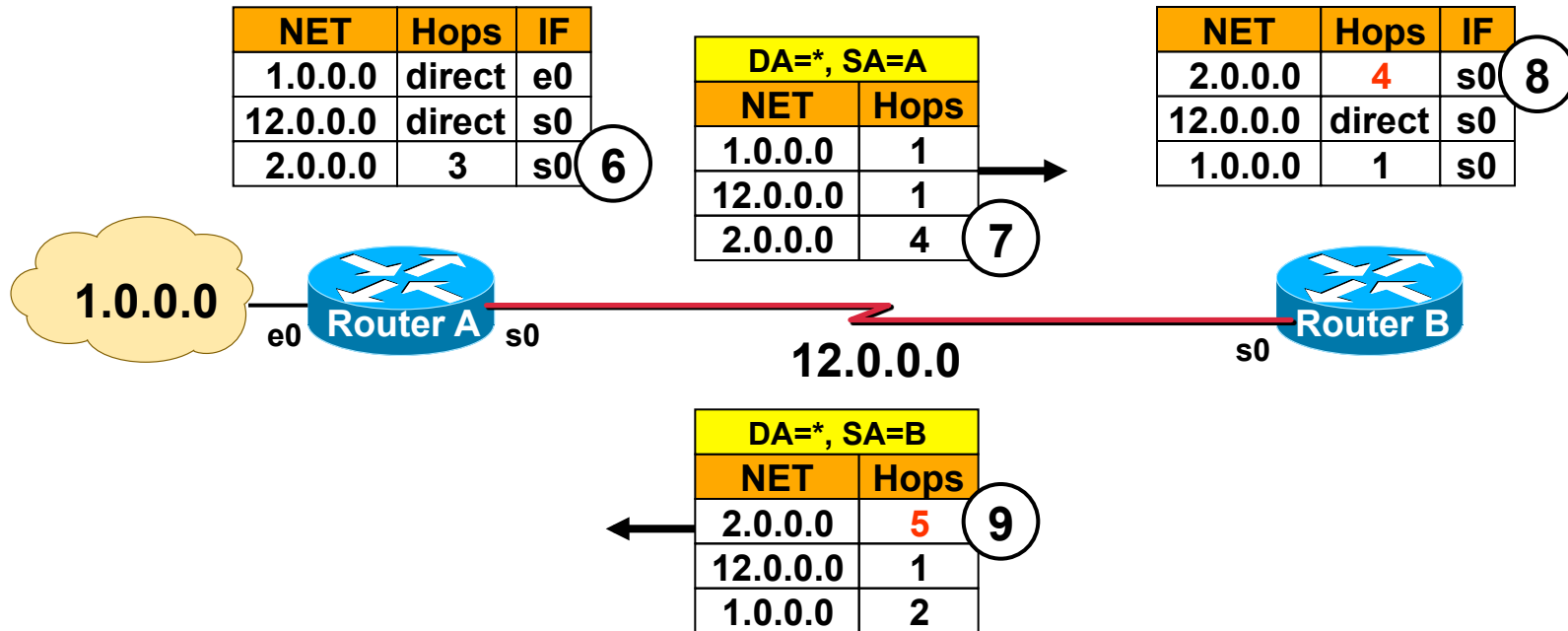
Topology Change (2)

(Without Split Horizon)



Topology Change (3)

(Without Split Horizon)



...Count to Infinity...

During count to infinity datagrams to network 2.0.0.0 are caught in a routing loop

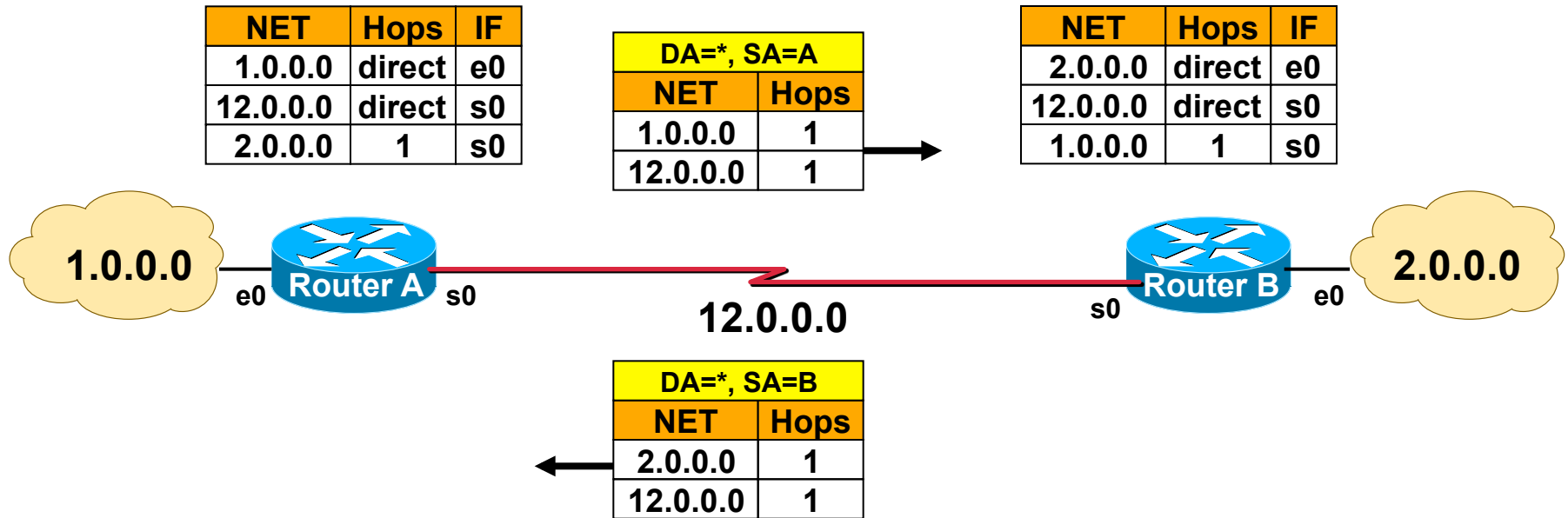
Agenda

- **Introduction to IP Routing**
- **RIP**
 - Introduction
 - Split Horizon
 - Count-To-Infinity
 - Max-Hop-Count
 - Poison Reverse
 - Hold Down
 - Some Details and Summary
 - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

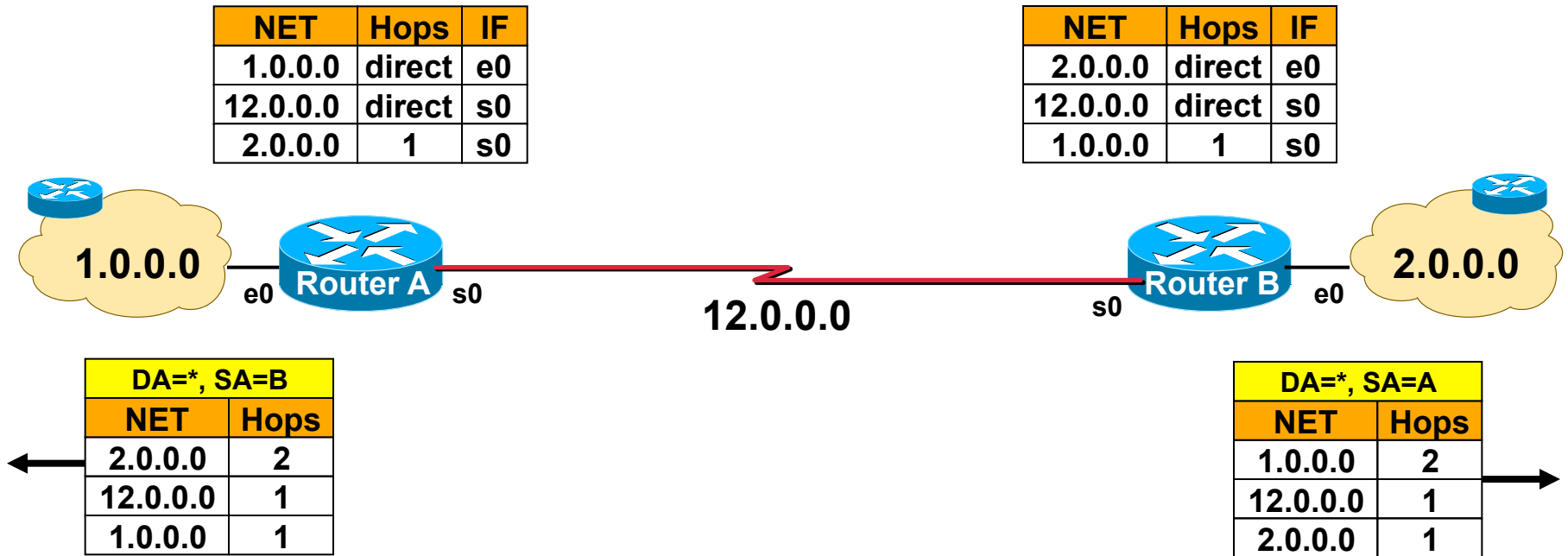
Split Horizon

- **A router will not send information about routes through an interface over which the router has learned about those routes**
 - Exactly THIS is split horizon
- **Idea: "Don't tell neighbor of routes that you learned from this neighbor"**
 - That's what humans (almost) always do:
Don't tell me what I've told you !
- **Split horizon**
 - Cannot 100% avoid all routing loops!
- **See RIP at work with split horizon on the following slides**

Periodic Updates With Split Horizon (1)



Periodic Updates With Split Horizon (2)

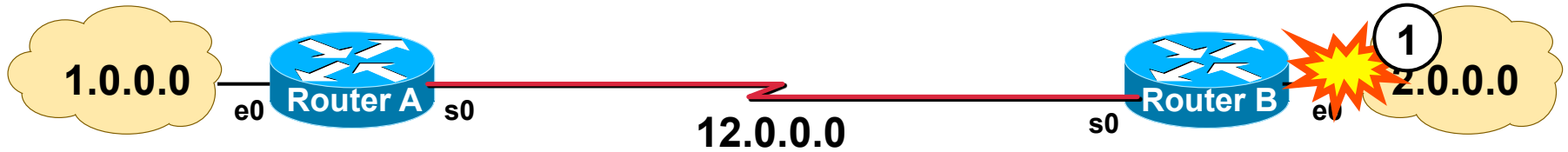


Topology Change (1)

(With Split Horizon)

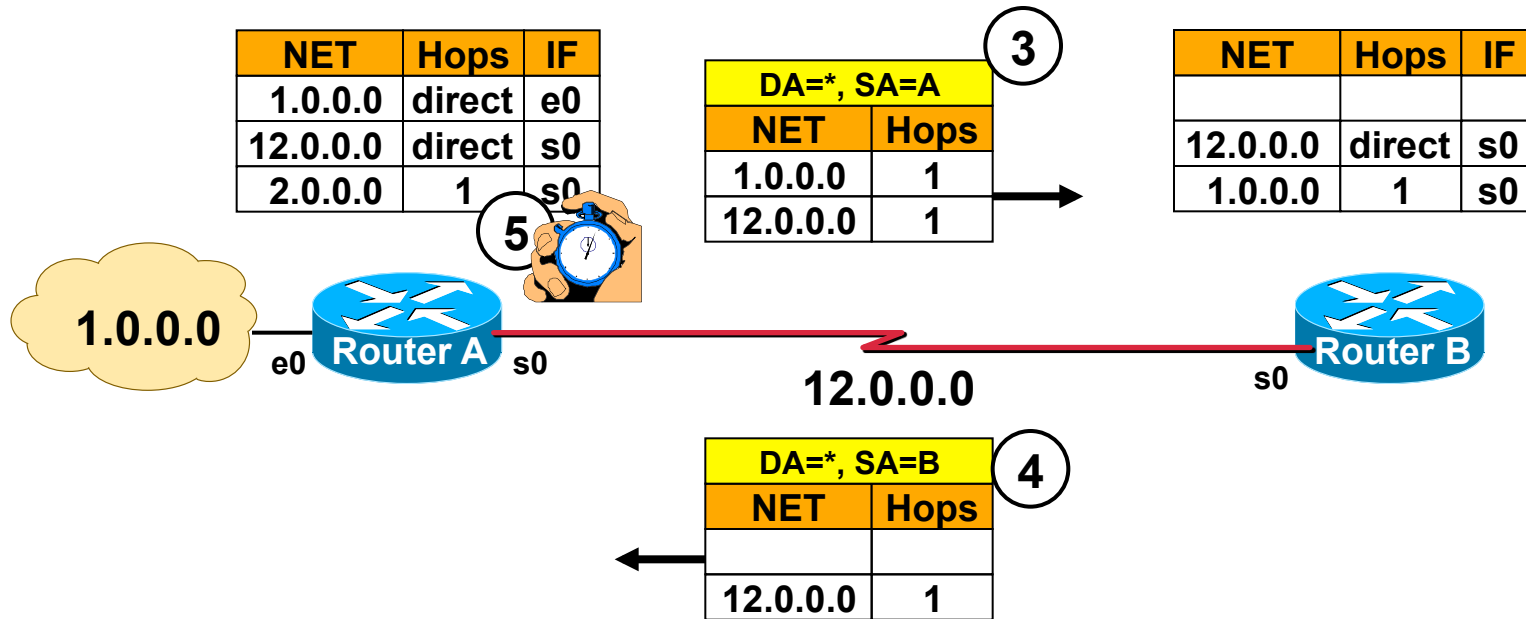
NET	Hops	IF
1.0.0.0	direct	e0
12.0.0.0	direct	s0
2.0.0.0	1	s0

NET	Hops	IF
2.0.0.0	???	??
12.0.0.0	direct	s0
1.0.0.0	1	s0

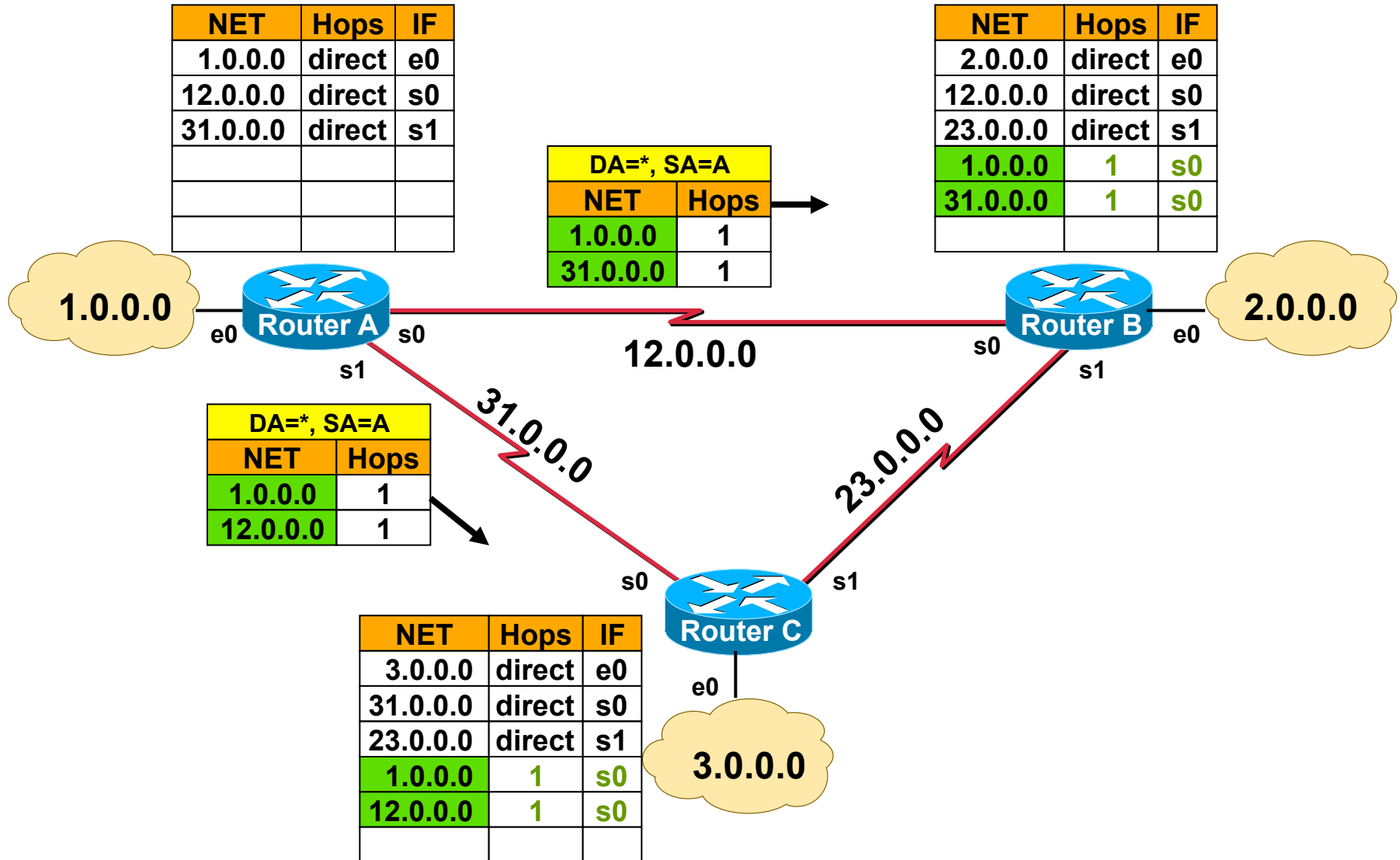


Topology Change (2)

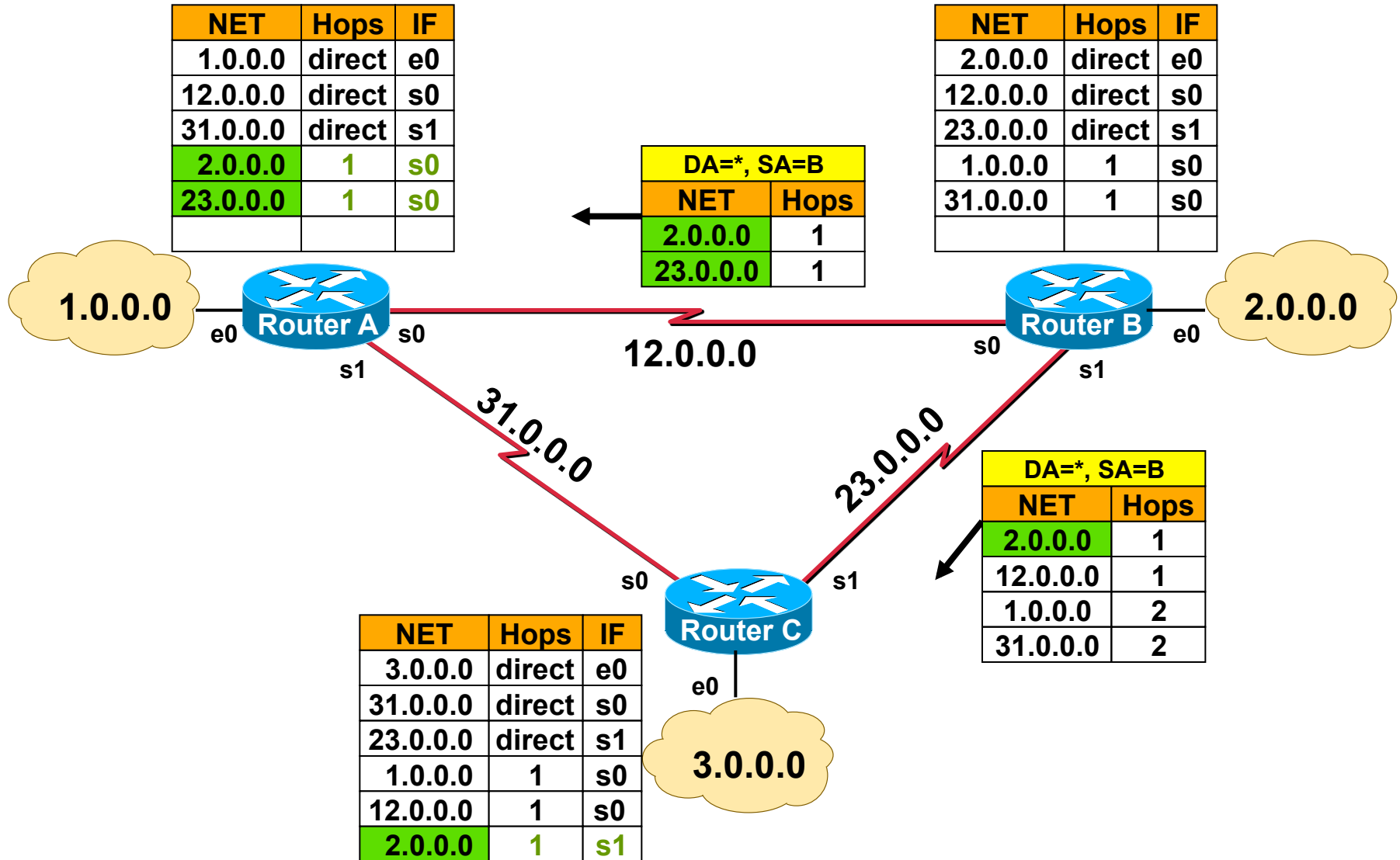
(With Split Horizon)



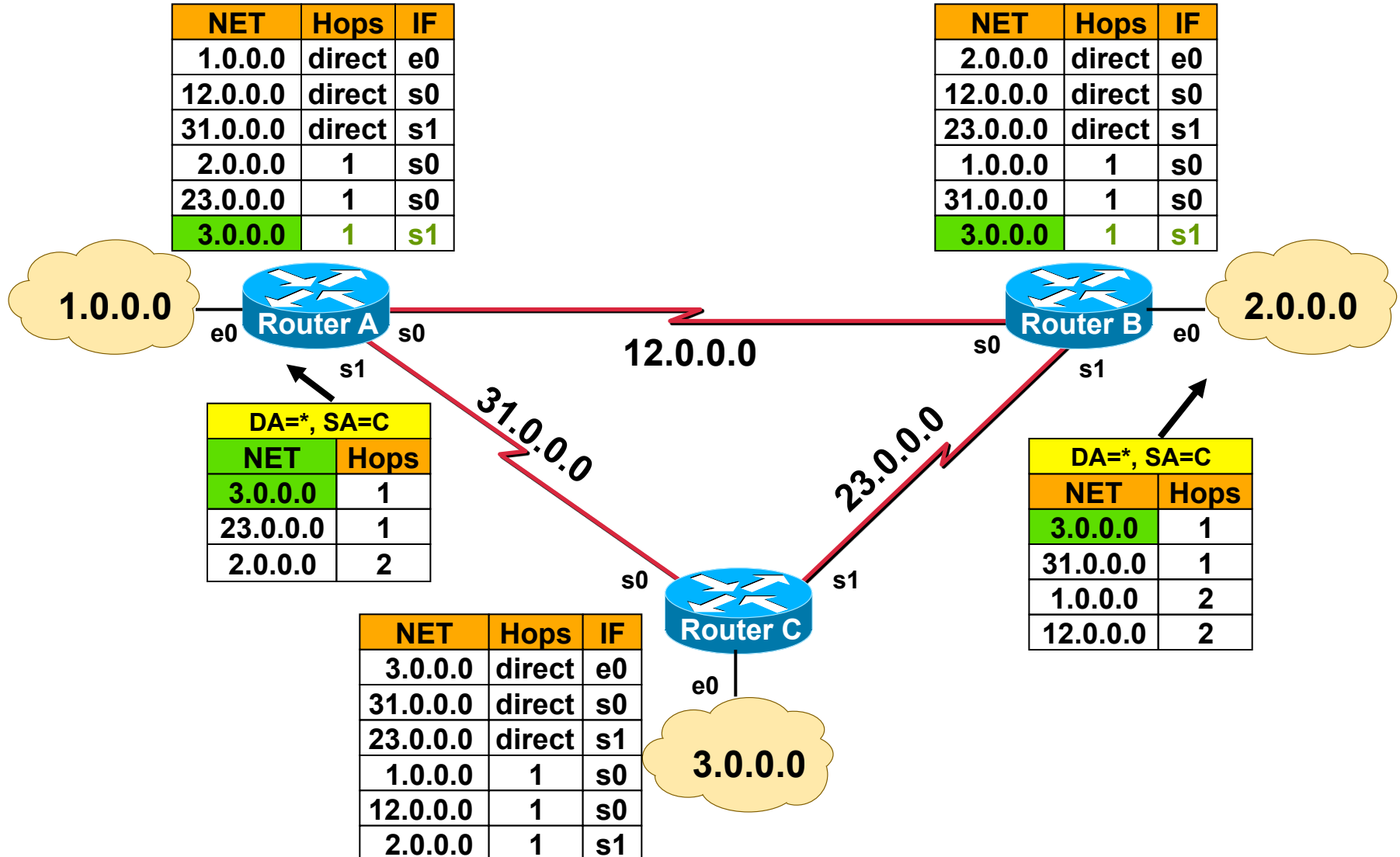
RIP At Work (Update Router A)



RIP At Work (Update Router B)



RIP At Work (Update Router C)



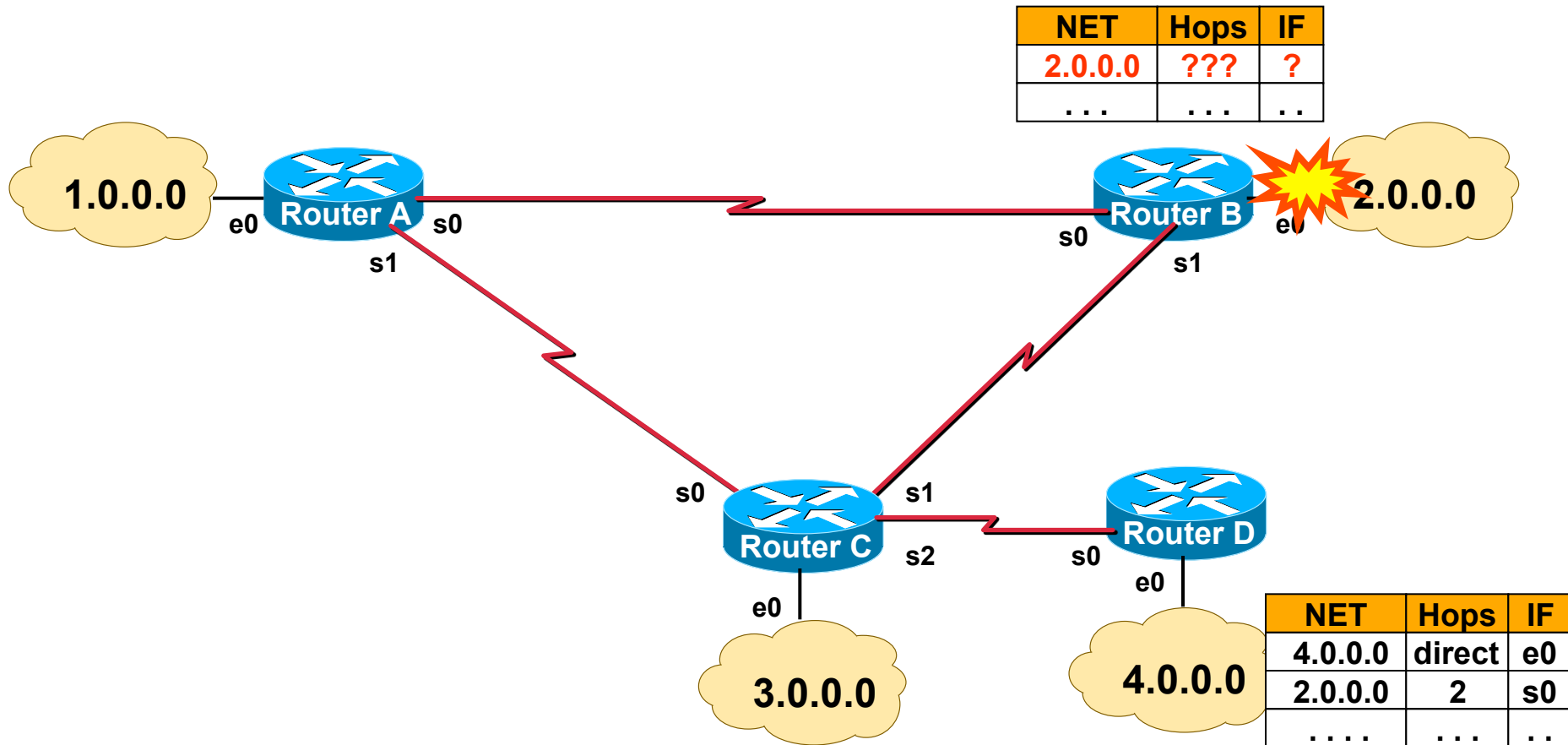
Agenda

- **Introduction to IP Routing**
- **RIP**
 - Introduction
 - Split Horizon
 - Count-To-Infinity
 - Max-Hop-Count
 - Poison Reverse
 - Hold Down
 - Some Details and Summary
 - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

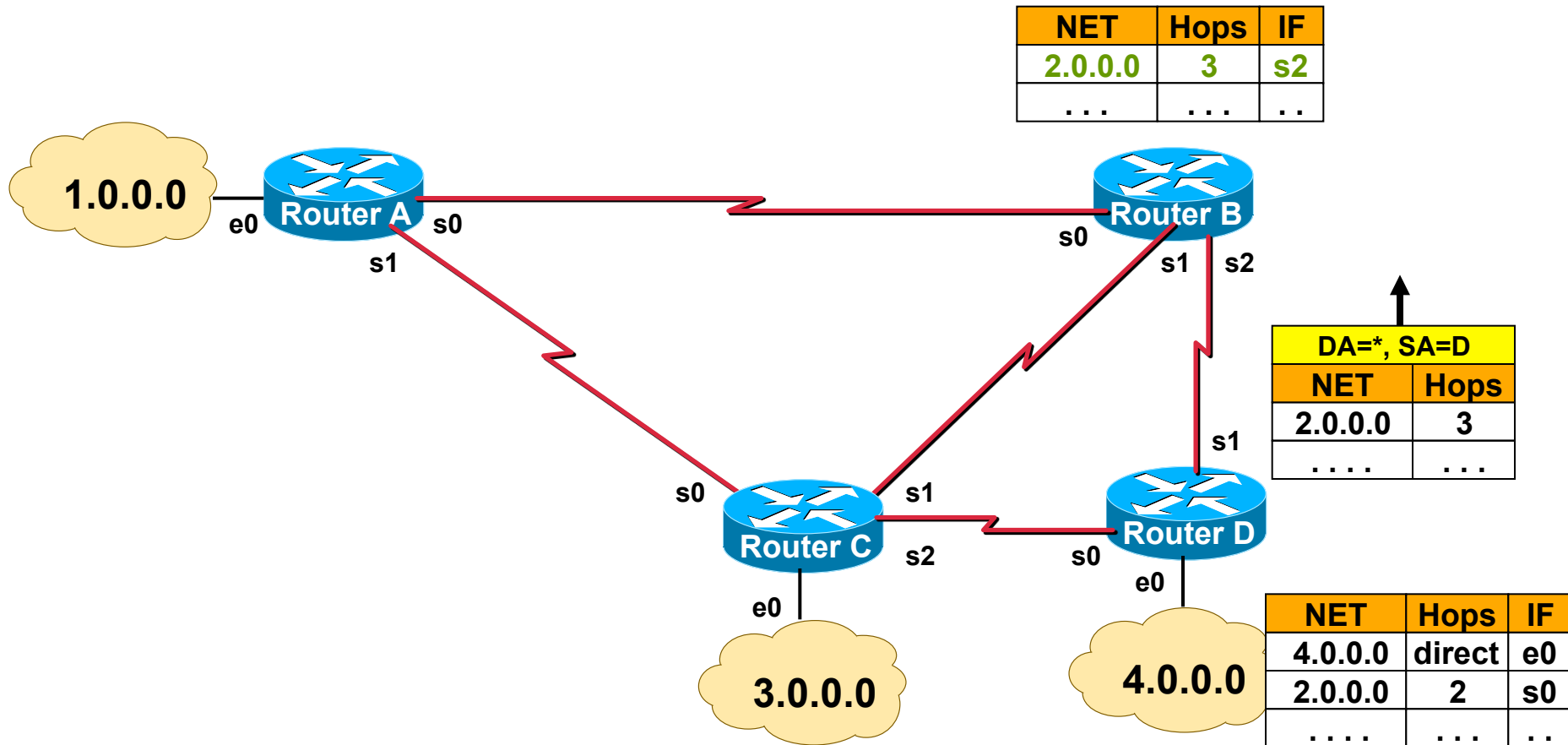
Count To Infinity

- **Main problem with distance vector protocols**
- **Unforeseeable situations can still lead to count to infinity**
 - Access lists
 - Disconnection and connections
 - Router malfunctions
 -
- **During that time, routing loops occur!**
- **We need an additional element**
 - Maximum Hop Count = 16 for RIP
 - Hop count =16 can also be used as unreachability message

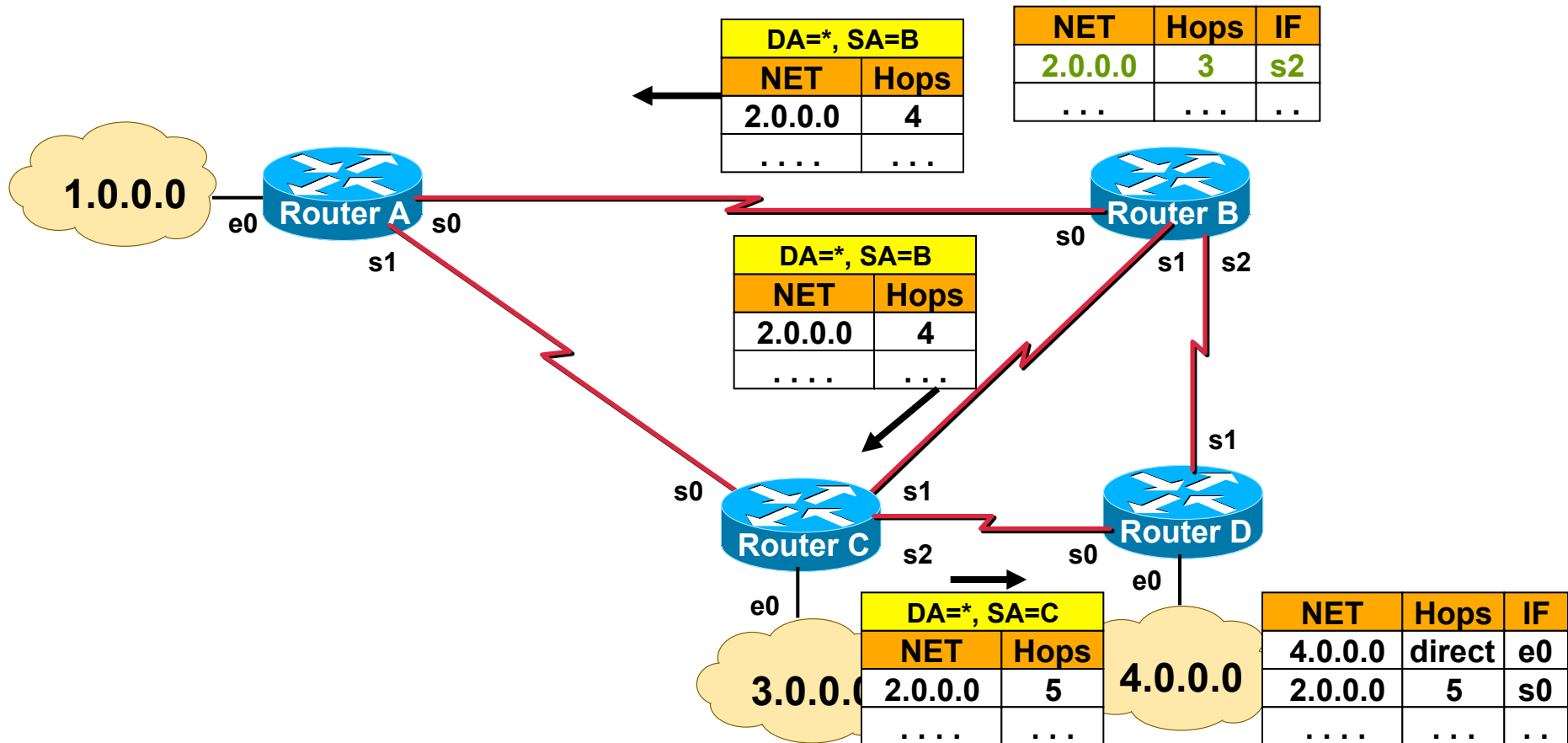
Count To Infinity (1)



Count To Infinity (2)



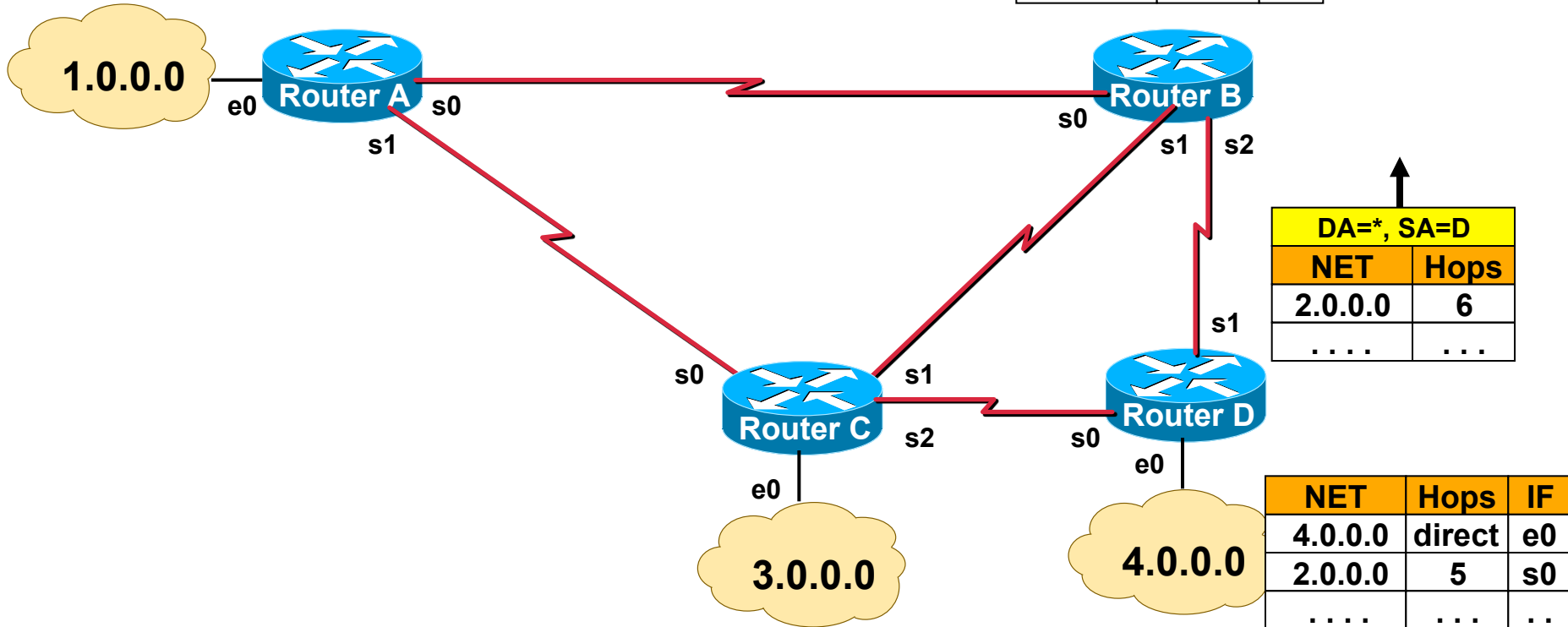
Count To Infinity (3)



Count To Infinity (4)

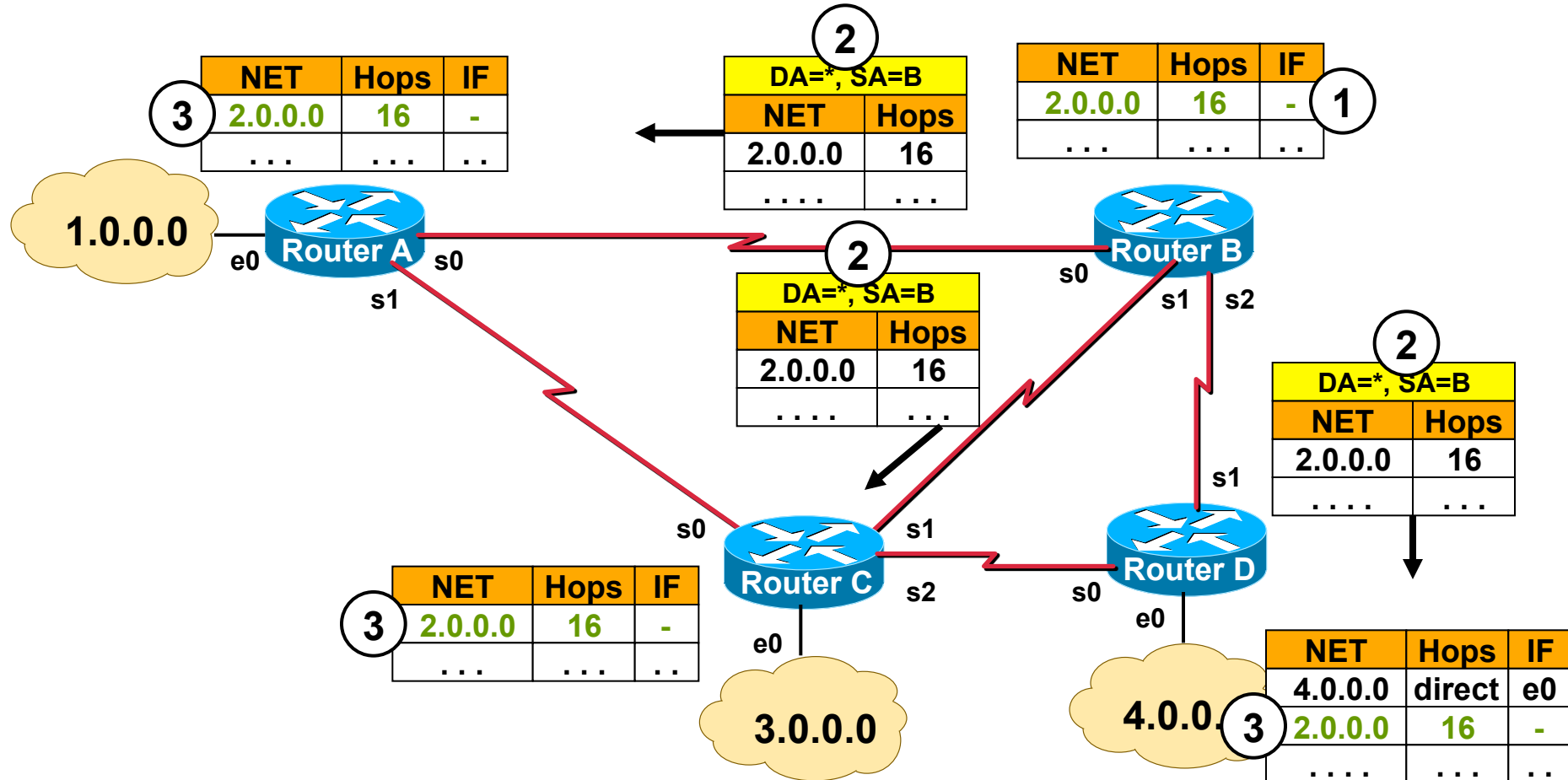
Count to Infinity situations cannot be avoided in any situation (drawback of signpost principle)

Basic solution: **Maximum Hop Count = 16**



Maximum Hop Count = 16

Reaching hop count 16, the route is marked as **INVALID** (1) and propagated for a certain time (2) to inform neighbors about unreachability of network 2.0.0.0.

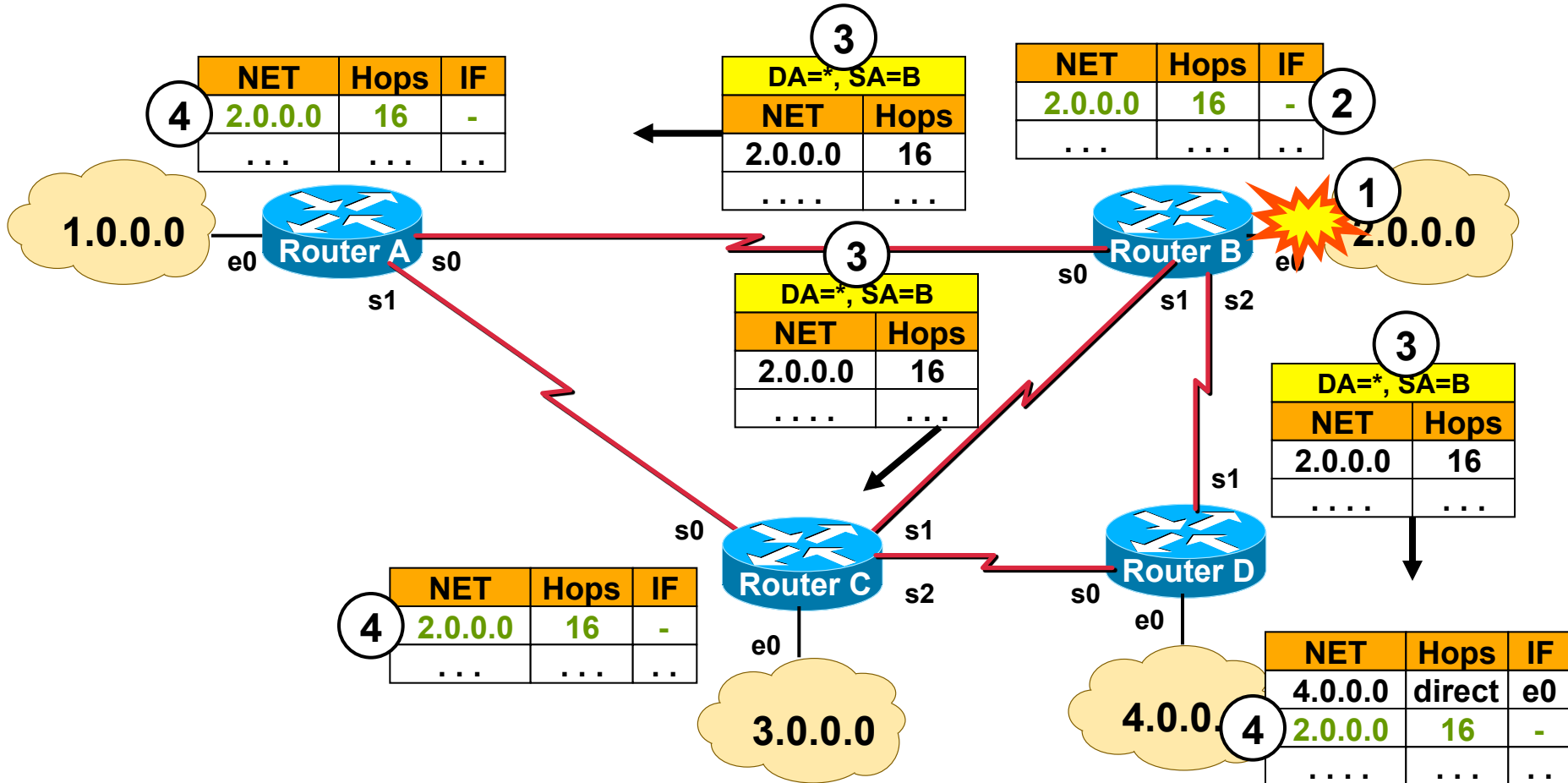


Agenda

- **Introduction to IP Routing**
- **RIP**
 - Introduction
 - Split Horizon
 - Count-To-Infinity
 - Max-Hop-Count
 - Poison Reverse
 - Hold Down
 - Some Details and Summary
 - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

Maximum Hop Count = 16

Upon network failure, the route is marked as **INVALID** (hop count 16) and propagated.



Maximum Hop Count

- **Defining a maximum hop count of 16 provides a basic safety factor**
- **But restricts the maximum network diameter**
- **Routing loops might still exist during 480 seconds (16×30s)**
- **Therefore several additional measures are necessary**
 - Split Horizon
 - Poison Reverse
 - Hold Down
 - Triggered Update

Agenda

- **Introduction to IP Routing**
- **RIP**
 - Introduction
 - Split Horizon
 - Count-To-Infinity
 - Max-Hop-Count
 - Poison Reverse
 - Hold Down
 - Some Details and Summary
 - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

Additional Measures (1)

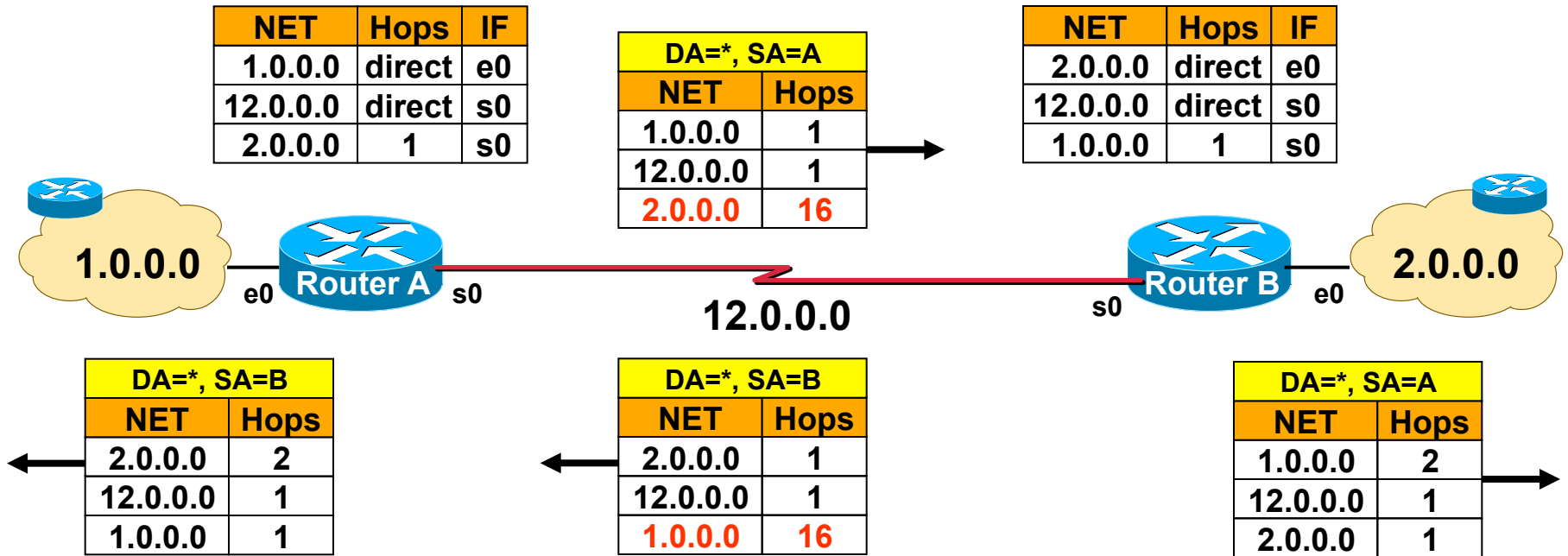
- **Split Horizon**

- Suppressing information that the other side should know better
 - Used during normal operation but cannot prevent routing loops !!!
- Remember: good news overwrite bad news
 - Unreachable information could be overwritten by uninformed routers (which are beyond scope of split horizon)

- **Poison Reverse**

- Alternate approach split horizon
- Declare learned routes as unreachable
- "Bad news is better than no news at all"
- Stops potential loops due to corrupted routing updates

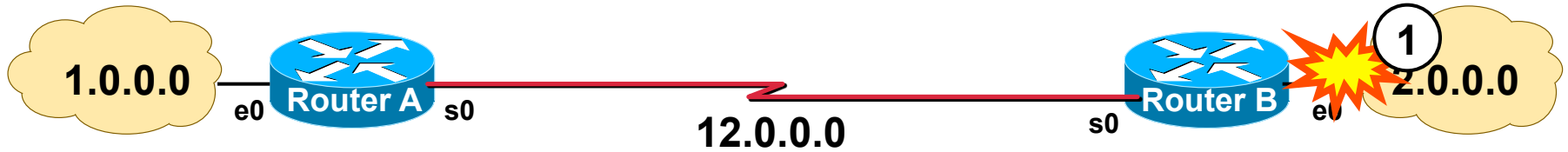
Poison Reverse At Work (1)



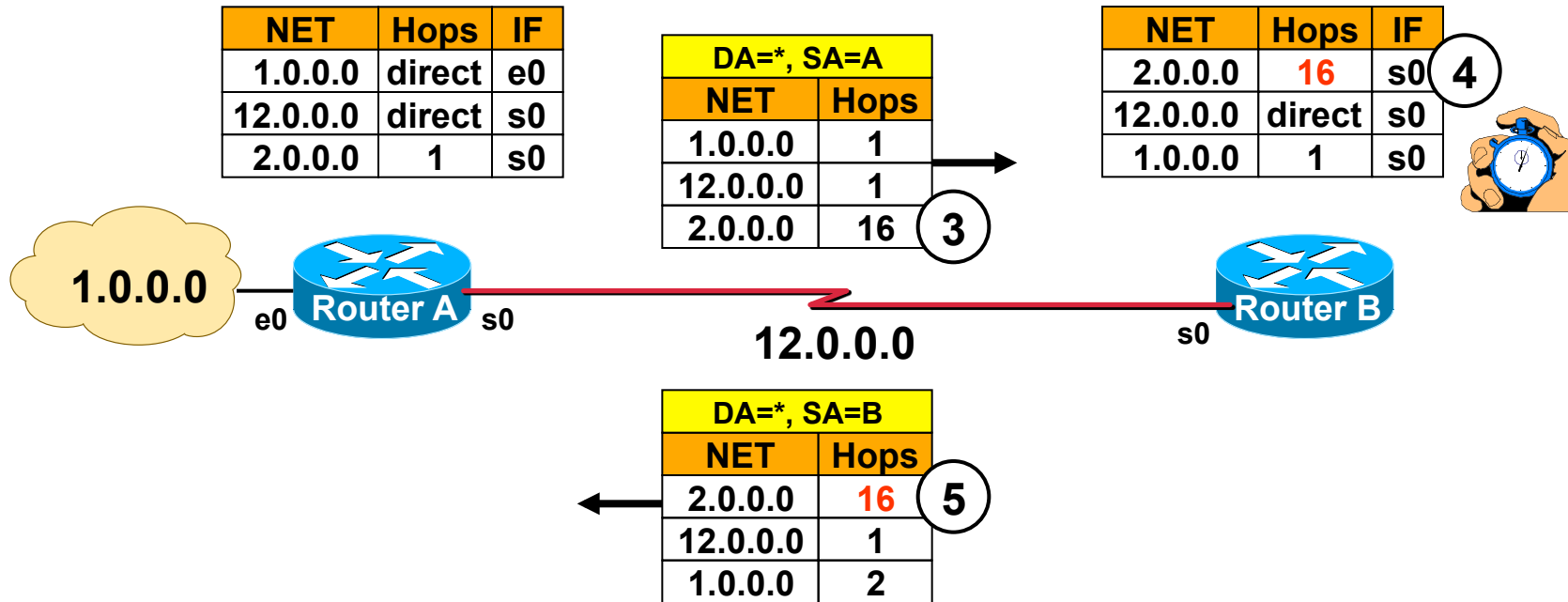
Poison Reverse At Work (2)

NET	Hops	IF
1.0.0.0	direct	e0
12.0.0.0	direct	s0
2.0.0.0	1	s0

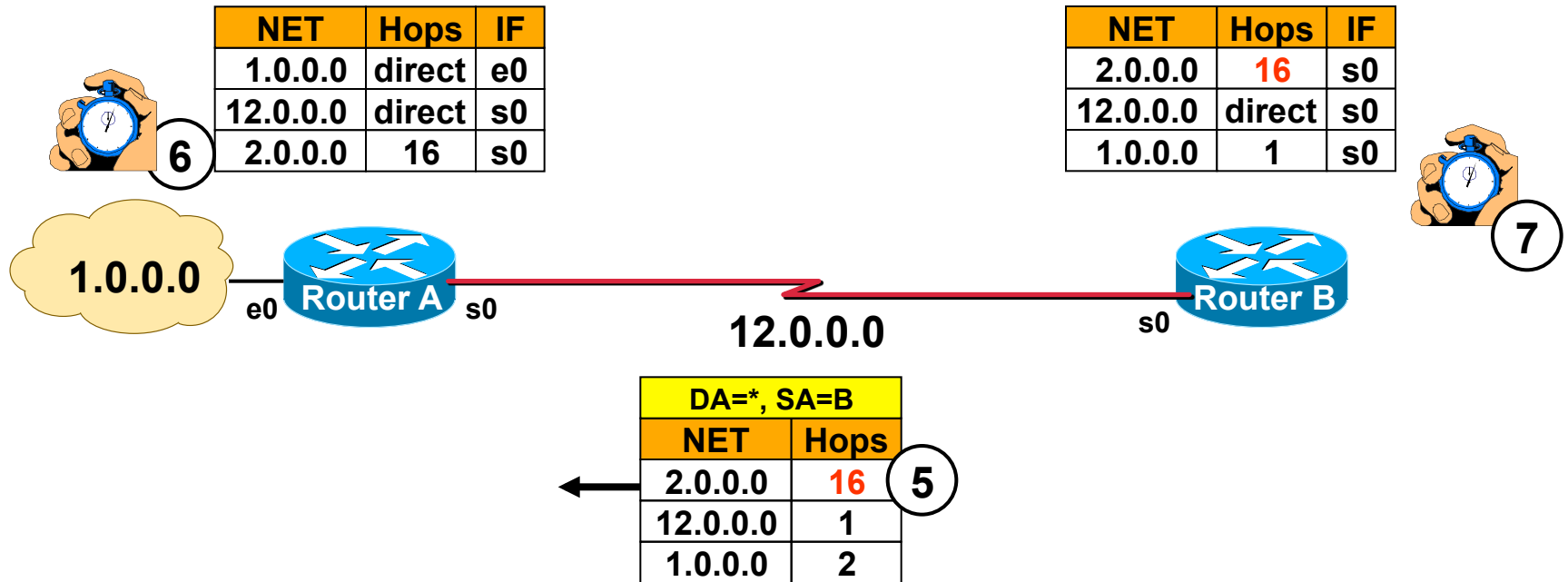
NET	Hops	IF
2.0.0.0	???	??
12.0.0.0	direct	s0
1.0.0.0	1	s0



Poison Reverse At Work (3)



Poison Reverse At Work (4)



Agenda

- **Introduction to IP Routing**
- **RIP**
 - Introduction
 - Split Horizon
 - Count-To-Infinity
 - Max-Hop-Count
 - Poison Reverse
 - Hold Down
 - Some Details and Summary
 - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

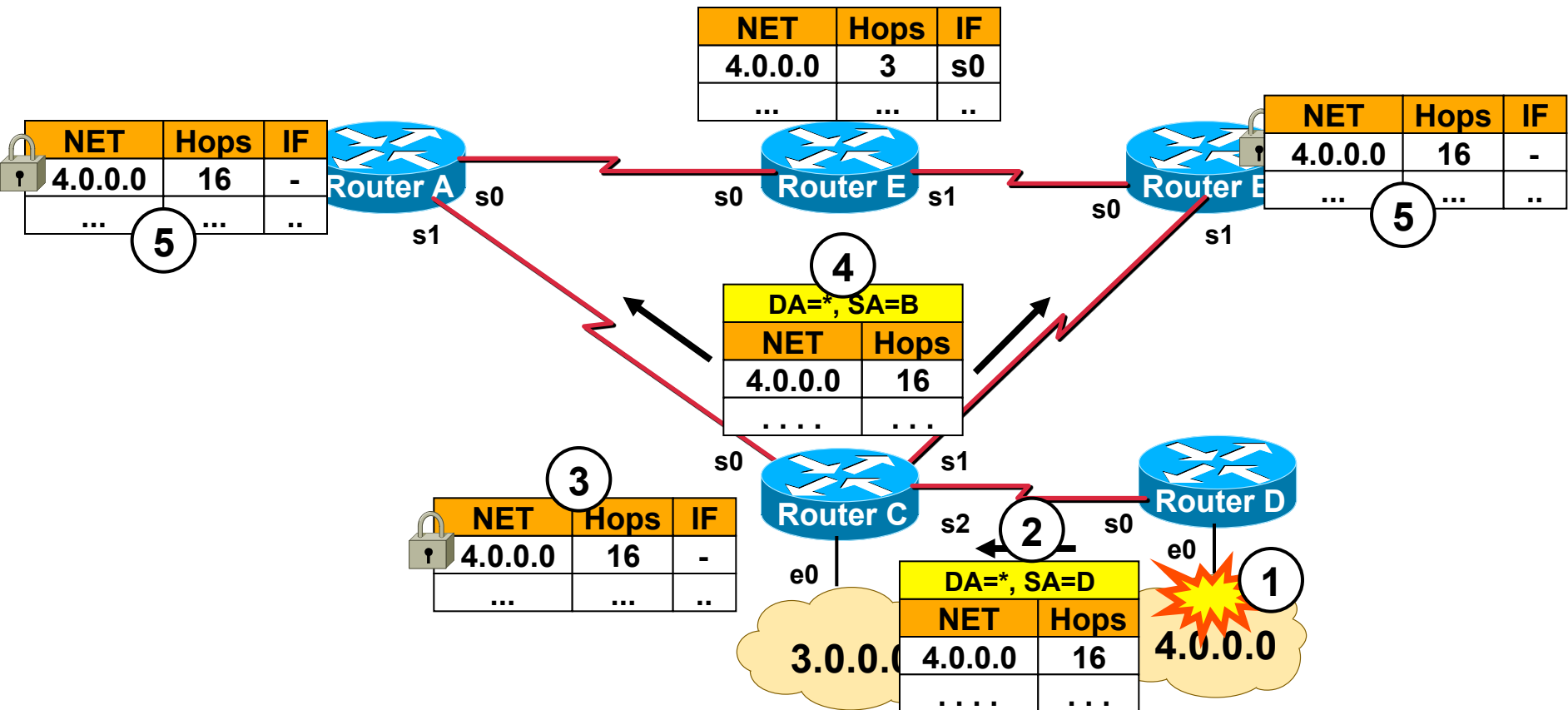
Additional Measures (2)

- **Hold Down**

- Guarantees propagation of bad news throughout the network
- Routers in hold down state ignore good news for 180 seconds
- Basic idea:
 - Network-failure message requires a specific amount of time to spread across the whole network (like a wave)
 - With Hold Down, all routers get the chance to receive the network-failure message
 - Inconsistent routing-tables and routing-loops are avoided

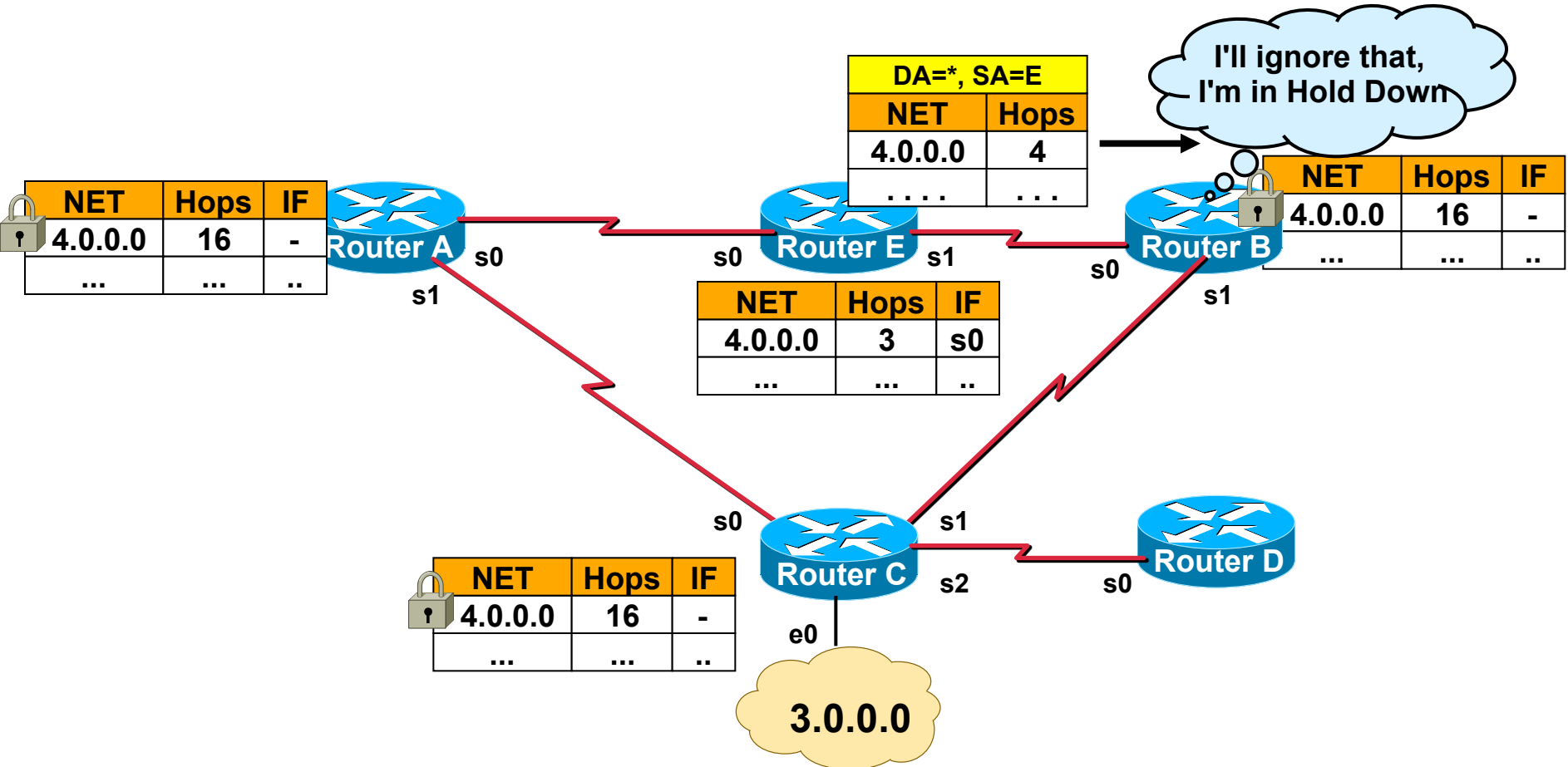
Hold Down (1)

- Router C receives unreachable message (4.0.0.0, 16) from router D
- Router C declares 4.0.0.0 as invalid (16) and enters **hold-down state**



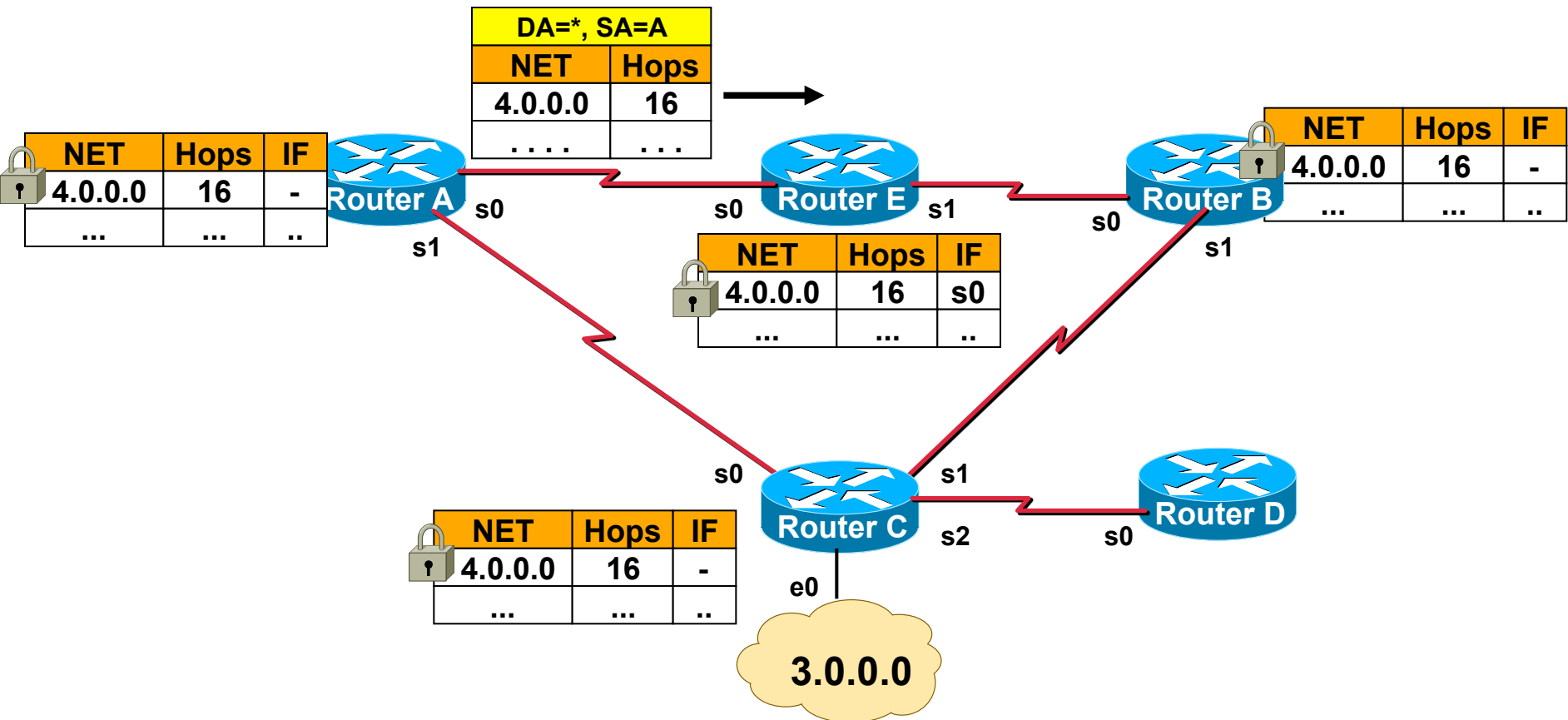
Hold Down (2)

- Information about network 4.0.0.0 with better metric is ignored for 180 seconds



Hold Down (3)

- Time enough to propagate the unreachability of network 4.0.0.0



Agenda

- **Introduction to IP Routing**
- **RIP**
 - Introduction
 - Split Horizon
 - Count-To-Infinity
 - Max-Hop-Count
 - Poison Reverse
 - Hold Down
 - Some Details and Summary **FYI**
 - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

Triggered Update / Timer Synchronization

- **To reduce convergence time, routing updates are sent immediately upon events (changes)**
 - New network connected to the router
 - Local network crashes
- **On receiving such a different routing update a router should also send immediately an update**
 - Called “Triggered Update”
- **In case of many routers on a single network**
 - Processing load might affect update timer
 - Router timers might get synchronized
 - Collisions will occur more often
- **Therefore either use**
 - External timer or add a small random time to the update timer (30 seconds + RIP_JITTER = 25...35 seconds)

RIP Timers Cisco

- **UPDATE (30 seconds)**
 - Period to send routing update
- **INVALID (180 seconds)**
 - Aging time before declaring a route invalid ("16") in the routing table
- **HOLDDOWN (180 seconds)**
 - After a route has been invalidated, how long a router will wait before accepting an update with better metric
- **FLUSH (240 seconds)**
 - Time before a non-refreshed routing table entry is removed

RIP Disadvantages

- **Big routing traffic overhead**
 - Contains nearly entire routing table
 - WAN links (!)
- **Slow convergence**
- **Small network diameter**
- **No discontinuous subnetting**
- **Only equal-cost load balancing supported**
 - (if you are lucky)

Summary RIPv1

- **First important distance vector implementation (not only for IP)**
- **Main problem: Count to infinity**
 - Maximum Hop Count
 - Split Horizon
 - Poison Reverse
 - Hold Down
- **Classful, Slow, Simple**

Agenda

- **Introduction to IP Routing**
- **RIP**
 - Introduction
 - Split Horizon
 - Count-To-Infinity
 - Max-Hop-Count
 - Poison Reverse
 - Hold Down
 - Some Details and Summary
 - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

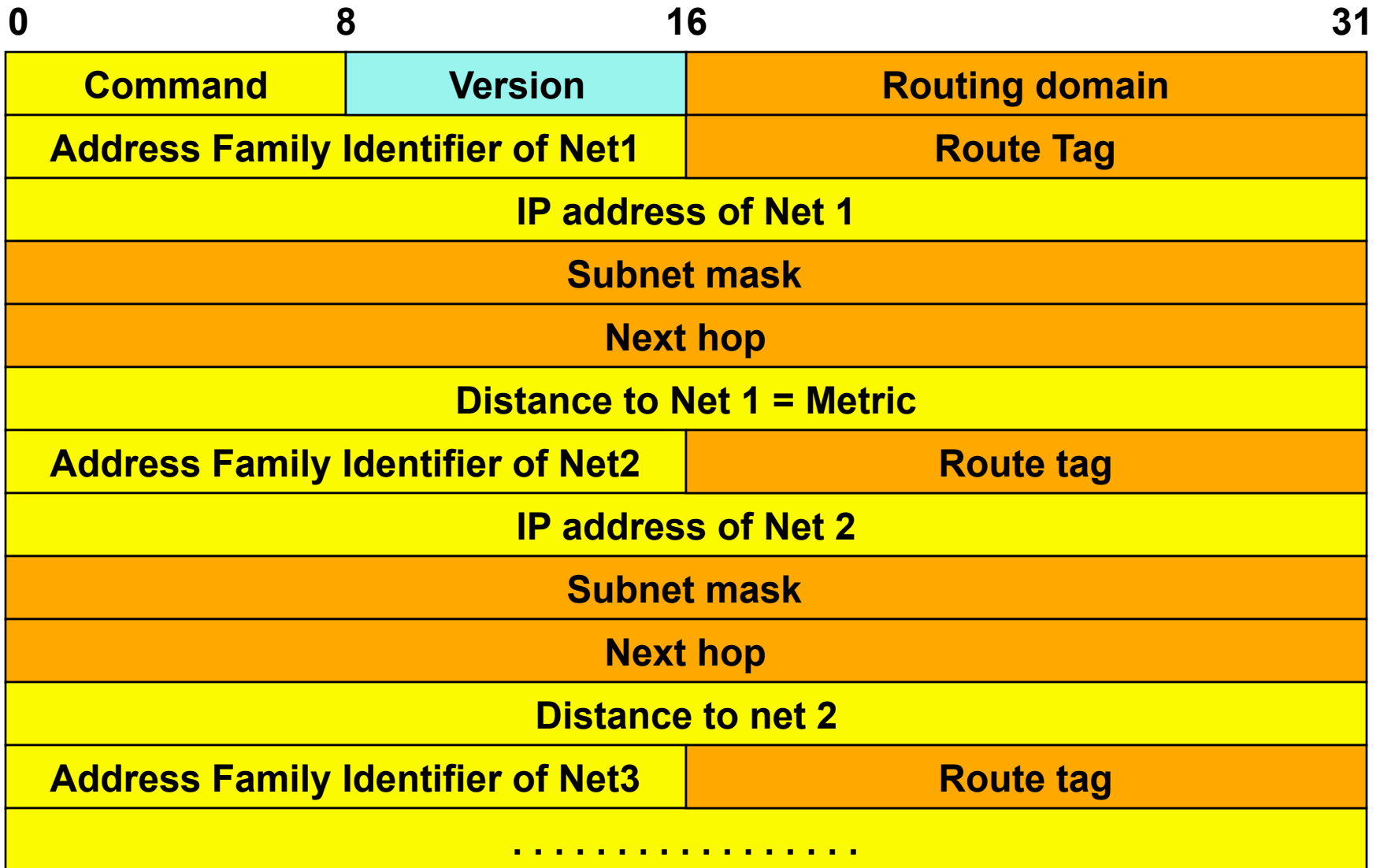
Why RIPv2?

- Need for Subnet information and VLSM
- Need for Multicast Routing Updates
 - RIPv1 used DA=255.255.255.255
 - Seen by each IP host
 - Slows down other IP stations
 - RIPv2 uses DA=224.0.0.9
 - Only RIPv2 routers will receive it
- Need for Next Hop Addresses for each route
- Need for External Route Tags

RIPv2

- **RFC 2453 specifies a new, extended RIP version:**
 - RIPv2 is RFC category “Standard”
 - RIPv1 is RFC category “Historic”
- **RIPv2 is an alternative choice to OSPF**
 - OSPF has the touch to be more complicated!
- **Several new features are supported:**
 - Transmission of subnet-masks
 - Transmission of next hop redirect information
 - Routing domains and route tags
 - Route advertisements via EGP - protocols
 - Authentication
- **RIPv2 is a **classless** routing protocol**

RIPv2 Message Format



Some Special Message Fields of RIPv2

FYI

- **Route tag**

- To distinguish between internal routes (learned via RIP) and external routes (learned from other protocols like EGP)
- Typically **AS number** is used
 - Not used by RIPv2 process
 - External routing protocols (EGP) may use the route tag to exchange information across a RIP domain

Some Special Message Fields of RIPv2

FYI

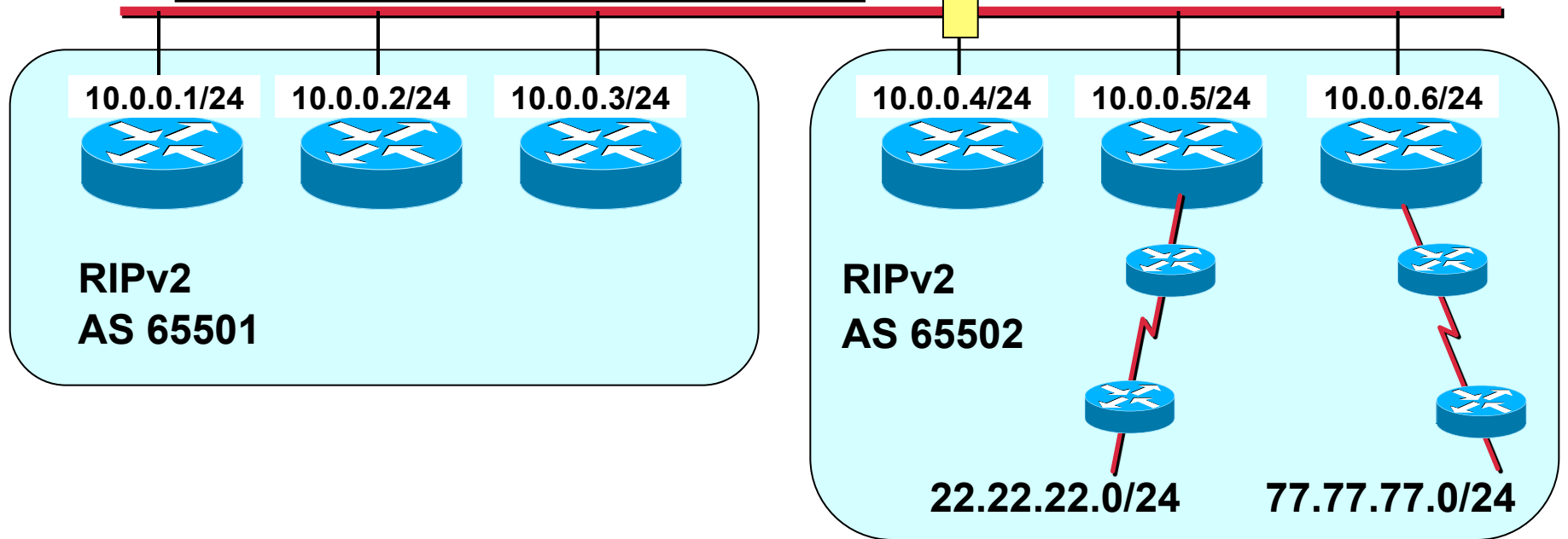
- **Next hop**

- Datagrams for the network specified in the "IP address" - field have to be redirected to that router whose IP address is specified in the "next hop" field
 - This next-hop router must be located in the same subnet as the sender of the routing-update
 - A next hop value of 0.0.0.0 indicates, that the sender-router acts as next hop itself for the given network
- Identifies a better next hop address than implicitly given by SA of the announcing router
- Especially useful on broadcast multi-access network for peering
 - Indirect routing on a broadcast segment would be ...silly.

Next Hop and Route Tag

FYI

2	2	
2		65502
22.22.22.0		
255.255.255.0		
10.0.0.5		
	1	
2		65502
77.77.77.0		
255.255.255.0		
10.0.0.6		
	3	



Authentication

- **Hackers might send invalid routing updates**
- **RIPv2 introduces password protection as authentication**
- **Initially only 16 plaintext characters (!)**
 - Authentication type 2
- **RFC 2082 proposes keyed MD-5 authentication**
 - Authentication type 2
 - Multiple keys can be defined
 - Updates contain a key-id and an unsigned 32 bit sequence number to prevent replay attacks
- **Cisco IOS supports**
 - MD5 authentication (Type 3, 128 bit hash)

Authentication

Command	Version	Unused or Routing Domain
0xFFFF		Authentication Type
Password		
Password		
Password		
Password		
Address Family Identifier	Route Tag	
IP Address		
Subnet Mask		
Next Hop		
Metric		
.....		

Up to 24 route entries

- **Cisco's implementation offers key chains**
 - Multiple keys (MD5 or plaintext)
 - Each key is assigned a lifetime (date, time and duration)
- **Can be used for migration**
 - Key management should rely on Network Time Protocol (NTP)

RIPv1 Inheritance

- **All timers are the same**
 - UPDATE
 - INVALID
 - HOLDDOWN
 - FLUSH
- **Same convergence protections**
 - Split Horizon
 - Poison Reverse
 - Hold Down
 - Maximum Hop Count (also 16 !!!)
- **Same UDP port 520**
 - Also maximum 25 routes per update
 - Equally 512 Byte payloads

RIPv1 Compatibility

FYI

- **RIPv1 Compatibility Mode**

- RIPv2 router uses broadcast addresses
- RIPv1 routers will ignore header extensions
- RIPv2 performs route summarization on address class boundaries
 - Disable: `(config-router)# no auto-summary`

- **RIPv1 Mode**

- RIPv2 sends RIPv1 messages

- **RIPv2 Mode**

- Send genuine RIPv2 messages

RIPv2 Summary

- **Most important: RIPv2 is classless**
 - Subnet masks are carried for each route
- **Multicasts and next hop field increase performance**
- **But still not powerful enough for large networks**

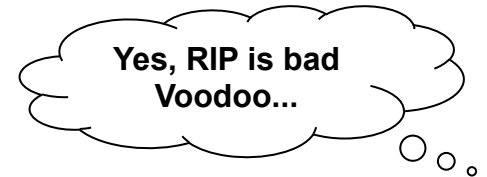
Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
 - Introduction
 - The Dijkstra Algorithm
 - Communication Procedures
 - LSA Broadcast Handling
 - Split Area
 - Broadcast Networks
 - Area Principles
 - Stub Areas
 - Route Summarization
 - Virtual Link
 - Summary
 - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

Open Shortest Path First

- **Official (IETF) successor of RIP**

- RIP is slow
- RIP is unreliable
- RIP produces too much routing traffic
- RIP only allows 15 hop routes



- **OSPF is a link-state routing protocol**

- “Open” means “not proprietary”
- Inherently fast convergence
- Designed for large networks
- Designed to be reliable

- **OSPF's father: John Moy**

- Version 1: RFC 1131
- Version 2: RFC 2328 (244 pages !!!)
 - V2 first released in RFC 1583 obsoleted by RFC 2178

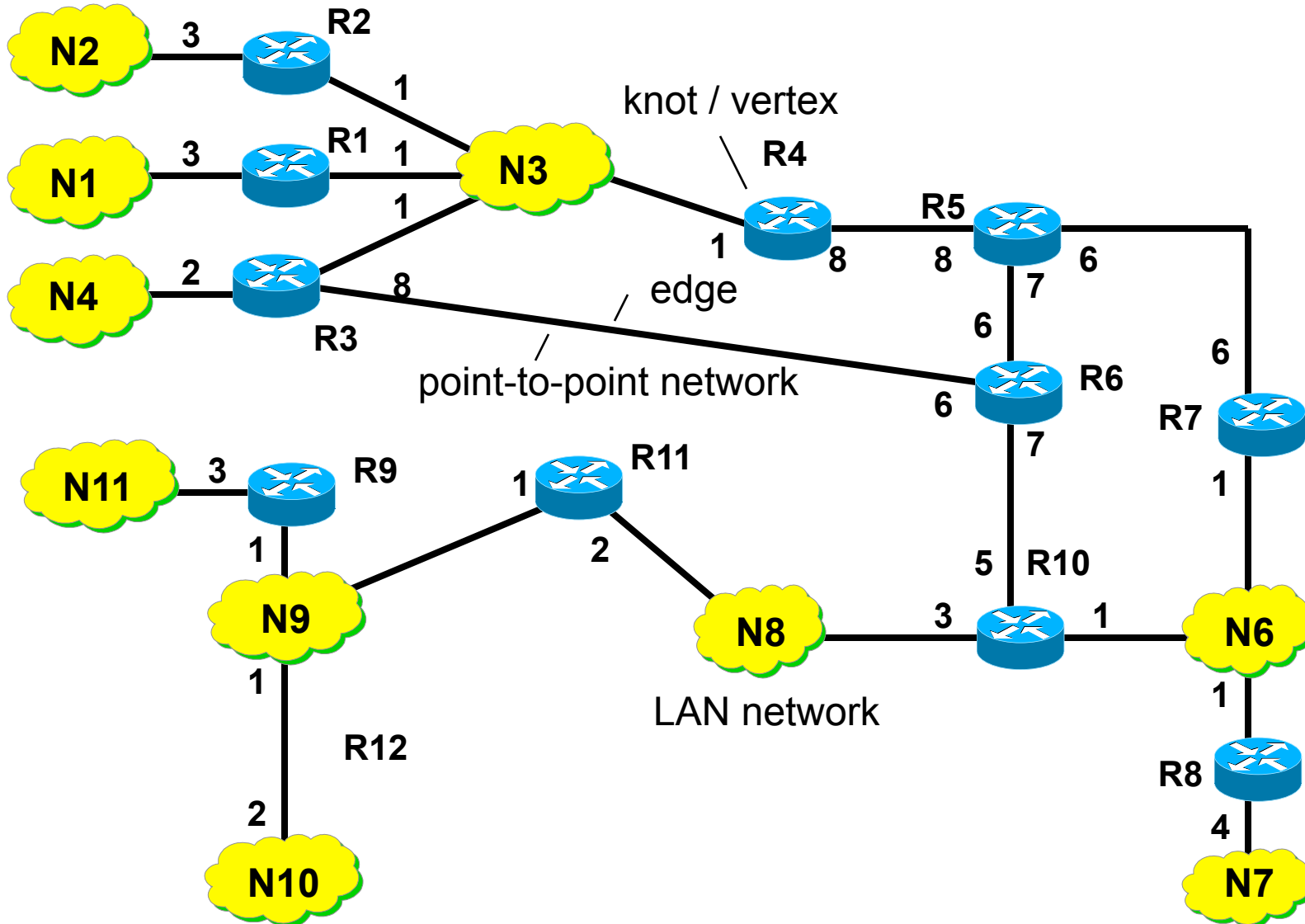
OSPF Base Principles

- Every router knows topology of the whole network including subnets and routers
 - “Roadmap”
- Topology (roadmap) stored in router’s OSPF database
- Shortest Path First (SPF) algorithm applied to find the best path
 - Invented by E. W. Dijkstra
 - Creates a (loop-free) tree with local router as source
 - Is used to find the best path by calculating very efficiently all paths to all destinations at once; best path is entered into the routing table
- Changes are flooded over the network to update the OSPF database
 - Like traffic announcements used by car navigation systems
 - LSA (Link State Advertisements)

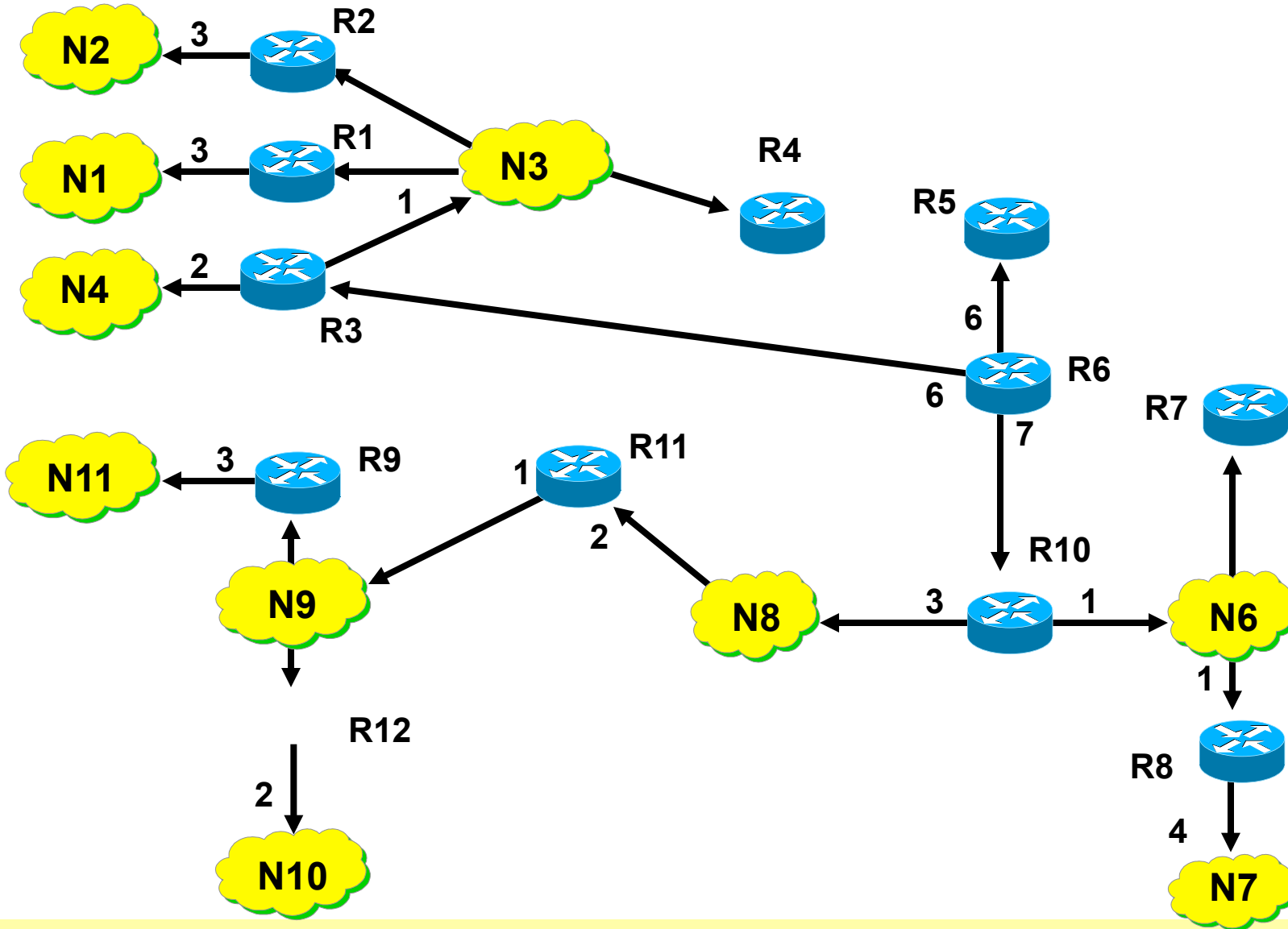
OSPF Topology Database

- **Every router maintains a topology database**
 - Like a "network roadmap"
 - Describes the whole network !!
 - Note: RIP provides only "signposts"
- **Database is based on a graph**
 - Where each knot (vertex) stands for a router
 - Where each edge stands for a subnet
 - Connecting the routers
 - Path-costs are assigned to the edges
- **Router uses the graph**
 - To calculate shortest paths to all subnets
 - Router itself is the root of the shortest path

OSPF Domain



Shortest Paths regarding Router R6



Routing Table Router 6

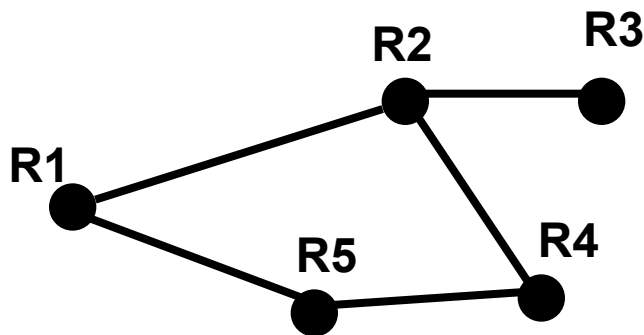
NET-ID	NEXT HOP	DISTANCE
N1	R3	10
N2	R3	10
N3	R3	7
N4	R3	8
N6	R10	8
N7	R10	12
N8	R10	10
N9	R10	11
N10	R10	13
N11	R10	14

OSPF Ideas

- **Metric: "Cost" = $10^8/\text{BW}$ (in bit/s)**
 - Therefore easily configurable per interface
- **OSPF routers exchange real topology information**
 - Stored in dedicated topology databases
- **Now routers have a "roadmap"**
 - Instead of signposts (RIP)
- **Incremental updates**
 - NO updates when there is NO topology change
- **Fast convergence**
 - Almost no routing traffic in absence of topology changes

What is Topology Information?

- The smallest topological unit is simply the information element **ROUTER-LINK-ROUTER**
- So the question is: Which router is linked to which other routers?
- Link-state



==

Link Database:

R1– R2
R1– R5
R2– R3
R2– R4
R4– R5

The Link Database
exactly describes
the roadmap

OSPF Routing Updates / LSA

- **The routing updates are actually link state updates**
 - Parts of link state database are exchanged
 - Instead of parts of routing table (RIP)
 - Link **S**tate **A**dvertisement (LSA)
- **Applying the SPF algorithm on the link state database**
 - Each router can create routing table entries by its own
- **LSAs are carried**
 - In small packets, forwarded by each router without much modifications through the whole OSPF domain (area)
 - Flooding principle
- **Much faster than RIP updates**
 - RIP must receive, examine, create, and send
- **Convergence time**
 - Detection time + LSA flooding + 5 seconds before computing the topology table = "a few seconds"

OSPF Areas – OSPF Performance

- **Large networks: "Divide and conquer" into areas**
 - LSA-procedures inside each area
 - But *distance-vector updates between areas*
- **Additional complexity because of performance optimizations**
 - Limit number of adjacencies in a multi-access network OSPF
 - Limit scope of flooding through "Areas"
 - Deal with stub areas efficiently
 - Learn external routes efficiently
 - Realized through different LSA types

Agenda


- **Introduction to IP Routing**
- **RIP**
- **OSPF**
 - Introduction
 - The Dijkstra Algorithm **FYI**
 - Communication Procedures
 - LSA Broadcast Handling
 - Split Area
 - Broadcast Networks
 - Area Principles
 - Stub Areas
 - Route Summarization
 - Virtual Link
 - Summary
 - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

About E. W. Dijkstra


- **Born in 1930 in Rotterdam**
- **Degrees in mathematics and theoretical physics from the University of Leyden and a Ph.D. in computing science from the University of Amsterdam**
 - Programmer at the Mathematisch Centrum, Amsterdam, 1952-62
 - Professor of mathematics, Eindhoven University of Technology, 1962-1984
 - Burroughs Corporation research fellow, 1973-1984
 - Schlumberger Centennial Chair in Computing Sciences at the University of Texas at Austin, 1984-1999
 - Retired as Professor Emeritus in 1999
 - 1972 recipient of the ACM Turing Award, often viewed as the Nobel Prize for computing
- **Died 6 August 2002**



**Edsger W. Dijkstra
(1930-2002)**



*“The question of whether
computers can think is
like the question of whether
submarines can swim”*



Edsger Wybe Dijkstra

Dijkstra's SP Algorithm

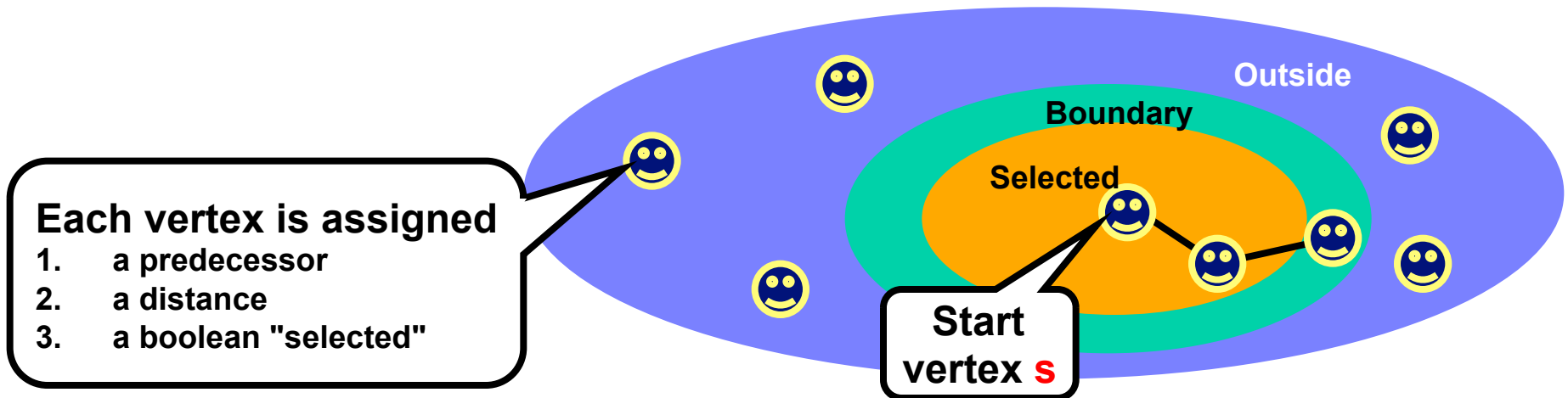
- **Famous paper "A note on two problems in connection with graphs" (1959)**
- **Single source SP problem in a directed graph**
- **Important applications include**
 - Network routing protocols (OSPF, IS-IS)
 - Traveler's route planner

Terms

- **Graph $G(V,E)$ consists of vertices V and edges E**
- **Edges are assigned costs c**
- **"Length" of graph $c(G) = \text{sum of all costs}$**
 - Assumed to be positive ("Distance Graph")
- **"Distance" between two vertices $d(v,v') = \min\{c(p)\}$, $p...$
path**
 - Can be infinite
- **p with $c(p) = d(v,v')$ is called shortest path $sp(v,v')$**

Definitions

- **Select start vertex s**
- **Three sets of vertices:**
 - **Selected** (sp already calculated)
 - **Boundary** (currently subject of calculation)
 - **Outside** (not yet examined)



The Algorithm

Initialize Vertices

v.predecessor = none
v.distance = ∞
v.selected = false

Select S

s.predecessor = s
s.distance = 0
s.selected = true

Add neighbors of S to boundary

Select V with lowest distance from boundary

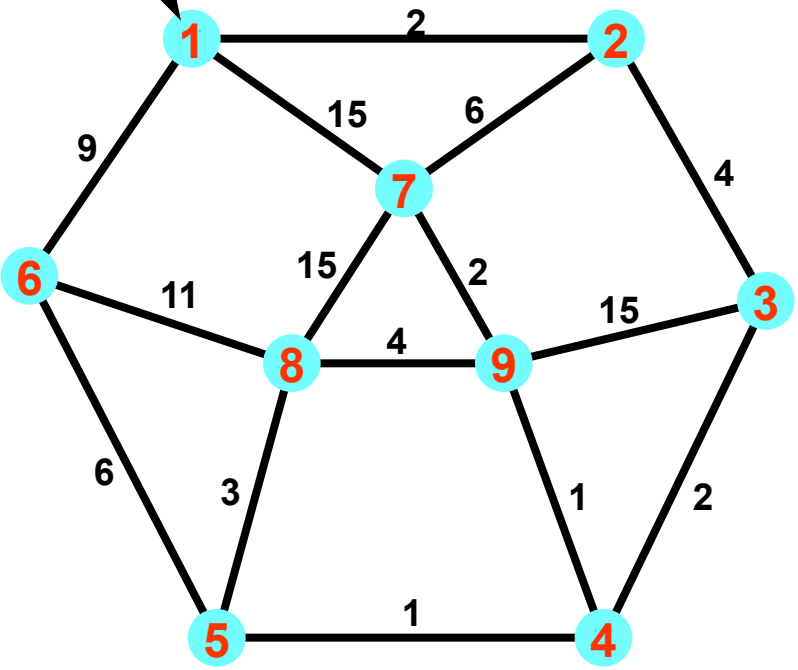
Add neighbors of V to boundary

For these neighbors calculate distance using V as predecessor
Previous vertices might get better total distance



Example 1

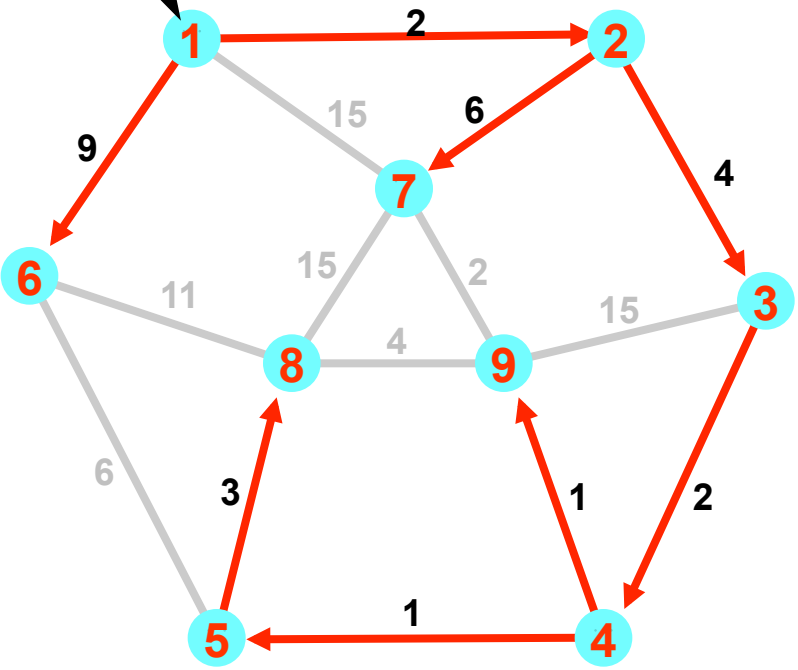
Start vertex **s**



Selected	vertex number	distance	predecessor
1	0	1	
2	2	1	
3	6	2	
7	8	2	
4	8	3	
6	9	1	
9	9	4	
5	9	4	
8	12	5	
2	2	1	
6	9	1	
7	8	2	
6	9	1	
6	9	1	
9	9	4	
5	9	4	
8	12	5	
6	9	1	
7	8	2	
4	8	3	
8	23	7	
5	9	4	
8	13	9	
7	15	1	
3	6	2	
4	8	3	
9	10	7	
9	9	4	
8	20	6	
9	21	3	
8	23	7	
5	9	4	

Result

Start vertex **s**



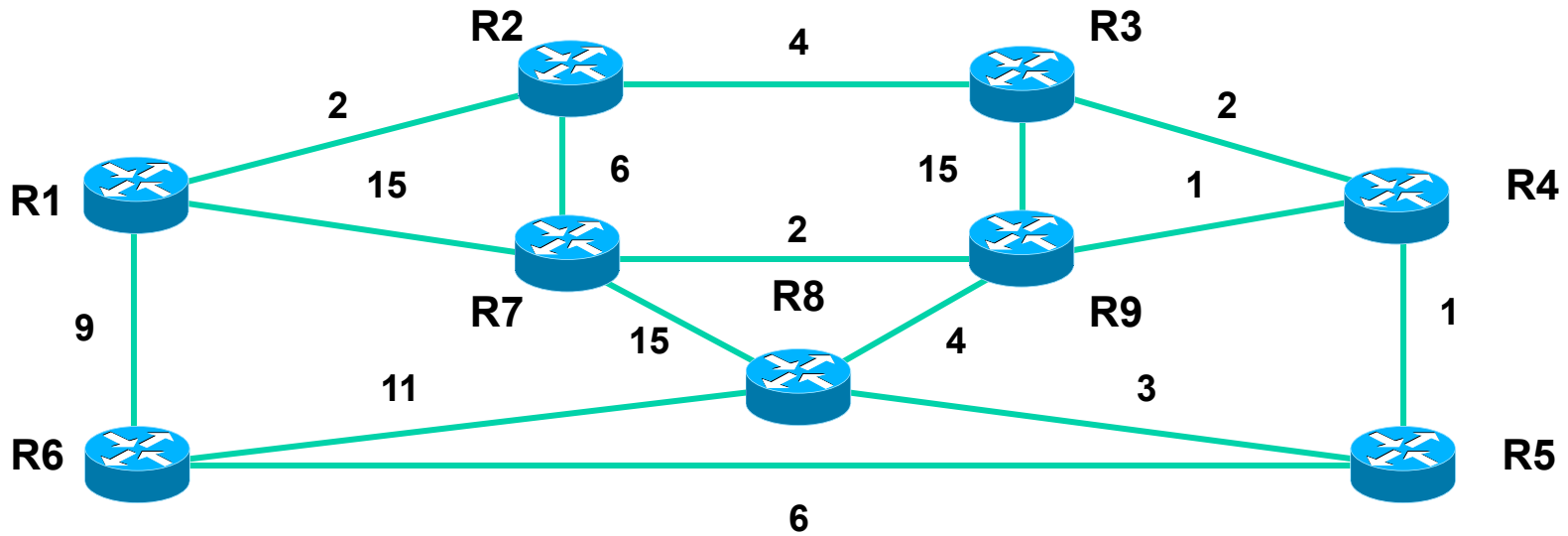
Selected		
1	0	1
2	2	1
3	6	2
7	8	2
4	8	3
6	9	1
9	9	4
5	9	4
8	12	5

- Single source SP
- Minimal length
- Complete

Performance

- **Greedy algorithm**
- **Most critical: Implementation of boundary data structure**
 - No explicit structure: $O(|V|^2)$
 - Fibonacci heap: $O(|E| + |V| \log |V|)$
- **Alternatives**
 - Bellman-Ford (RIP) algorithm
 - Floyd-Warshall algorithm
 - A* algorithm
 - Extends SPF with a estimation function to enhance performance in certain situations

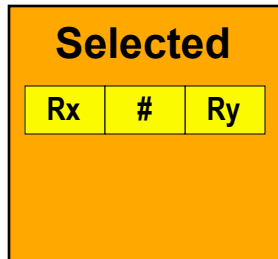
Example 2 for Dijkstra Algorithm in Action



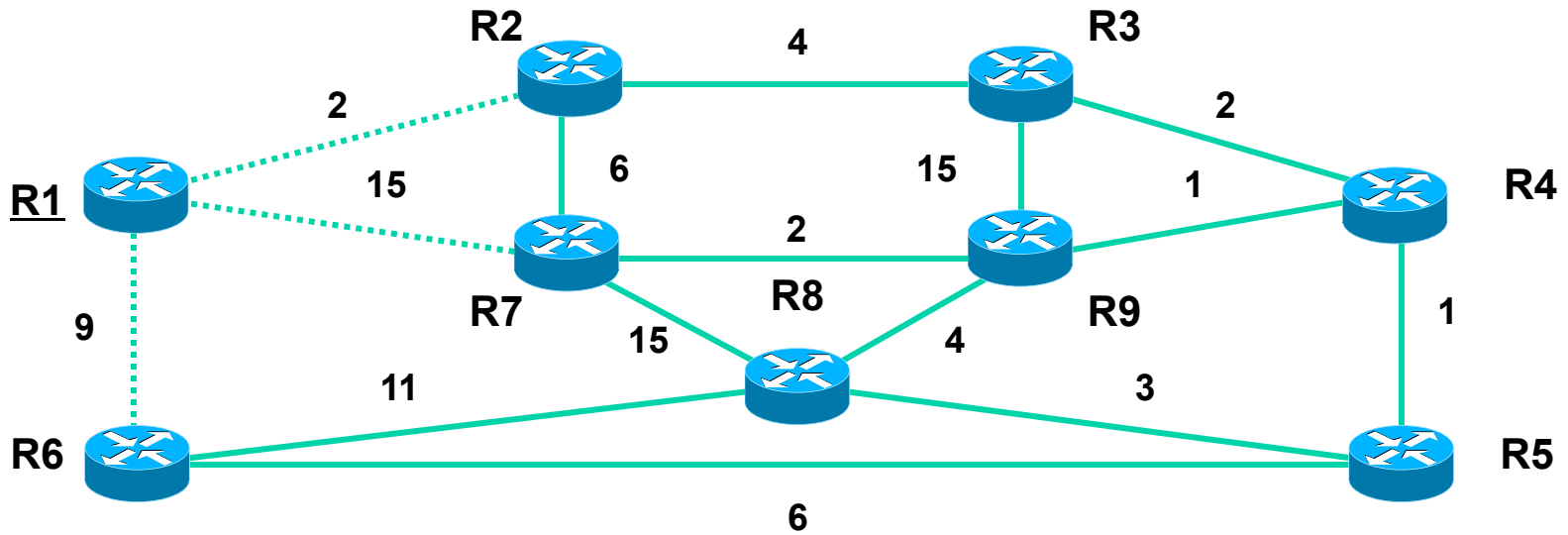
Summary Cost (Distance)

Router-Name (Vertex Number)

Router-Name of Predecessor



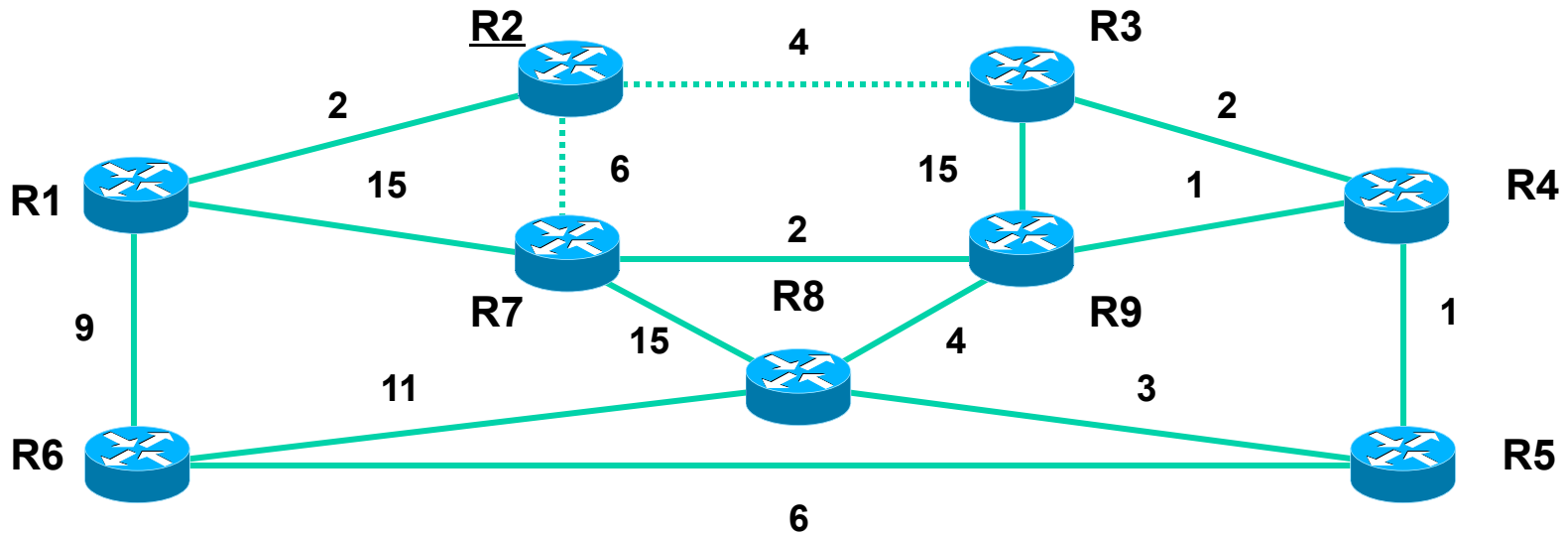
Select root (R1)



Selected		
R1	0	R1

Boundary								
R2	2	R1	R6	9	R1	R7	15	R1

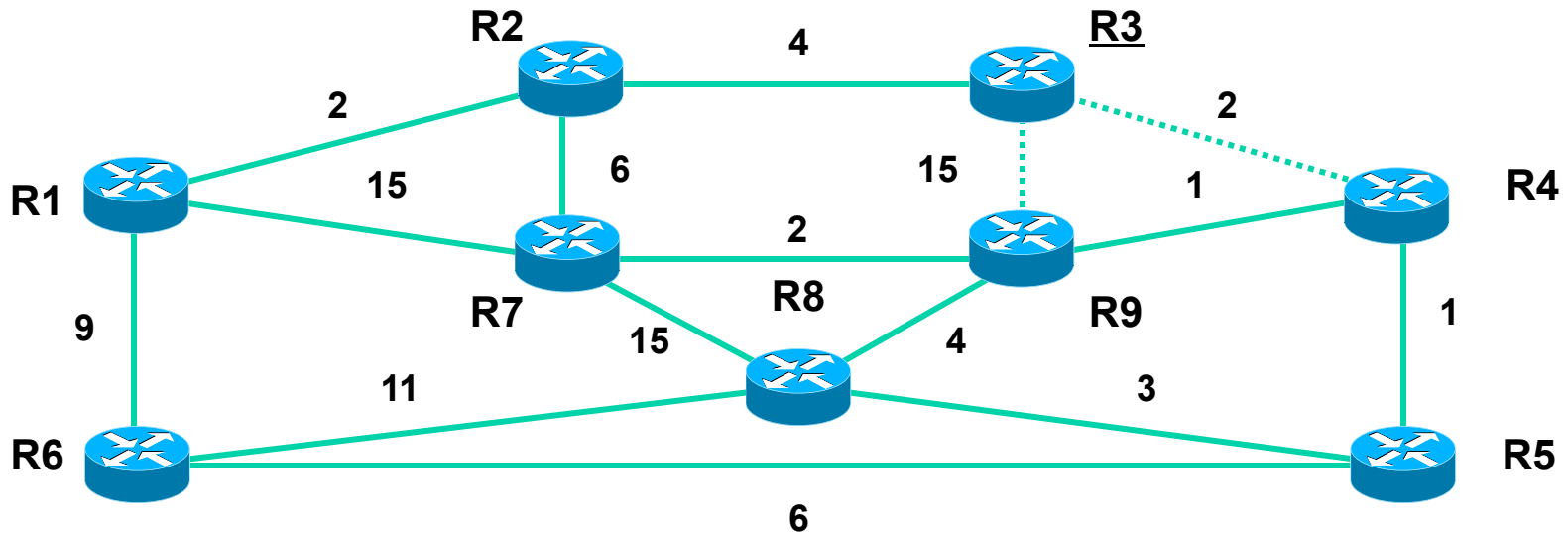
Select router with lowest cost in boundary (R2), calculate cost for neighbours R3, R7



Selected		
R1	0	R1
R2	2	R1

Boundary								
R2	2	R1	R6	9	R1	R7	15	R1
R6	9	R1	R7	8	R2	R3	6	R2

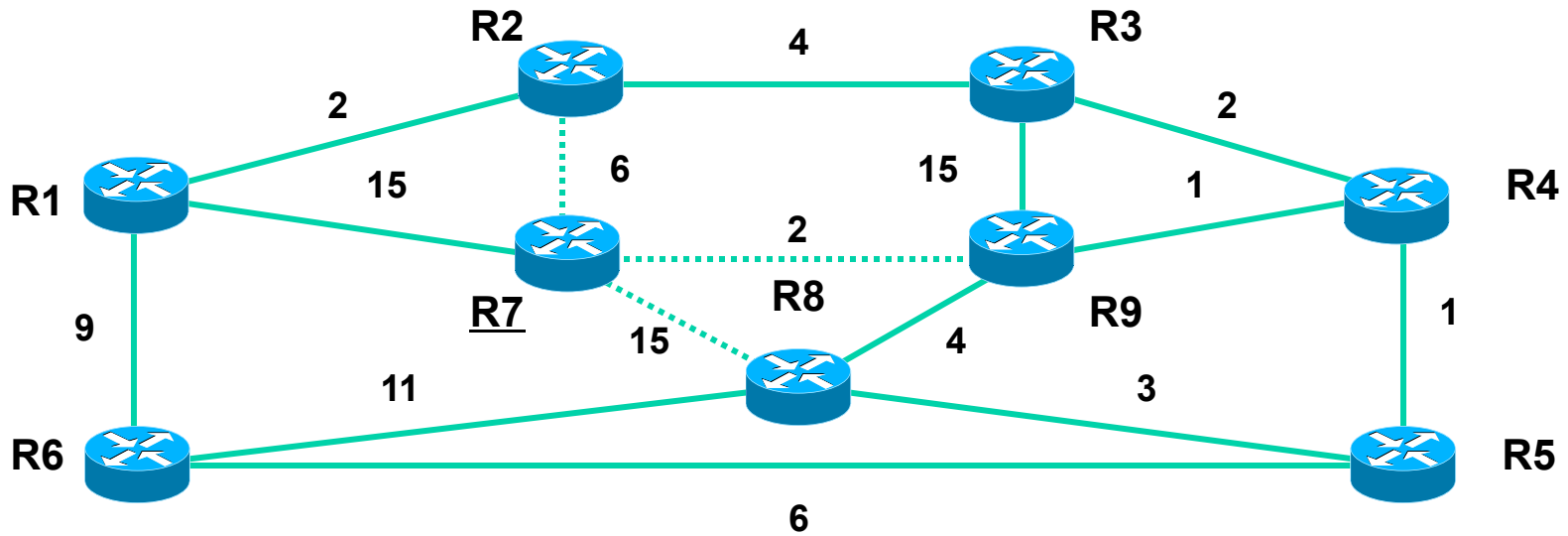
Select router with lowest cost in boundary (R3), calculate cost for neighbours R9, R4



Selected		
R1	0	R1
R2	2	R1
R3	6	R2

Boundary								
R2	2	R1	R6	9	R1	R7	15	R1
R6	9	R1	R7	8	R2	R3	6	R2
R6	9	R1	R7	8	R2	R9	21	R3
						R4	8	R3

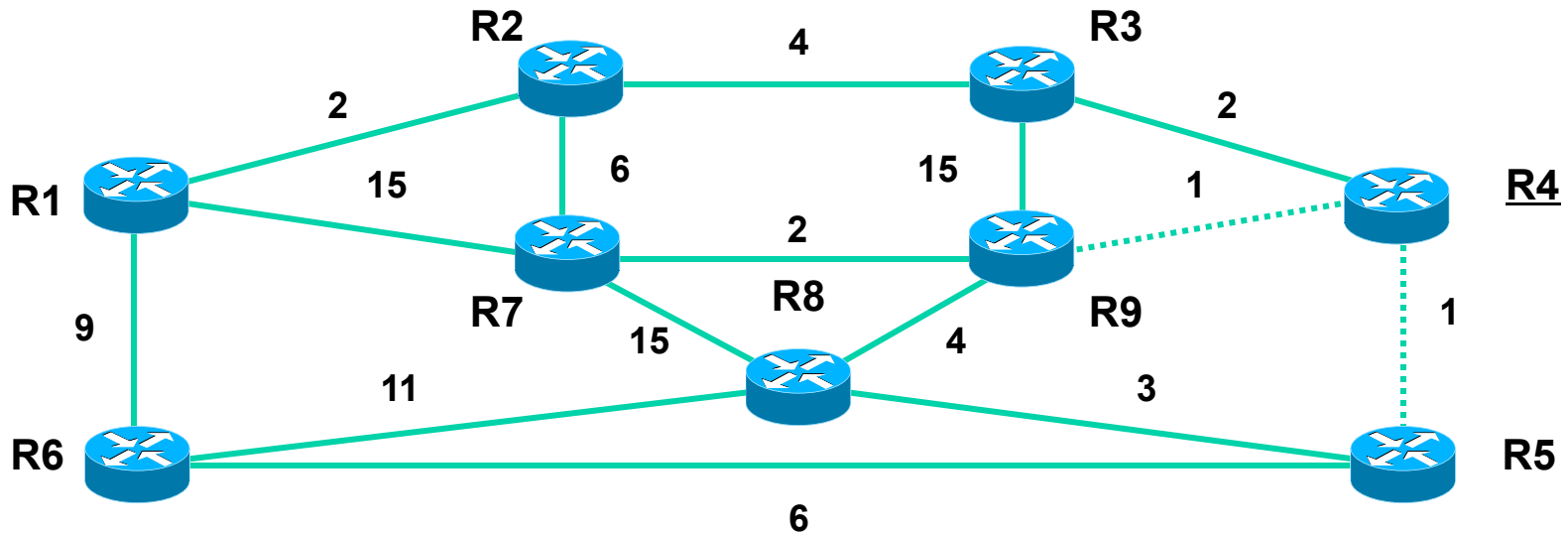
Select one router with lowest cost in boundary (R7), calculate cost for neighbours R8, R9



Selected		
R1	0	R1
R2	2	R1
R3	6	R2
R7	8	R2

Boundary								
R2	2	R1	R6	9	R1	R7	15	R1
R6	9	R1	R7	8	R2	R3	6	R2
R6	9	R1	R7	8	R2	R9	21	R3
R6	9	R1	R4	8	R3	R9	10	R7
			R8	23	R7			

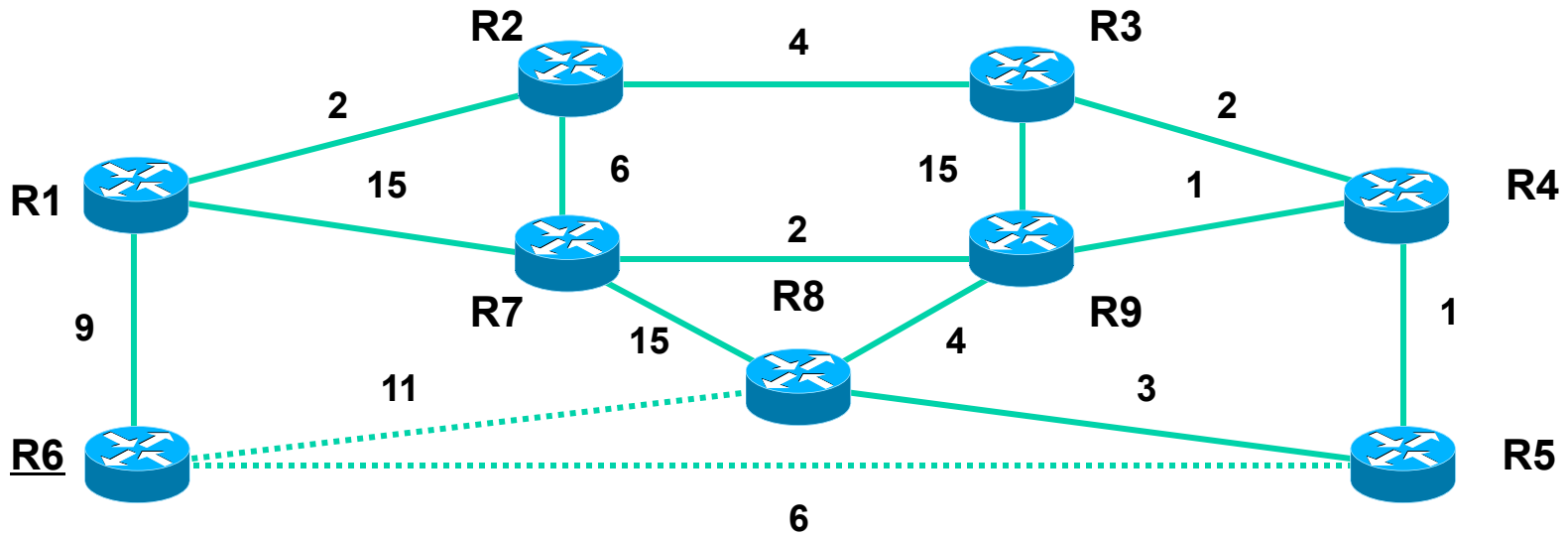
Select router with lowest cost in boundary (R4), calculate cost for neighbours R9, R5



Selected		
R1	0	R1
R2	2	R1
R3	6	R2
R7	8	R2
R4	8	R3

Boundary								
R2	2	R1	R6	9	R1	R7	15	R1
R6	9	R1	R7	8	R2	R3	6	R2
R6	9	R1	R7	8	R2	R9	21	R3
R6	9	R1	R4	8	R3	R9	10	R7
R8	23	R7	R8	23	R7	R9	9	R4
R5	9	R4	R5	9	R4	R8	23	R7
R5	9	R4						

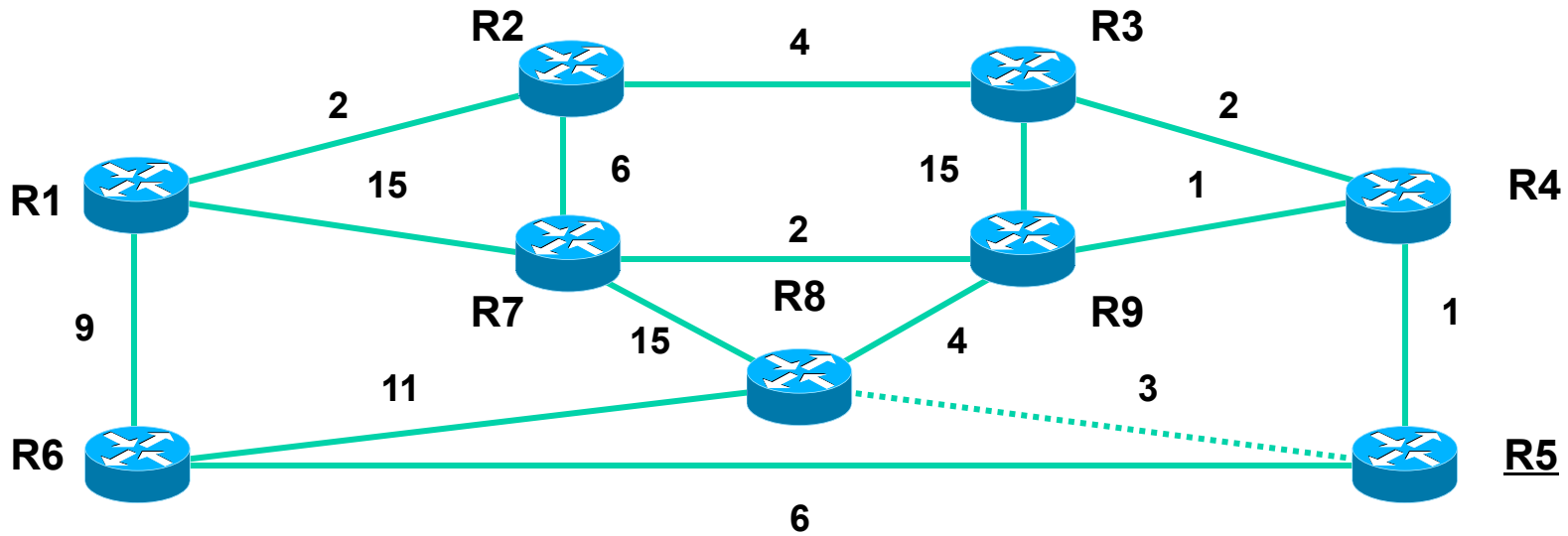
Select one router with lowest cost in boundary (R6), calculate cost for neighbours R5 and R8



Selected		
R1	0	R1
R2	2	R1
R3	6	R2
R7	8	R2
R4	8	R3
R6	9	R1

Boundary											
R2	2	R1	R6	9	R1	R7	15	R1			
R6	9	R1	R7	8	R2	R3	6	R2			
R6	9	R1	R7	8	R2	R9	21	R3	R4	8	R3
R6	9	R1	R4	8	R3	R9	10	R7	R8	23	R7
R6	9	R1	R8	23	R7	R9	9	R4	R5	9	R4
R9	9	R4	R8	20	R6	R5	9	R4			

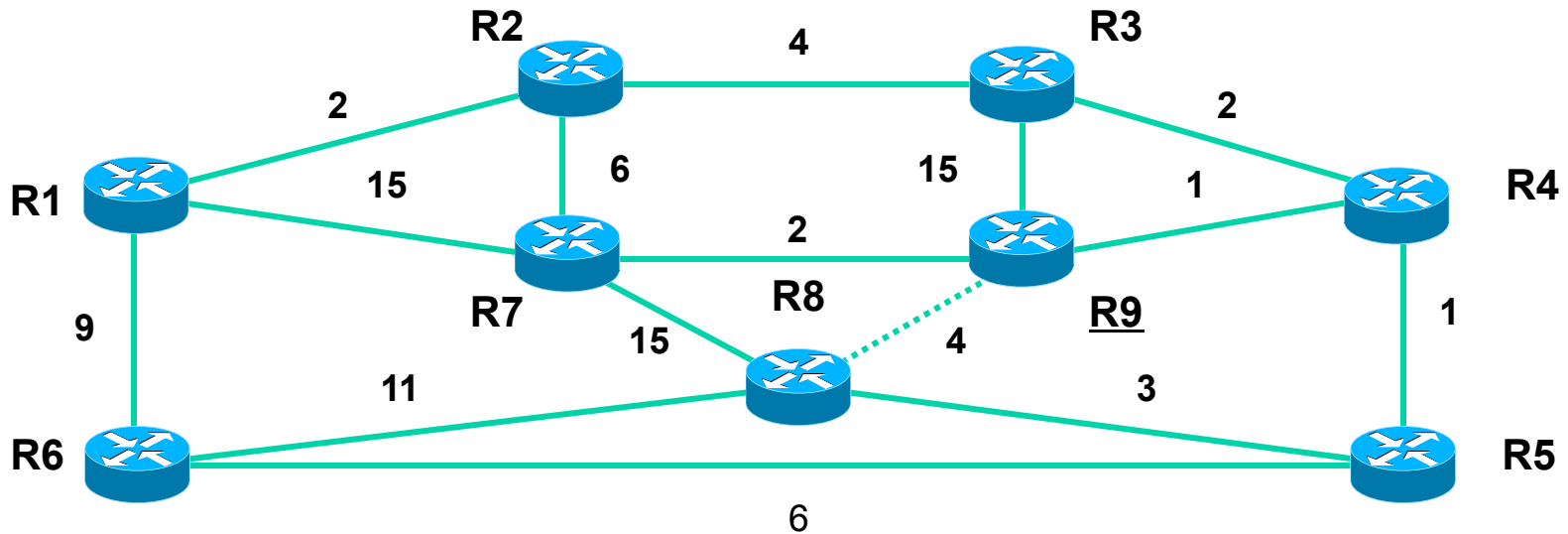
Select one neighbour with lowest cost in boundary (R5), calculate cost for neighbour R8



Selected		
R1	0	R1
R2	2	R1
R3	6	R2
R7	8	R2
R4	8	R3
R6	9	R1
R5	9	R4

Boundary											
R2	2	R1	R6	9	R1	R7	15	R1			
R6	9	R1	R7	8	R2	R3	6	R2			
R6	9	R1	R7	8	R2	R9	21	R3	R4	8	R3
R6	9	R1	R4	8	R3	R9	10	R7	R8	23	R7
R6	9	R1	R8	23	R7	R9	9	R4	R5	9	R4
R9	9	R4	R8	20	R6	R5	9	R4			
R9	9	R4	R8	12	R5						

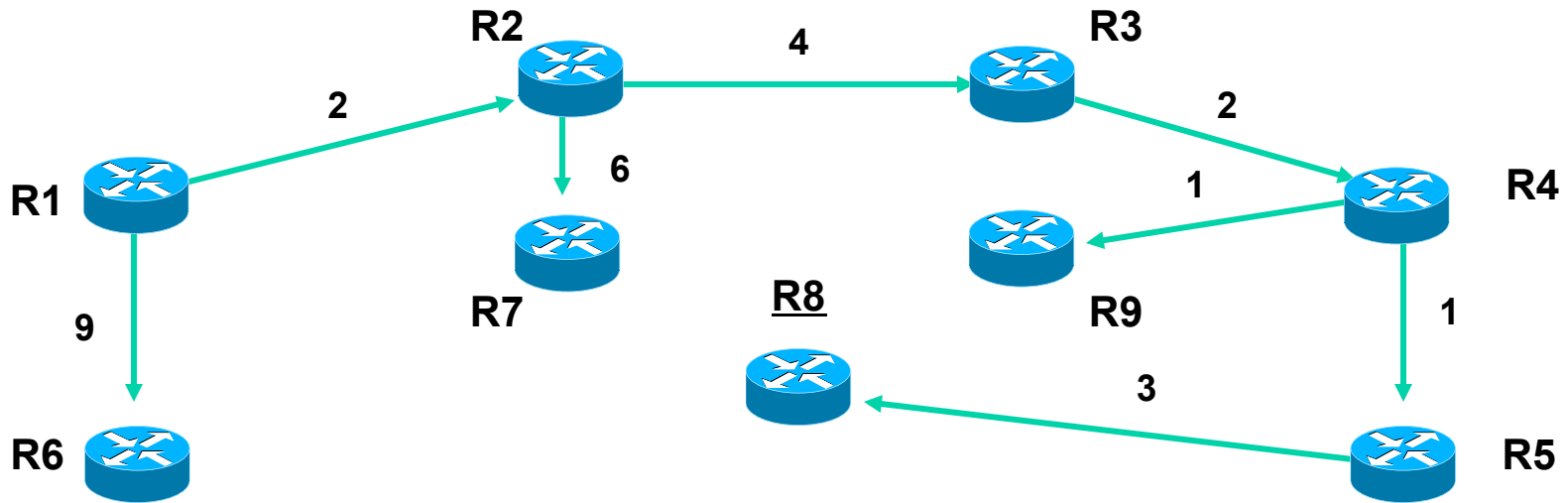
Select router with lowest cost in boundary (R9), calculate cost for neighbours R8



Selected		
R1	0	R1
R2	2	R1
R3	6	R2
R7	8	R2
R4	8	R3
R6	9	R1
R5	9	R4
R9	9	R4

Boundary								
R2	2	R1	R6	9	R1	R7	15	R1
R6	9	R1	R7	8	R2	R3	6	R2
R6	9	R1	R7	8	R2	R9	21	R3
R4	8	R3	R8	23	R7	R8	23	R7
R6	9	R1	R4	8	R3	R9	10	R7
R6	9	R1	R8	23	R7	R9	9	R4
R9	9	R4	R8	20	R6	R5	9	R4
R9	9	R4	R8	12	R5			
R8	12	R5						

Select last router in boundary (R8), algorithm terminated, all shortest paths found



Selected		
R1	0	R1
R2	2	R1
R3	6	R2
R7	8	R2
R4	8	R3
R6	9	R1
R5	9	R4
R9	9	R4
R8	12	R5

Boundary											
R2	2	R1	R6	9	R1	R7	15	R1			
R6	9	R1	R7	8	R2	R3	6	R2			
R6	9	R1	R7	8	R2	R9	21	R3	R4	8	R3
R6	9	R1	R4	8	R3	R9	10	R7	R8	23	R7
R6	9	R1	R8	23	R7	R9	9	R4	R5	9	R4
R9	9	R4	R8	20	R6	R5	9	R4			
R9	9	R4	R8	12	R5						
R8	12	R5									

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
 - Introduction
 - The Dijkstra Algorithm
 - Communication Procedures
 - LSA Broadcast Handling
 - Split Area
 - Broadcast Networks
 - Area Principles
 - Stub Areas
 - Route Summarization
 - Virtual Link
 - Summary
 - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

Creating the Database

- The basic means for creating and maintaining the database are the so-called
Link States
- A link state stands for an intact (synchronized) local neighbourhood between two routers
 - The link state is created by these two routers
 - Other routers are notified about this link state via a special broadcast-mechanism ("traffic-news")
 - Flooding together with sequence numbers stored in topology database
 - Link states are verified continuously

How are Link States used?

- **Adjacent routers declare themselves as neighbours by setting the link state up (or down otherwise)**
 - The link-state can be checked with hello messages
 - Note: Link state down is not explicitly expressed, it is just the absence of the link to the former neighbour in the LSA announcement
- **Every link state change is published to all routers of the OSPF domain using Link State Advertisements (LSAs)**
 - Is a broadcast mechanism
 - LSAs are much shorter than routing tables
 - Because LSAs contain only the actual changes
 - That's why distance vector protocols are much slower
 - Whole topology map relies on correct generation and delivery of LSAs
 - Synchronization of a distributed database !!!

OSPF Communication Principle 1

- **OSPF messages are transported by IP**
 - ip protocol number 89
- **During initialization a router sends hello-messages to all directly reachable routers**
 - To determine its neighbourhood
 - Can be done automatically in broadcast networks and point-to-point connections by using the IP multicast-address 224.0.0.5 (all OSPF routers)
 - Non-broadcast networks: configuration of the neighbourhood-routers is required (e.g. X25)
- **This router also receives hello-messages from other routers**

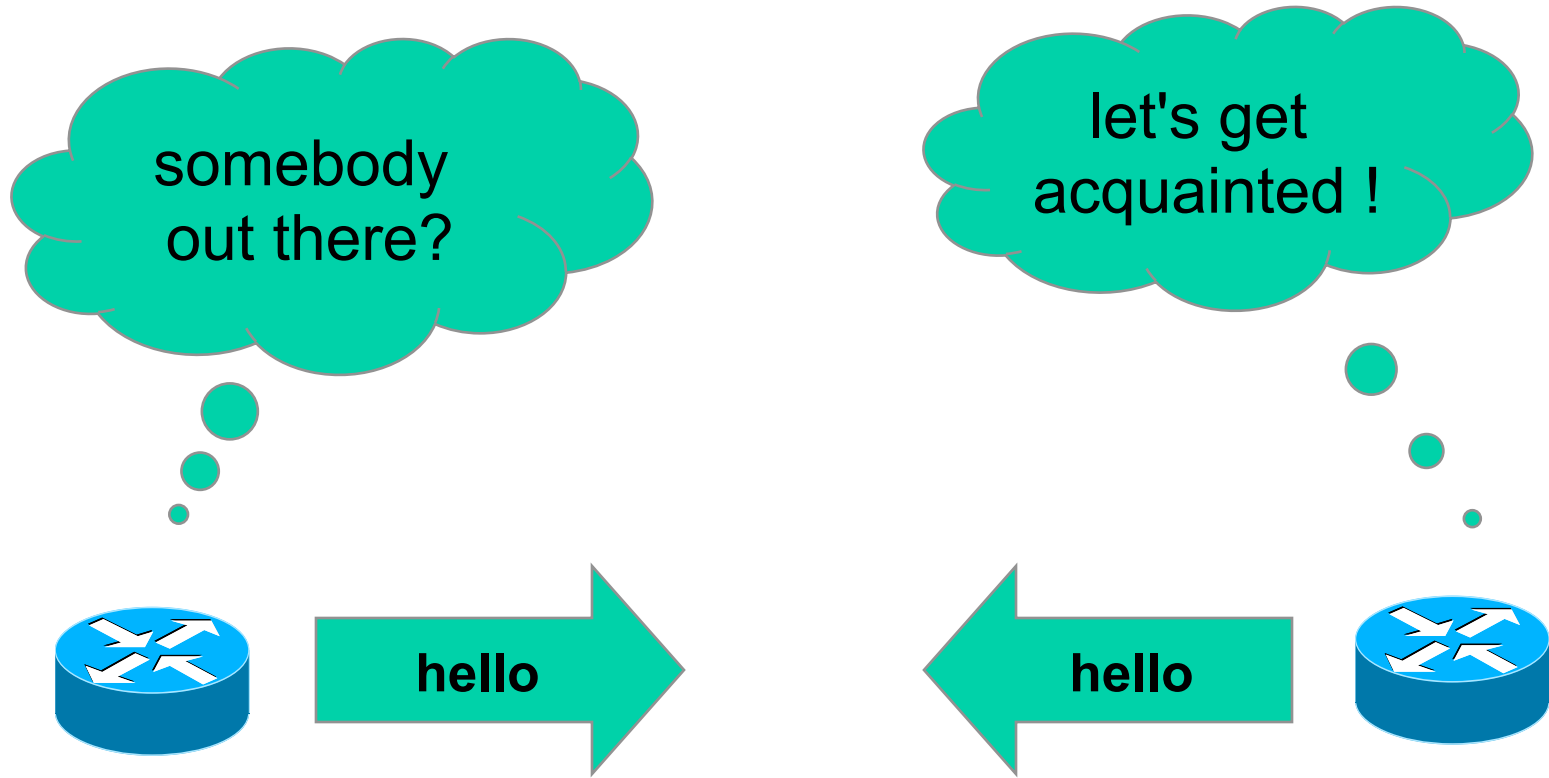
OSPF Communication Principle 2

- Each two acquainted routers send database description messages to each other, in order to publish their topology database
- Unknown or old entries are updated via link state request and link state update messages
 - Which synchronizes the topology databases
- After successful synchronization both routers declare their neighbourhood (adjacency) via router LSAs (using link state update messages)
 - Distributed across the whole network

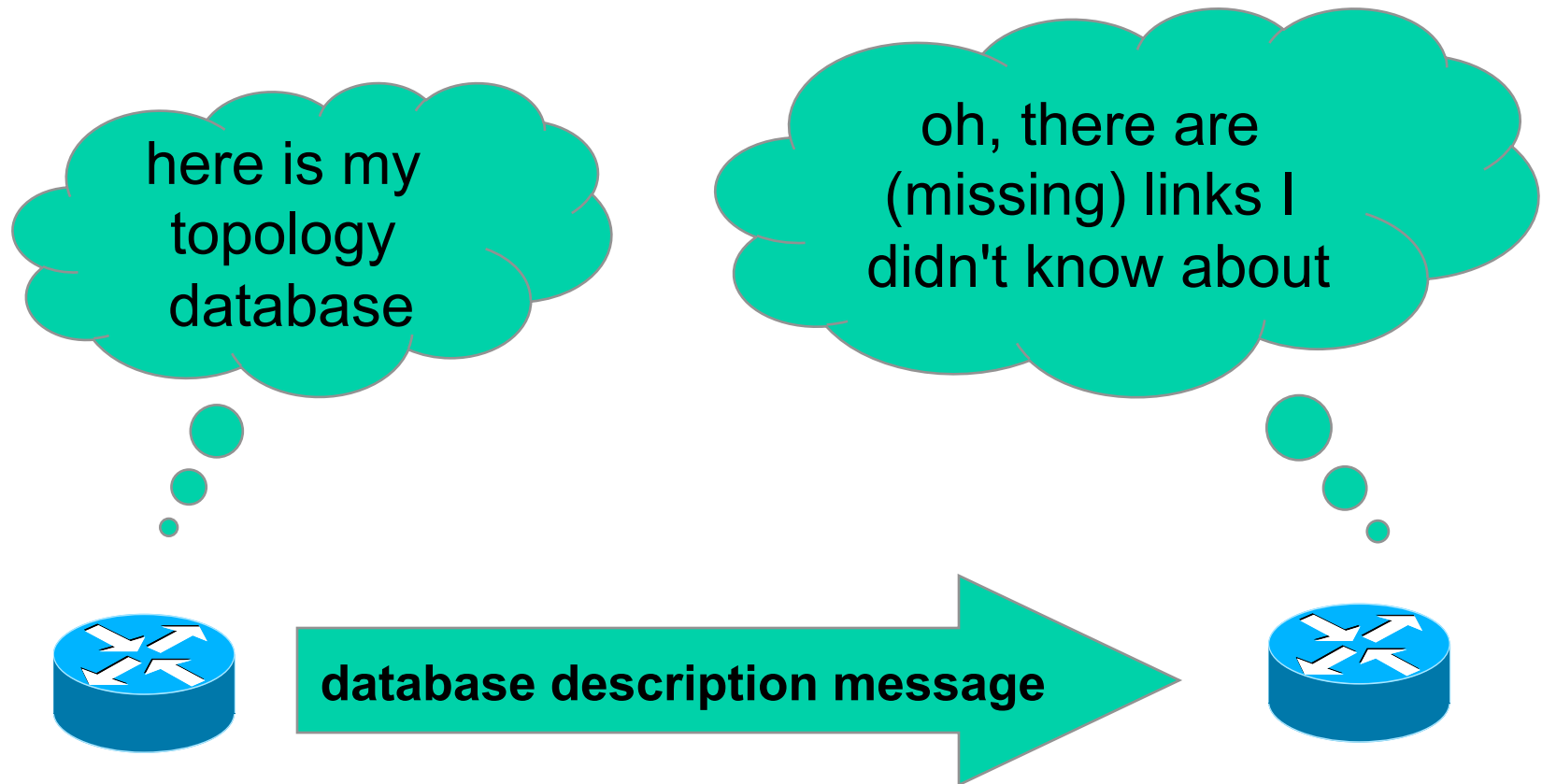
OSPF Communication Principle 3

- **Periodically, every router verifies its link state to its adjacent neighbours using hello messages**
- **From now only changes of link states are distributed**
 - Using link state update messages (LSA broadcast-mechanism)
- **If neighbourhood situation remains unchanged, the periodic hello messages represents the only routing overhead**
 - Note: additionally all Link States are refreshed every 30 minutes with LSA broadcast mechanism

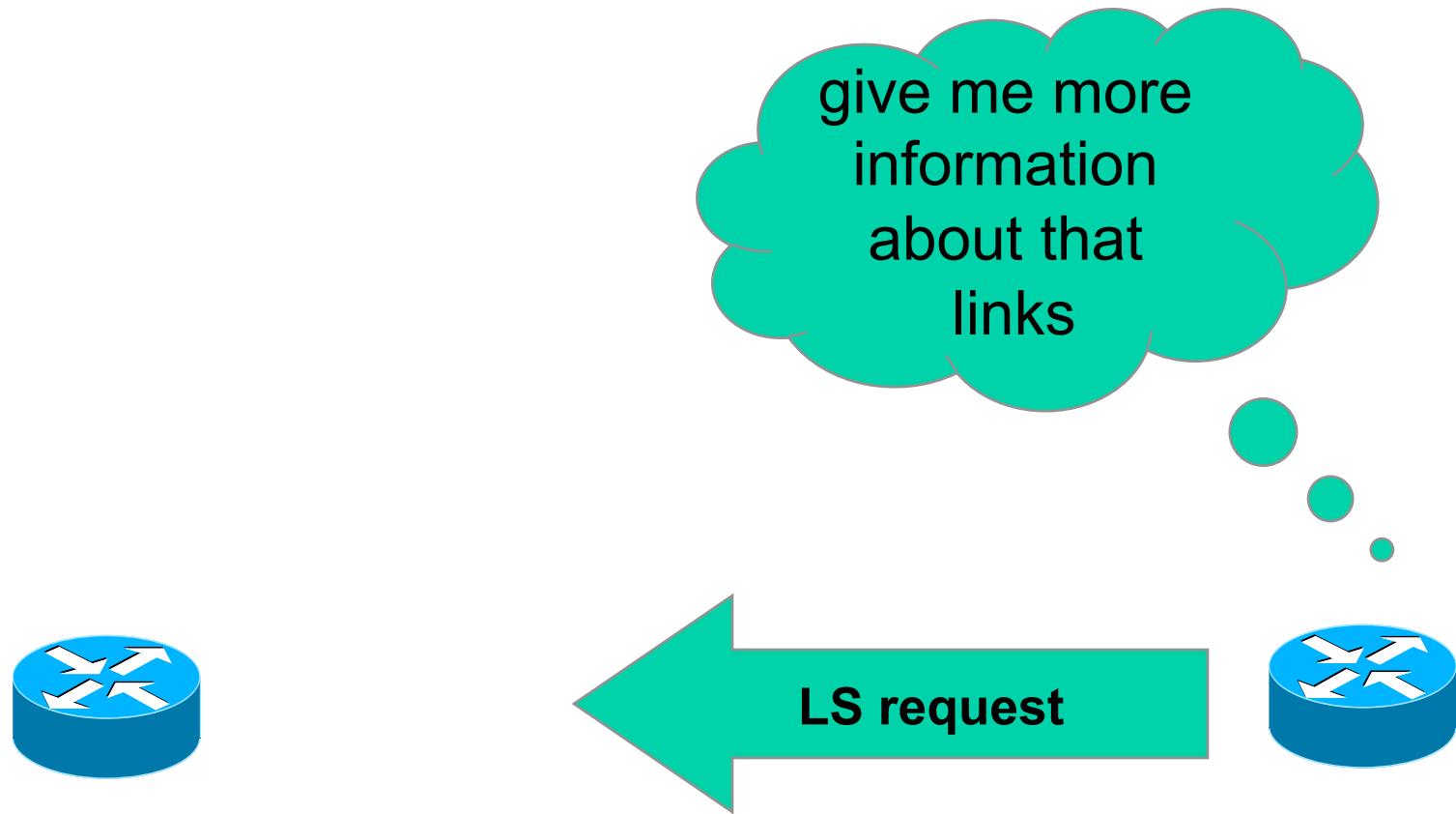
OSPF Communications Summary 1



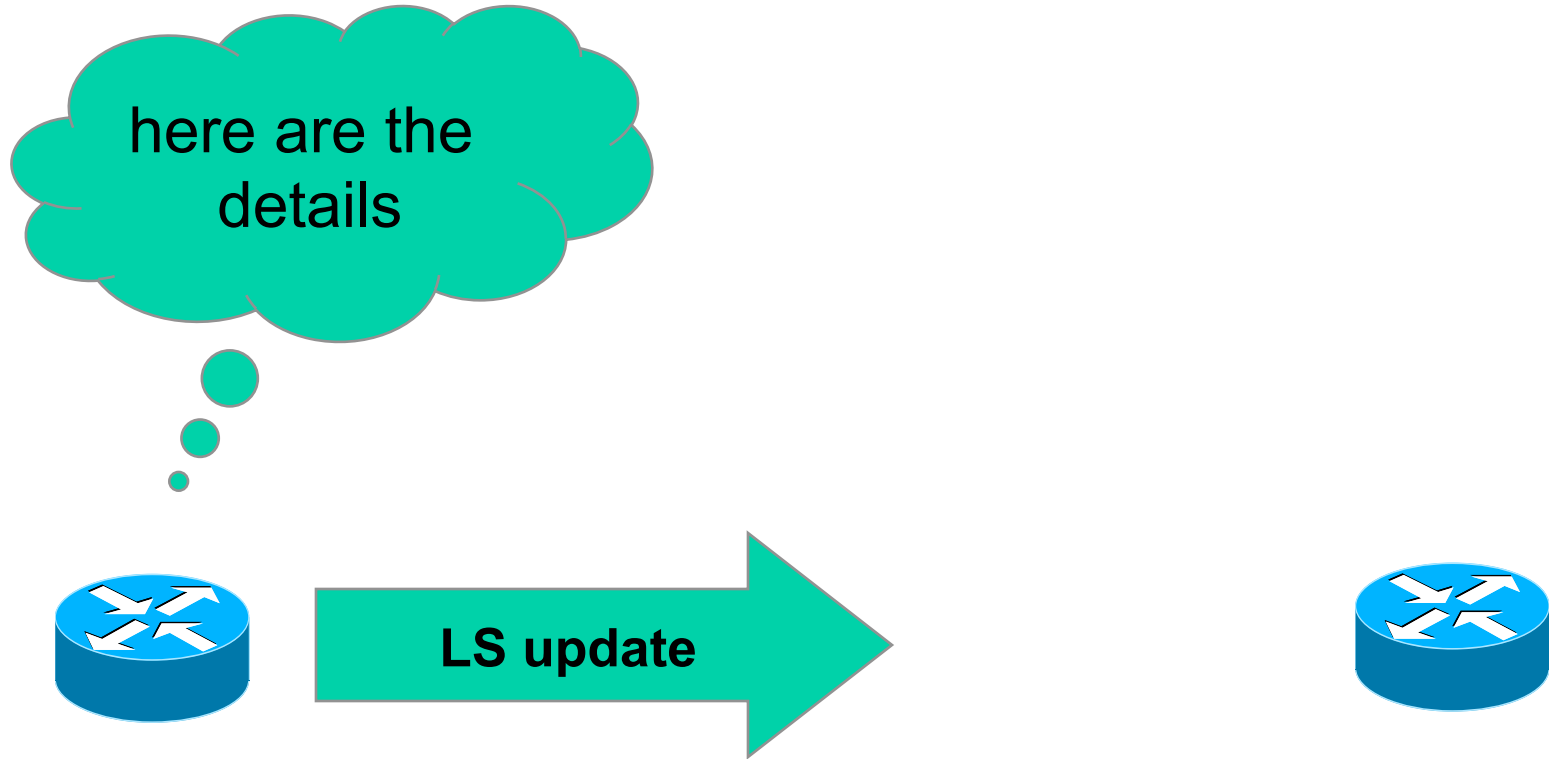
OSPF Communications Summary 2



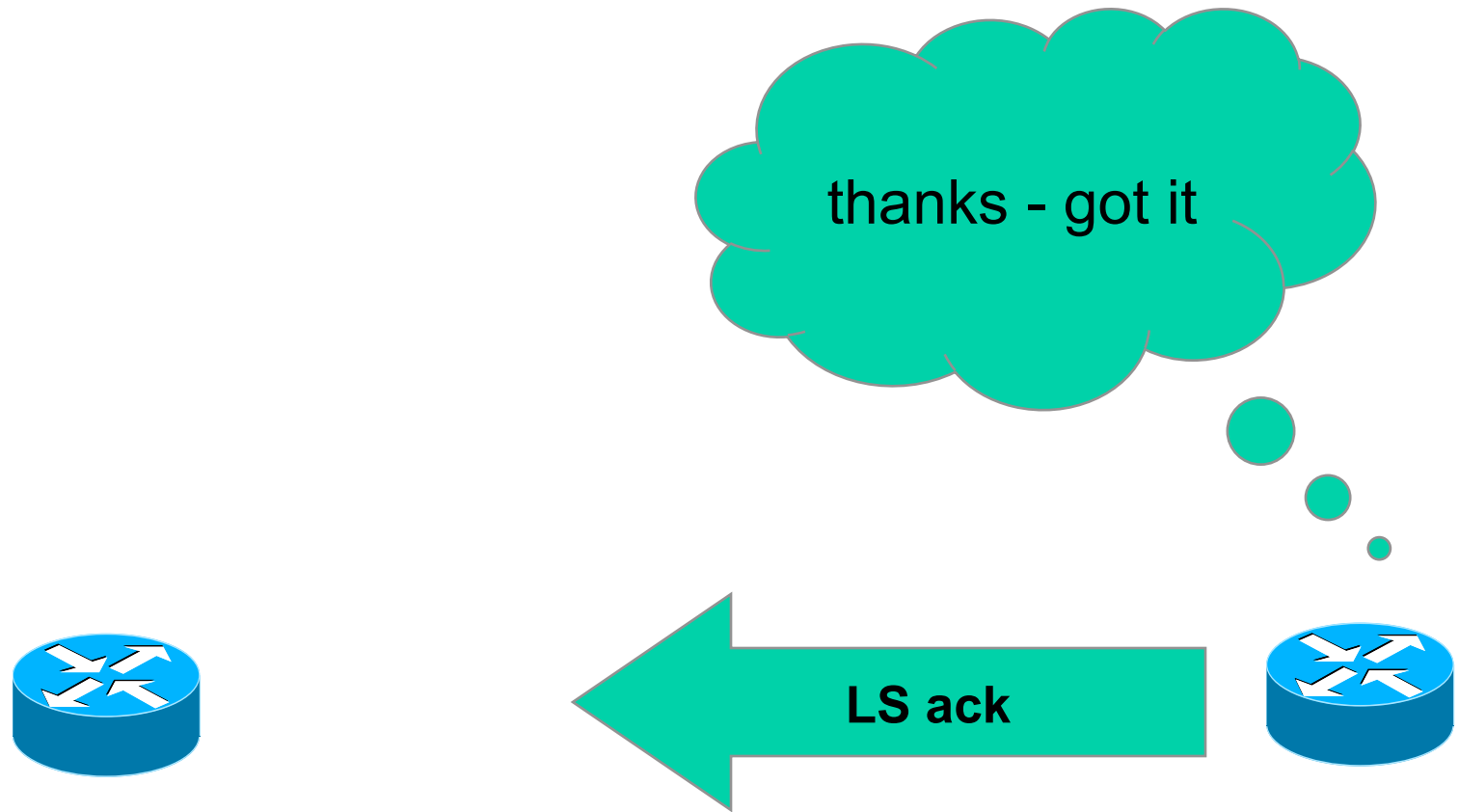
OSPF Communications Summary 3



OSPF Communications Summary 4



OSPF Communications Summary 5



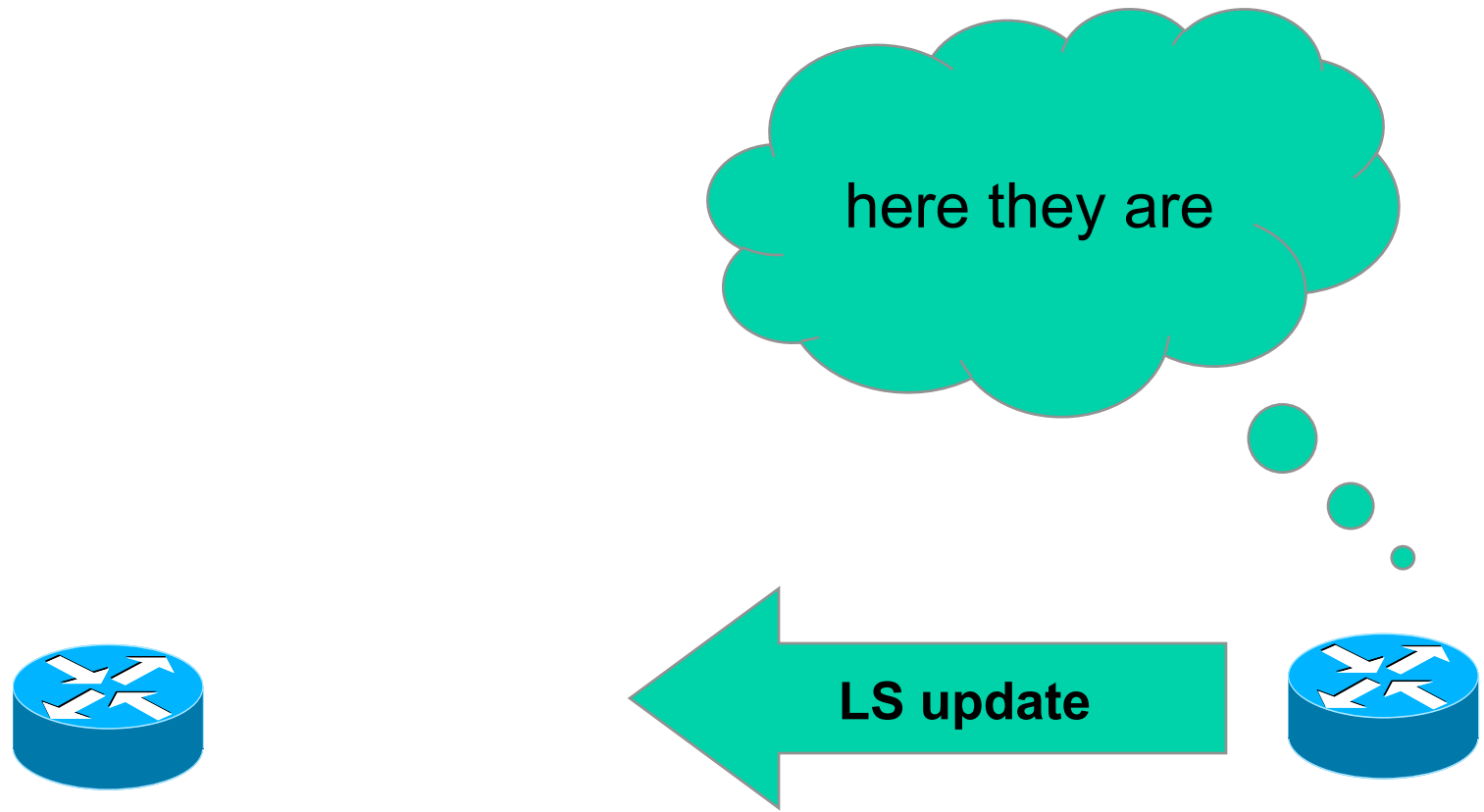
OSPF Communications Summary 6



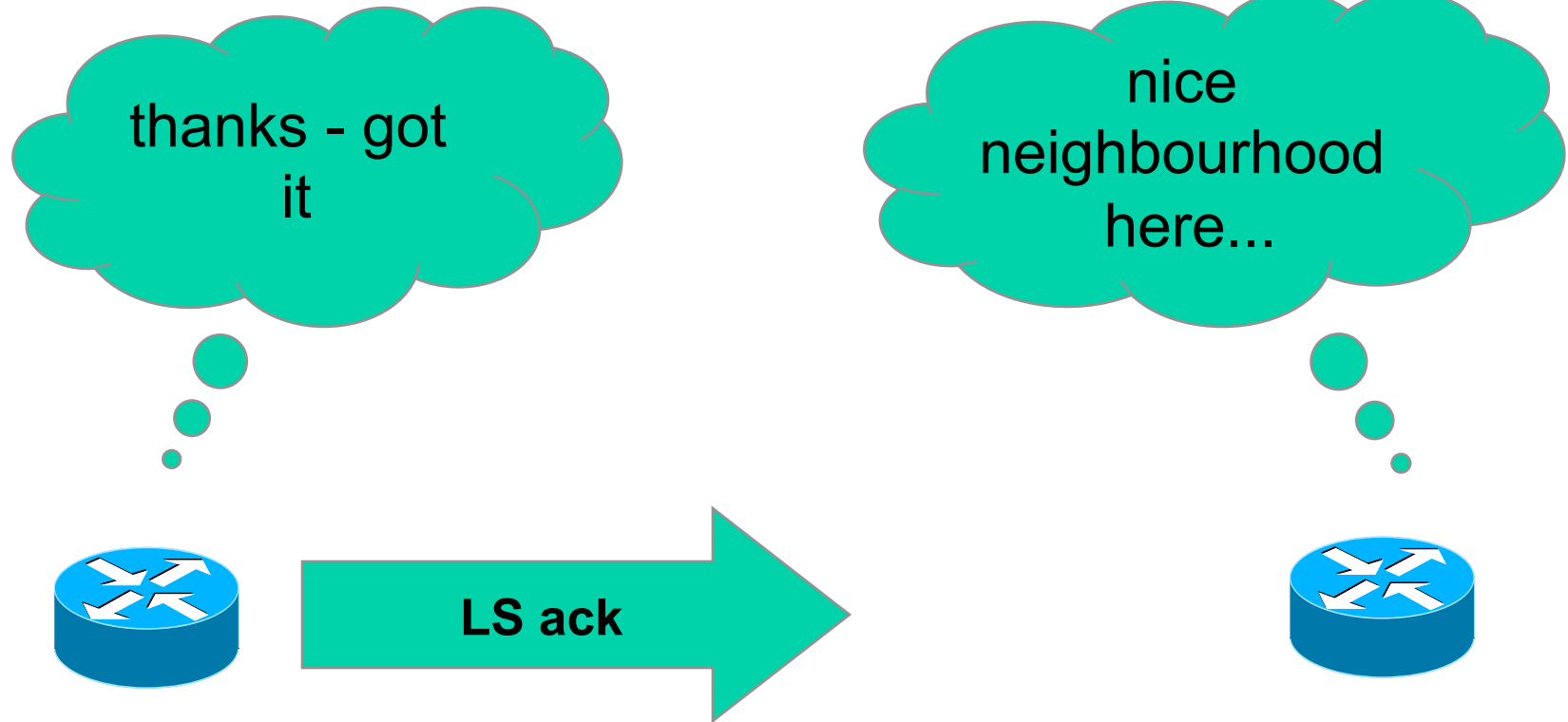
OSPF Communications Summary 7



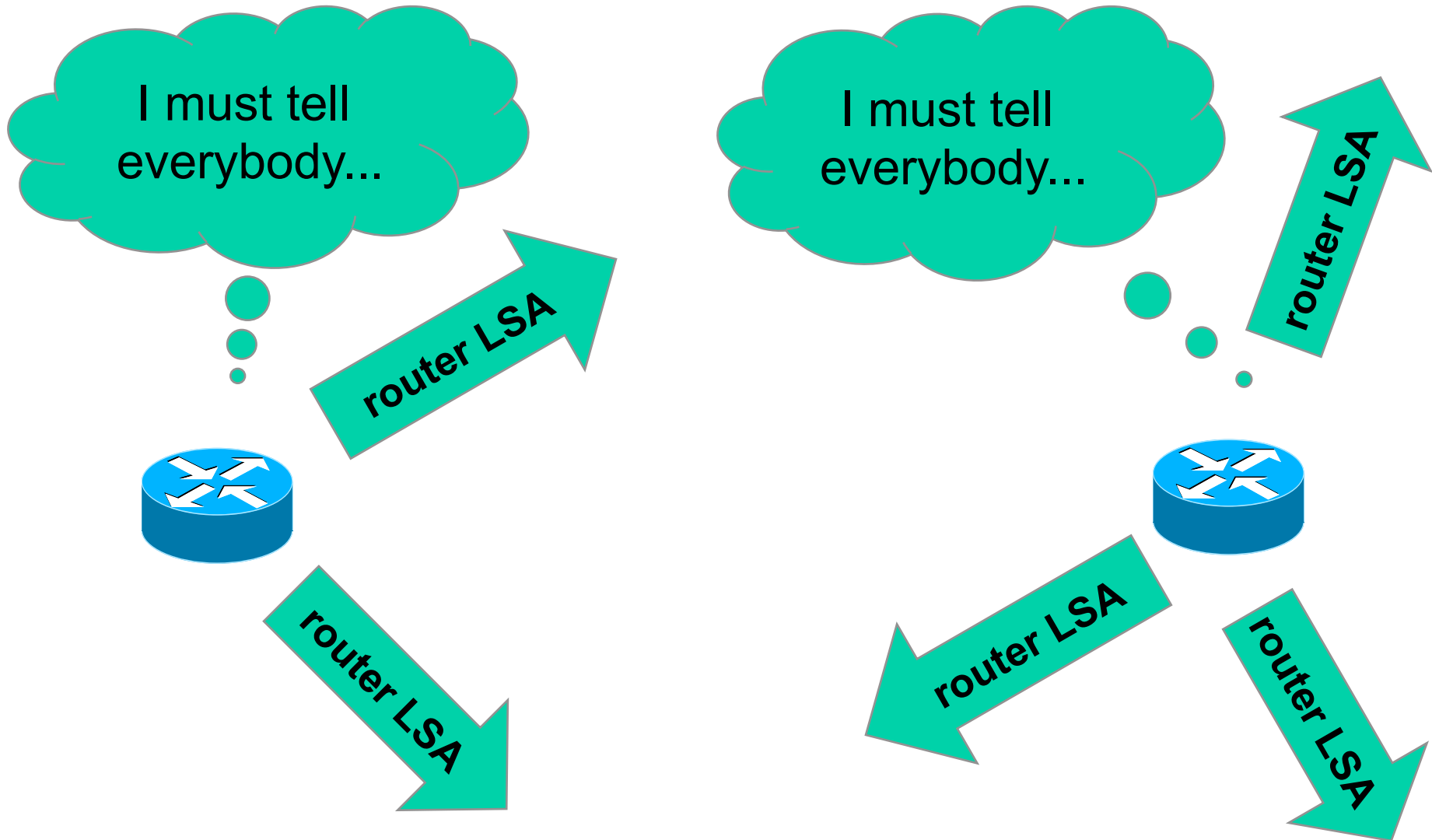
OSPF Communications Summary 8



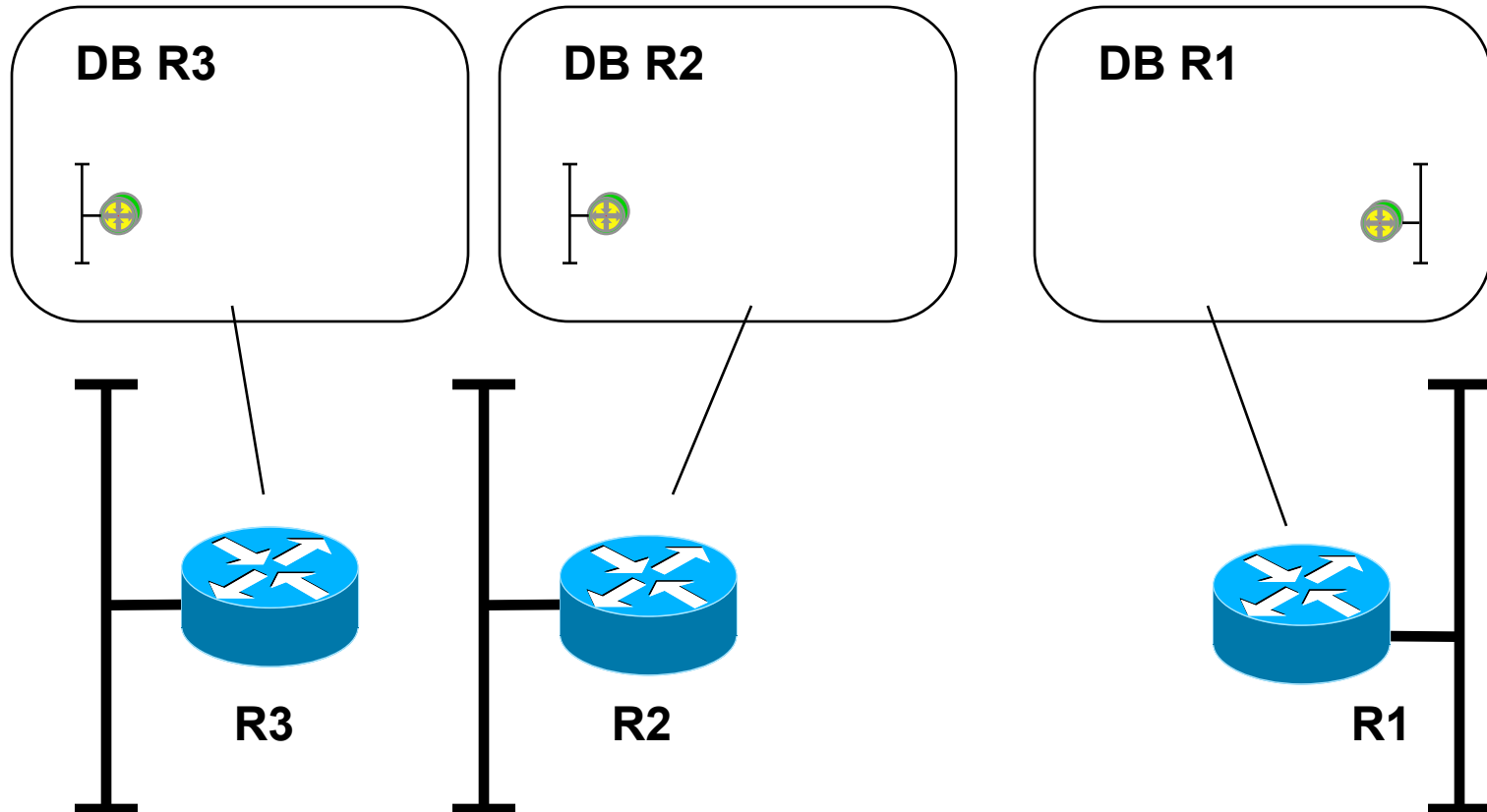
OSPF Communications Summary 9



OSPF Communications Summary 10

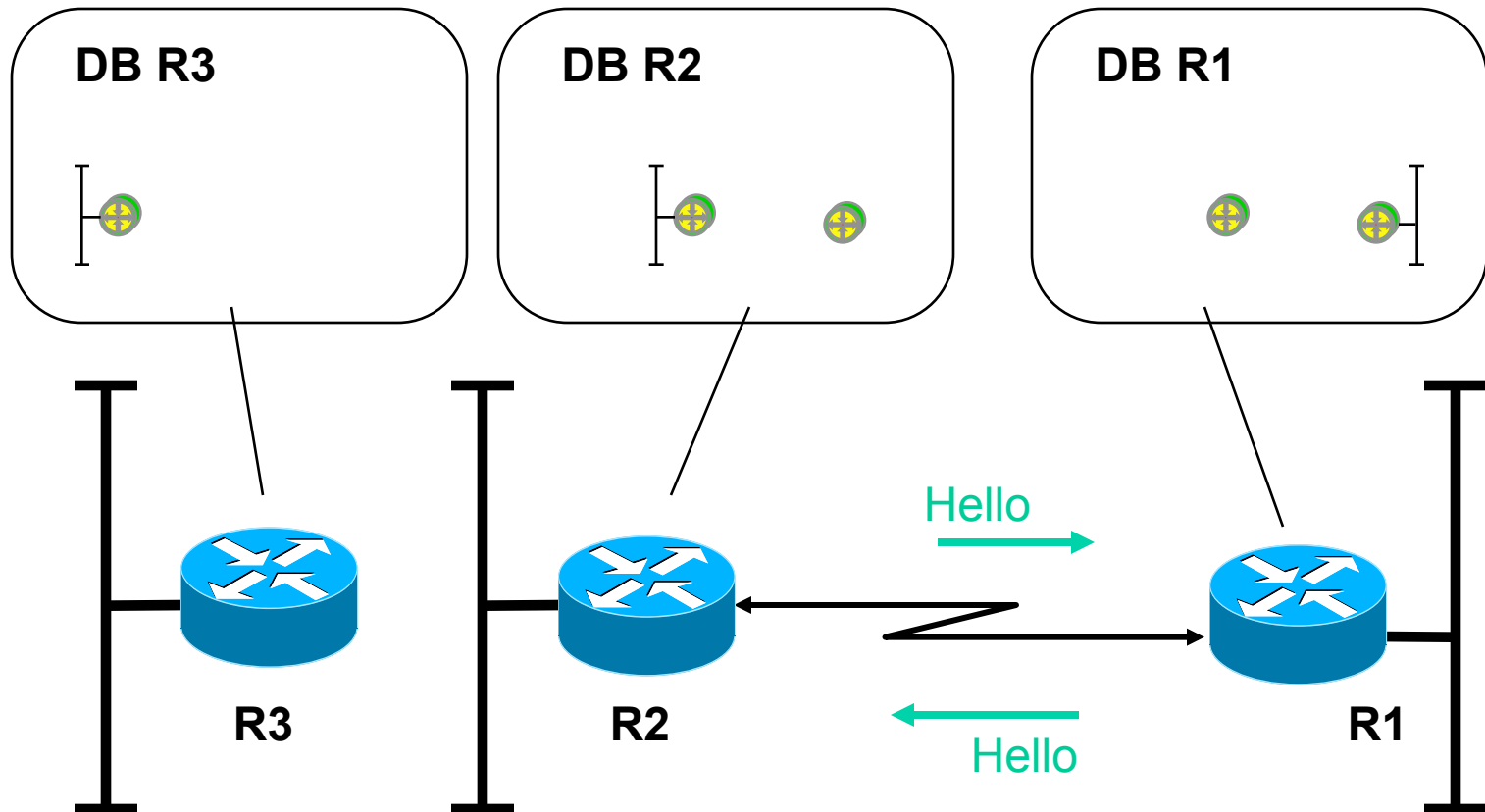


OSPF Start-up



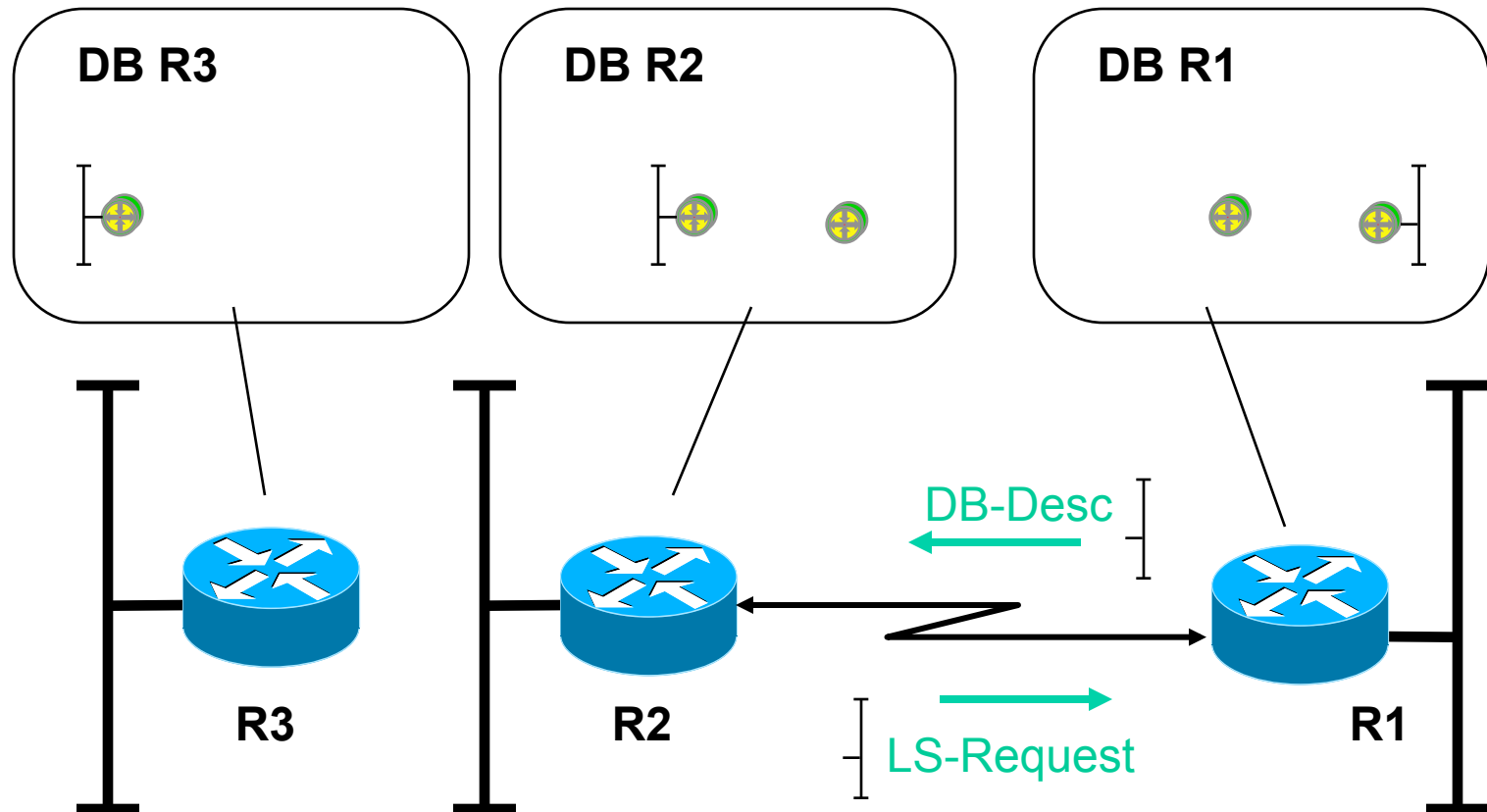
**starting position: all routers initialized,
no connection between R1-R2 or R2-R3**

OSPF Hello R1 - R2



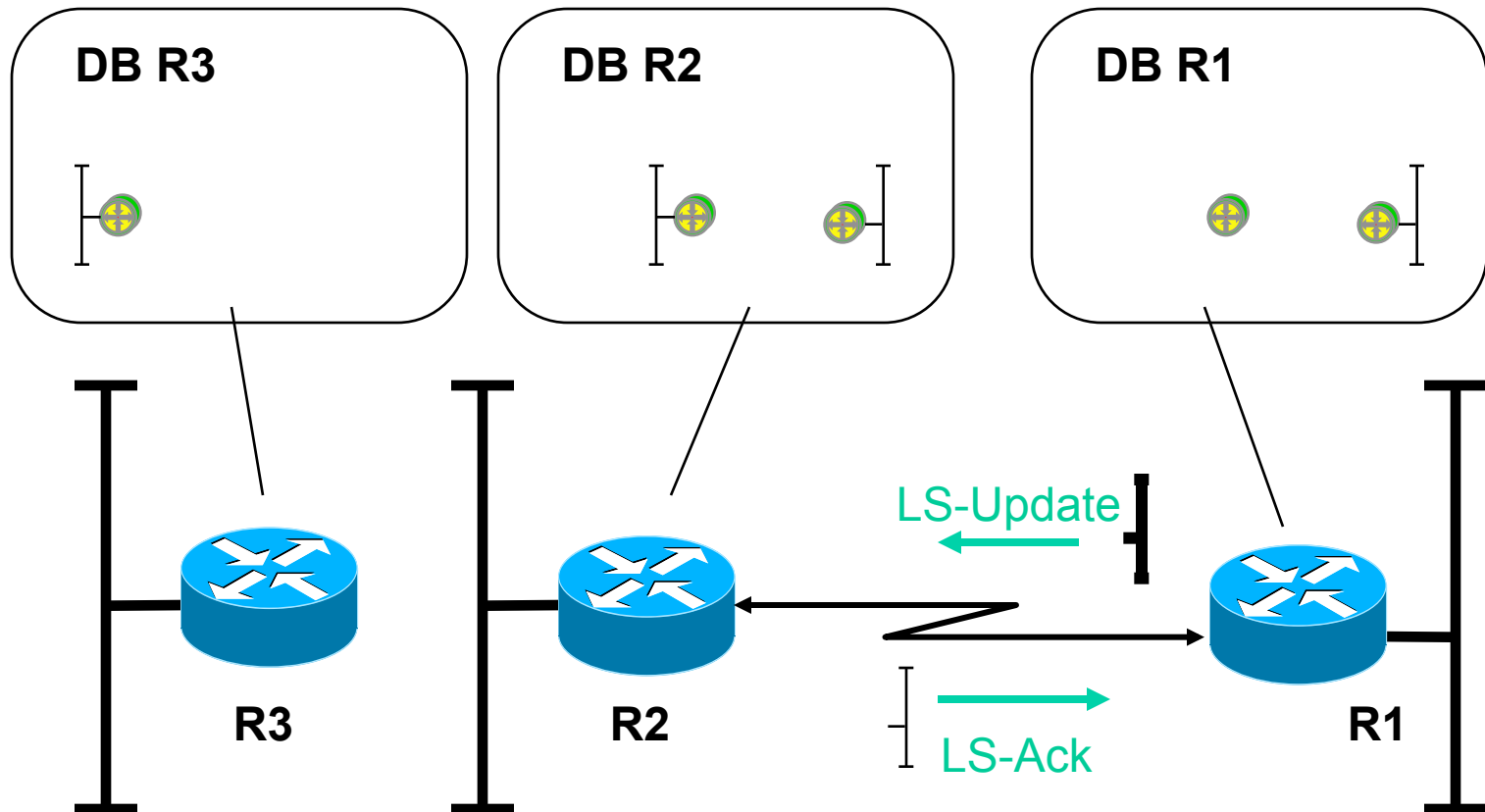
link between R1-R2 activated: get acquainted using hello messages

OSPF Data Base Description R1 -> R2



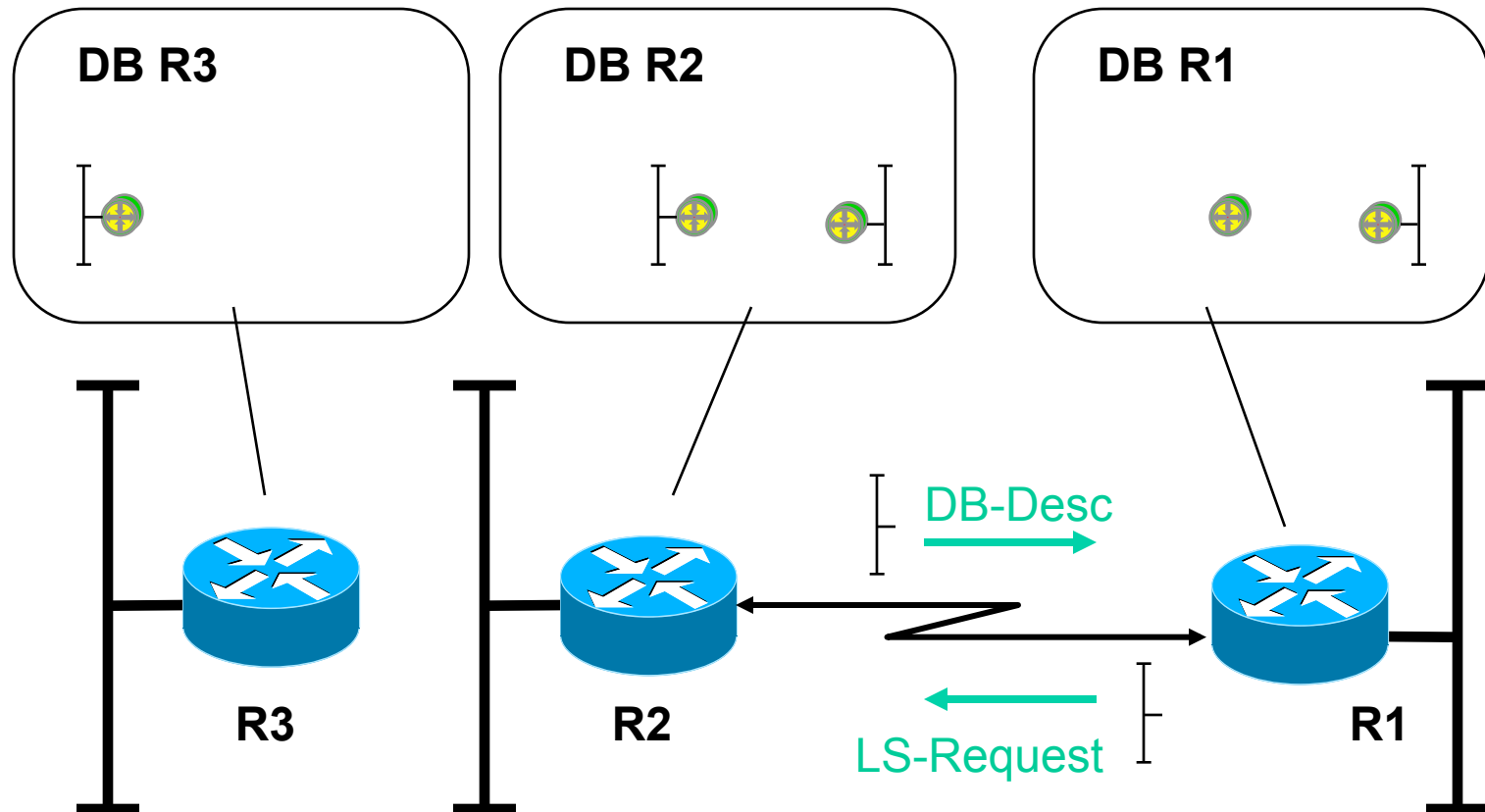
database synchronization: R1 master sends Database-Description, R2 slave sends Link-State Request

OSPF Data Base Update R1 -> R2



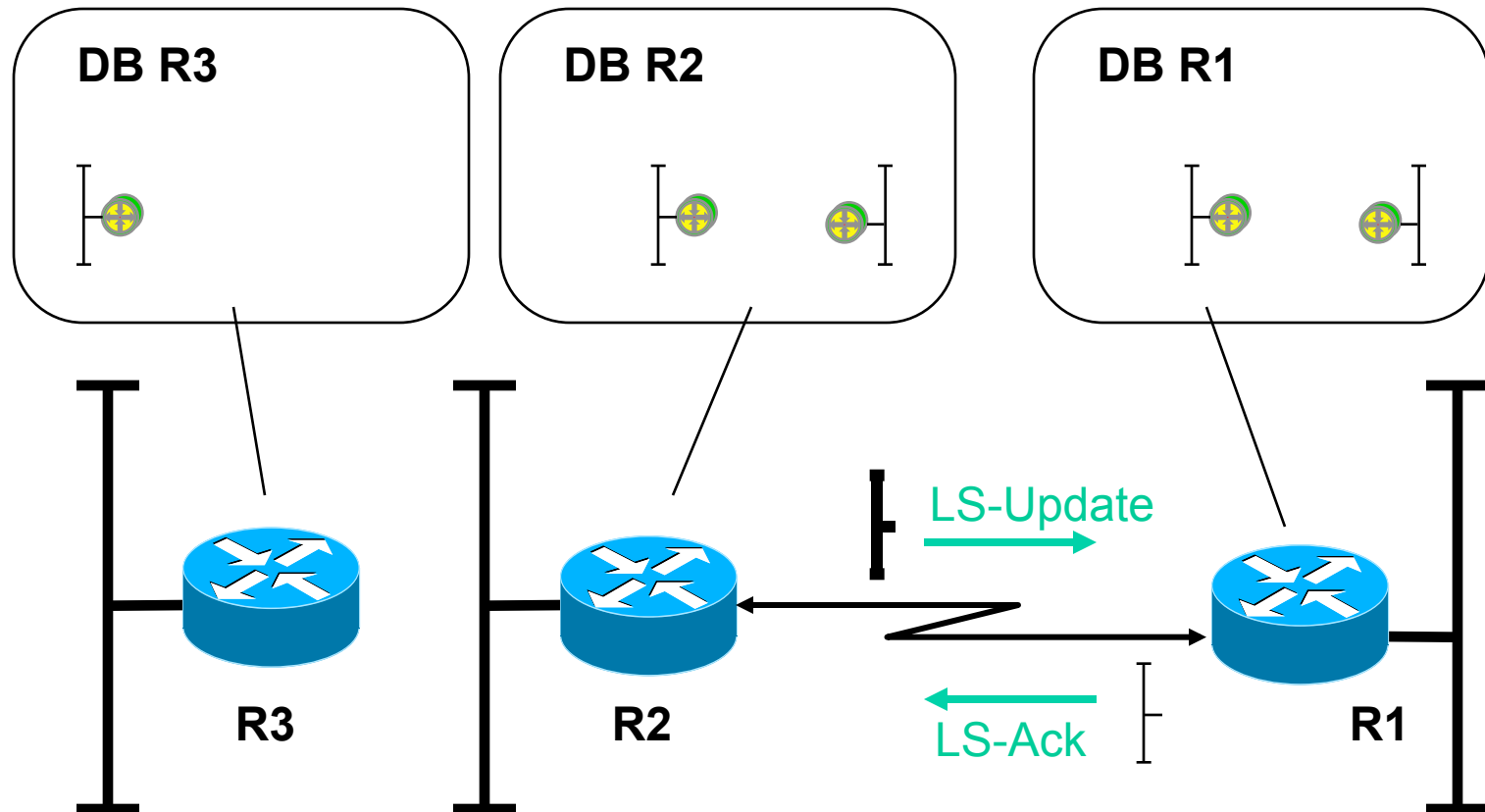
**database synchronization: R1 master
sends Link-State Update, R2 slave
sends Link-State Acknowledgement**

OSPF Data Base Description R2 -> R1



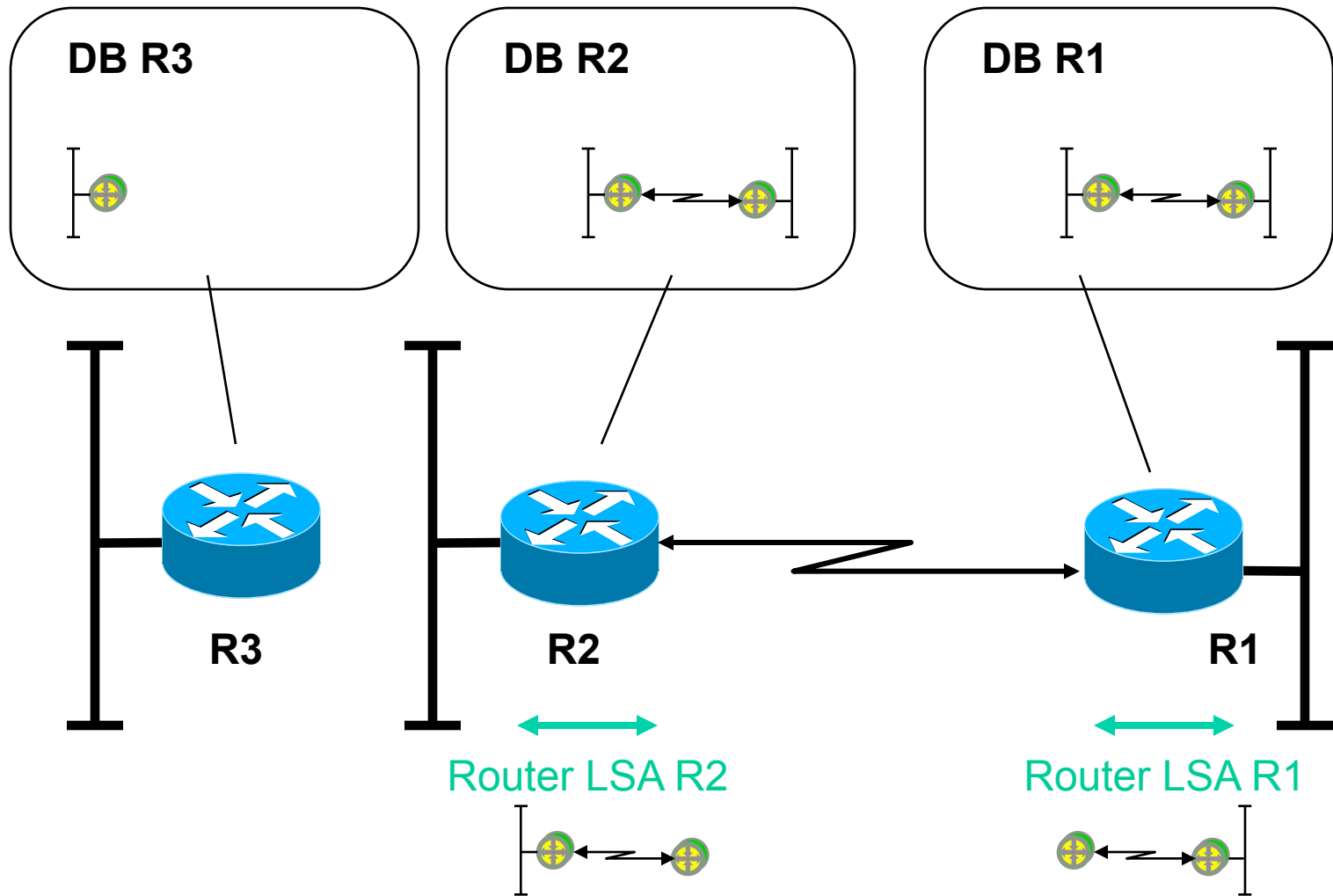
database synchronization: R2 master sends Database-Description, R1 slave sends Link-State Request

OSPF Data Base Update R2 -> R1



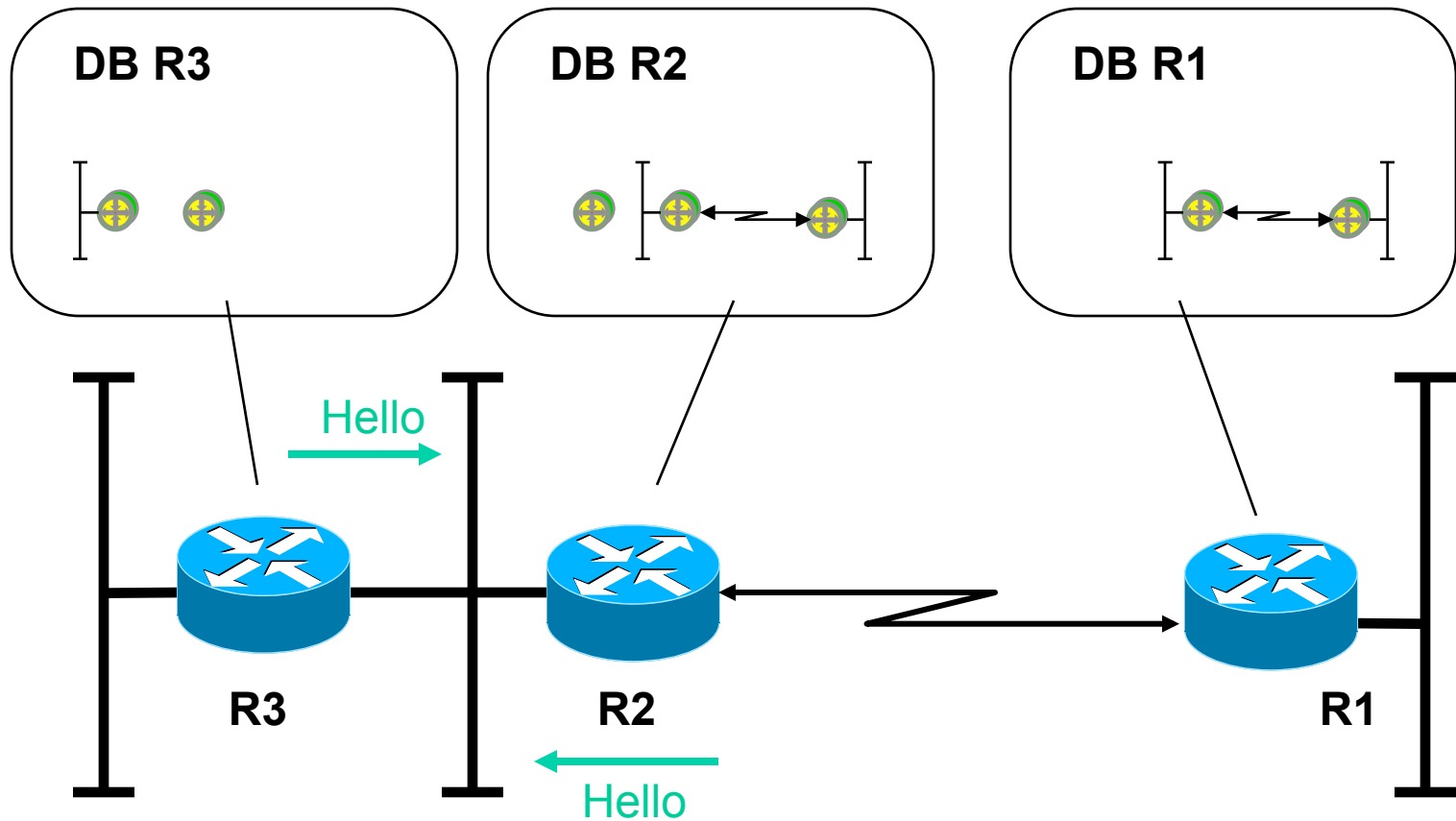
**database synchronization: R2 master
sends Link-State Update, R1 slave
sends Link-State Acknowledgement**

OSPF Router LSA Emission



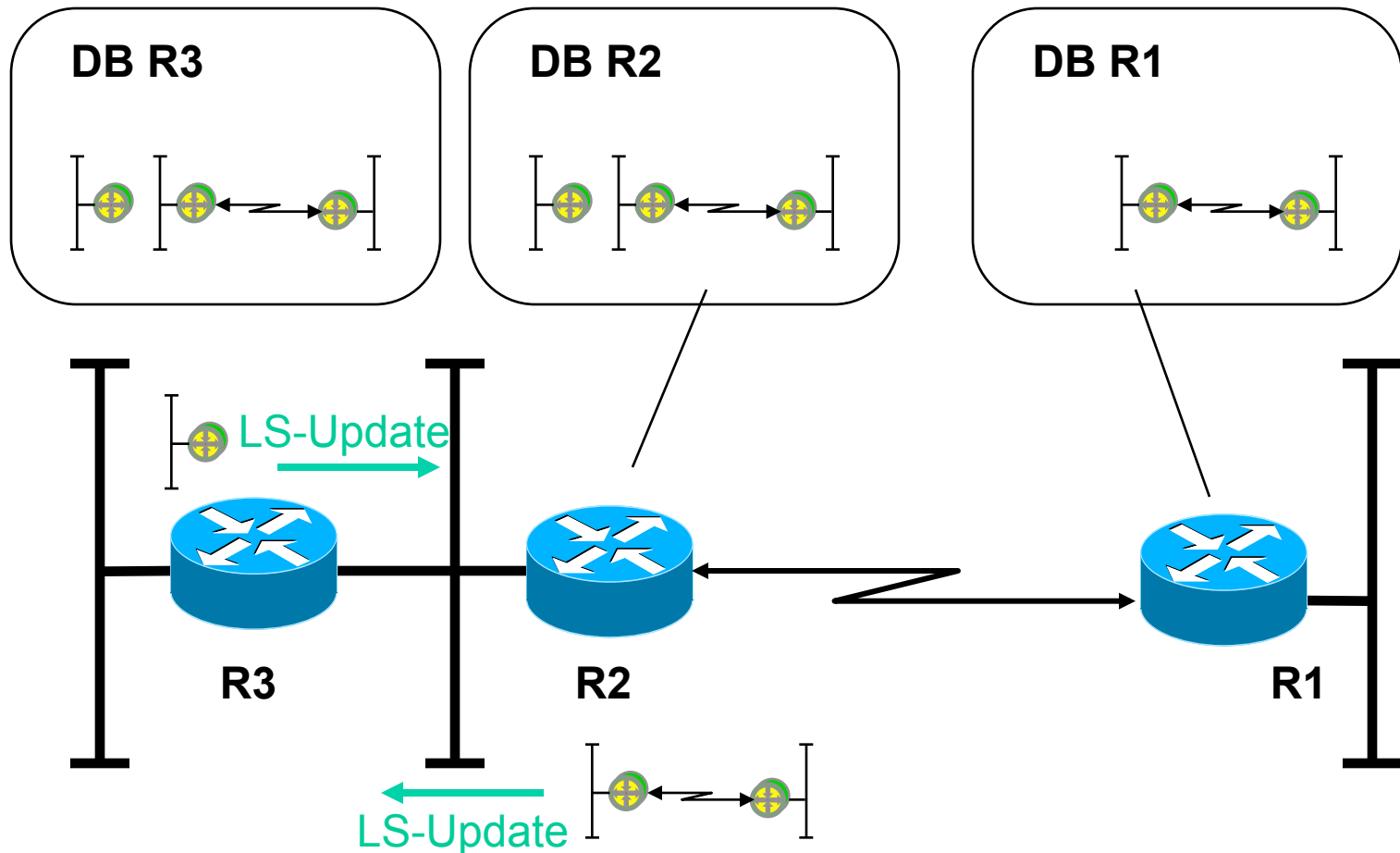
R1 and R2 have synchronized their database completely and notify other nodes about their links

OSPF Hello R2 - R3



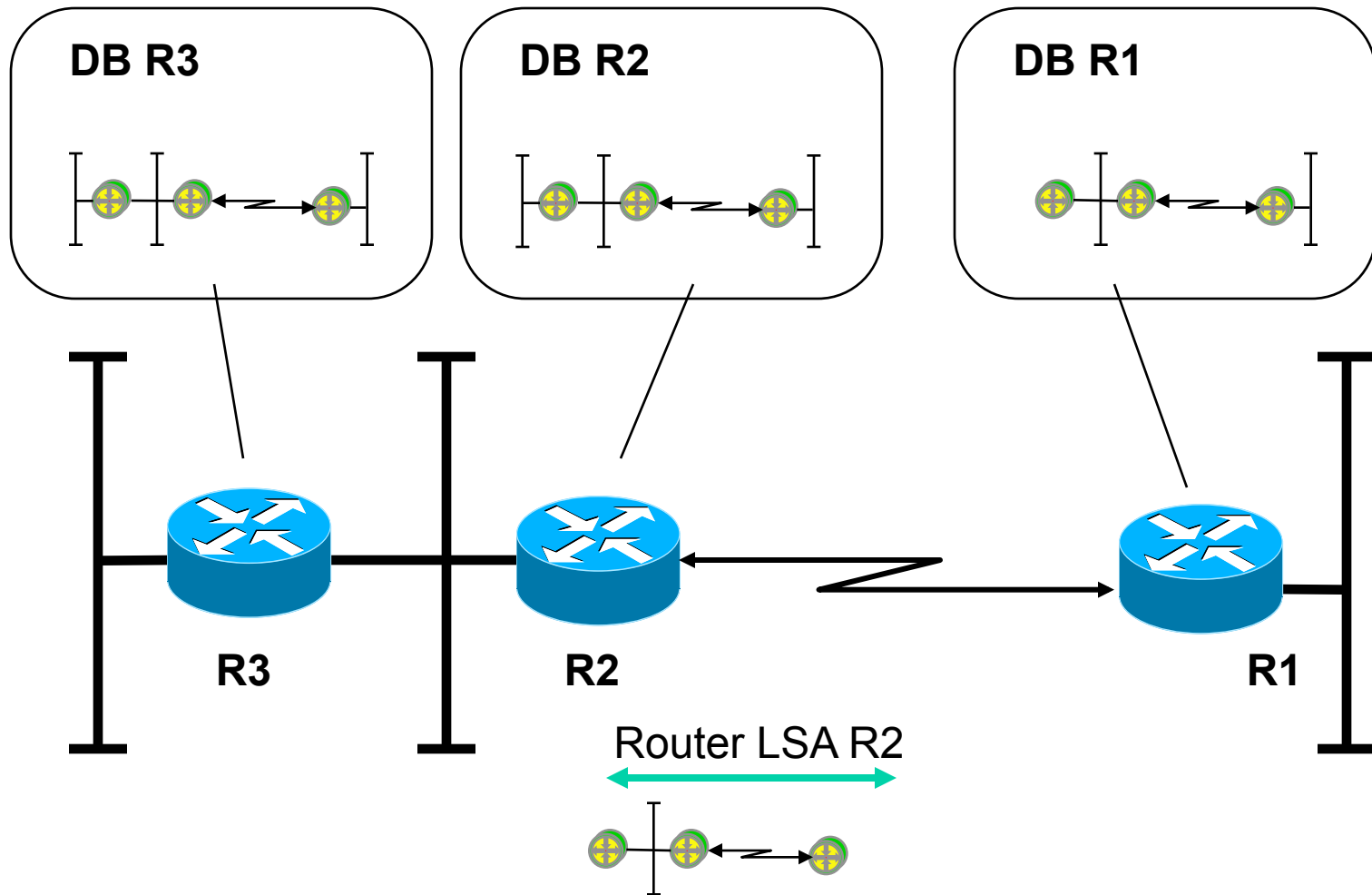
link between R2-R3 activated: get acquainted using Hello, determination of designated router

OSPF Database Update



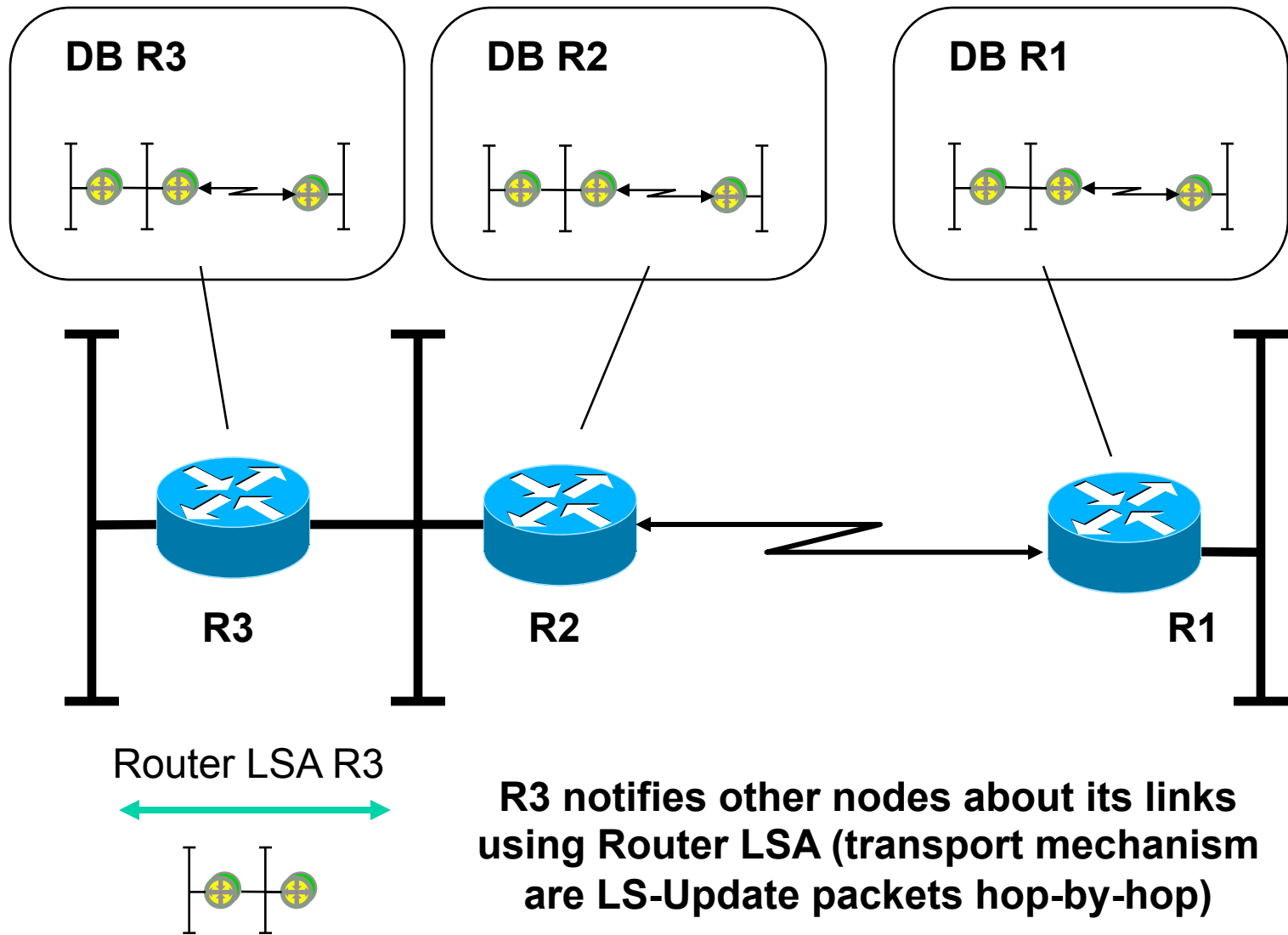
**R2 and R3 synchronize their databases
(DB-Des., LS-Req.,LS-Upd., LS-Ack.)**

OSPF Router LSA Emission R2

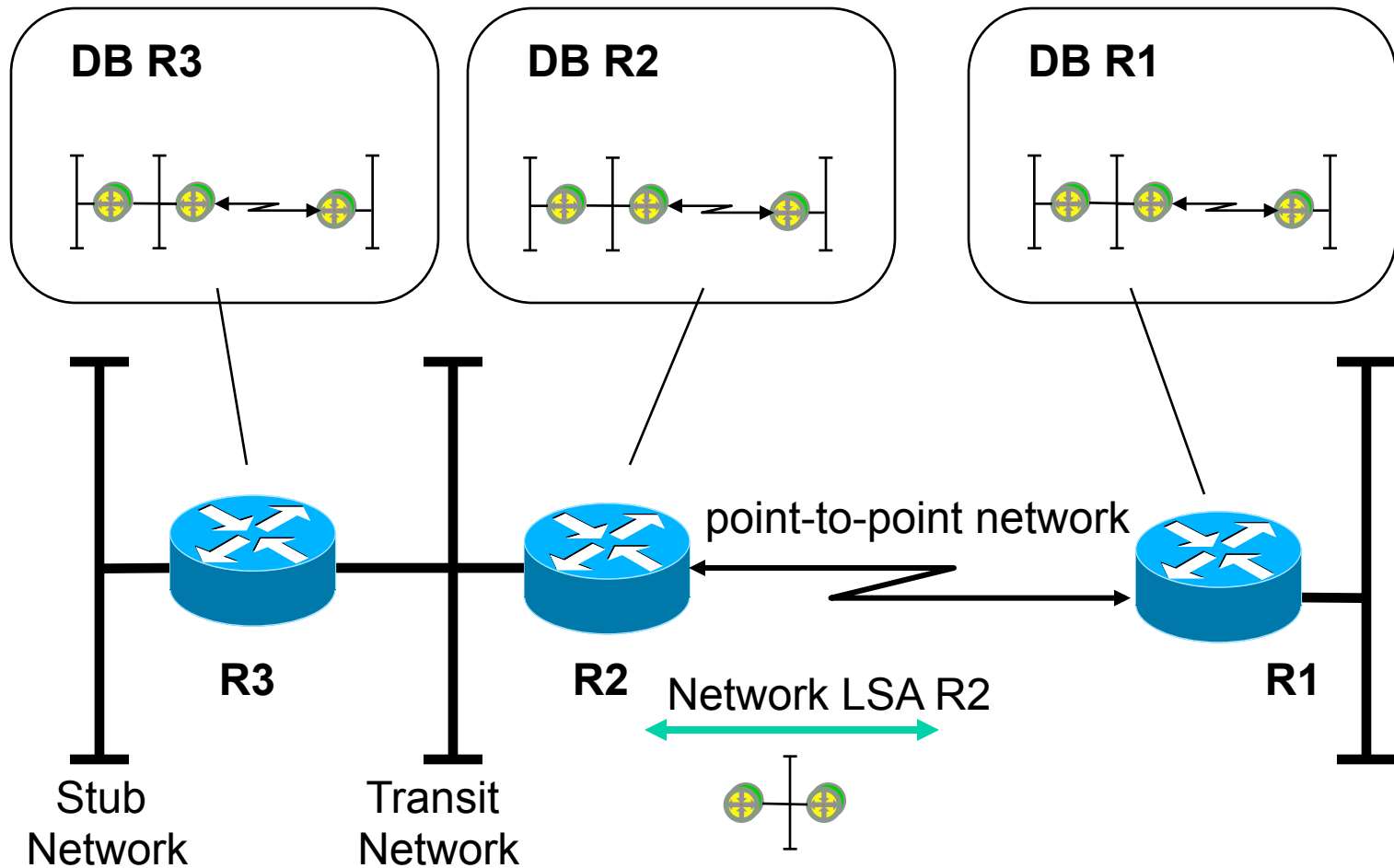


R2 notifies other nodes about its links using Router LSA, (transport mechanism are LS-Update packets hop-by-hop)

OSPF Router LSA Emission R3



OSPF Network LSA R2



Designated Router R2 notifies other nodes about the multi-access network using Network-LSA (transport mechanism are LS-Update packets hop-by-hop)

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
 - Introduction
 - The Dijkstra Algorithm
 - Communication Procedures
 - LSA Broadcast Handling
 - Split Area
 - Broadcast Networks
 - Area Principles
 - Stub Areas
 - Route Summarization
 - Virtual Link
 - Summary
 - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

LSA Broadcast Mechanism (1)

- **Flooding mechanism**

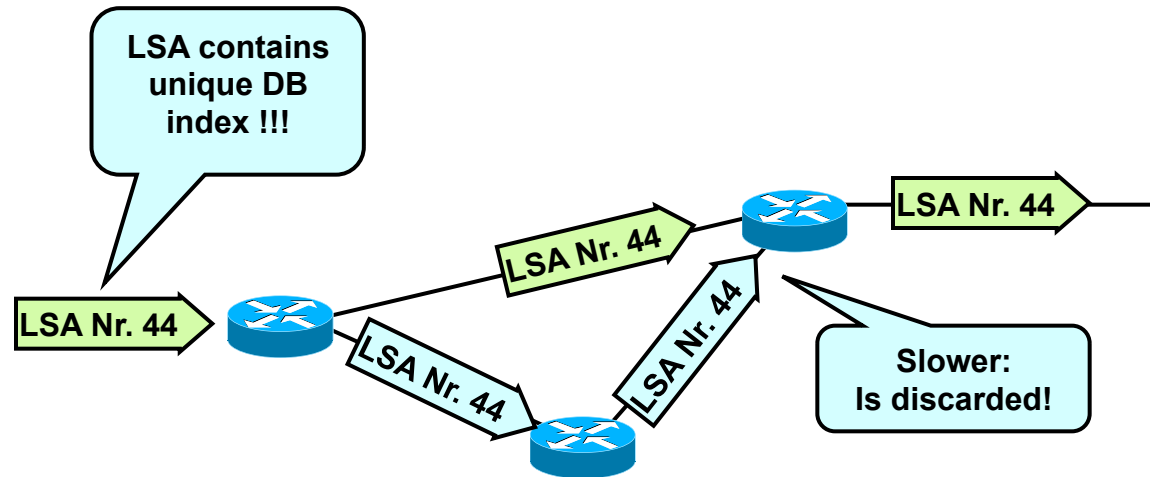
- Receive of LSA on incoming interface
- Forwarding of LSA on all other interfaces except incoming interface
- Well known principle to reach all parts of a meshed network
 - Remember: Transparent bridging – Ethernet switching for unknown destination MAC address
- “Hot-Potato” method

- **Avoidance of broadcast storm:**

- With the help of LSA sequence numbers carried in LSA packets and topology database
 - Remember: In case of Ethernet switching we had STP to avoid the broadcast storm
 - In our case we want to establish topology database so we do not have any STP information; SPF information and hence routing tables will result from existence of consistent topology databases
 - “Chicken-Egg” problem

LSA Sequence Number

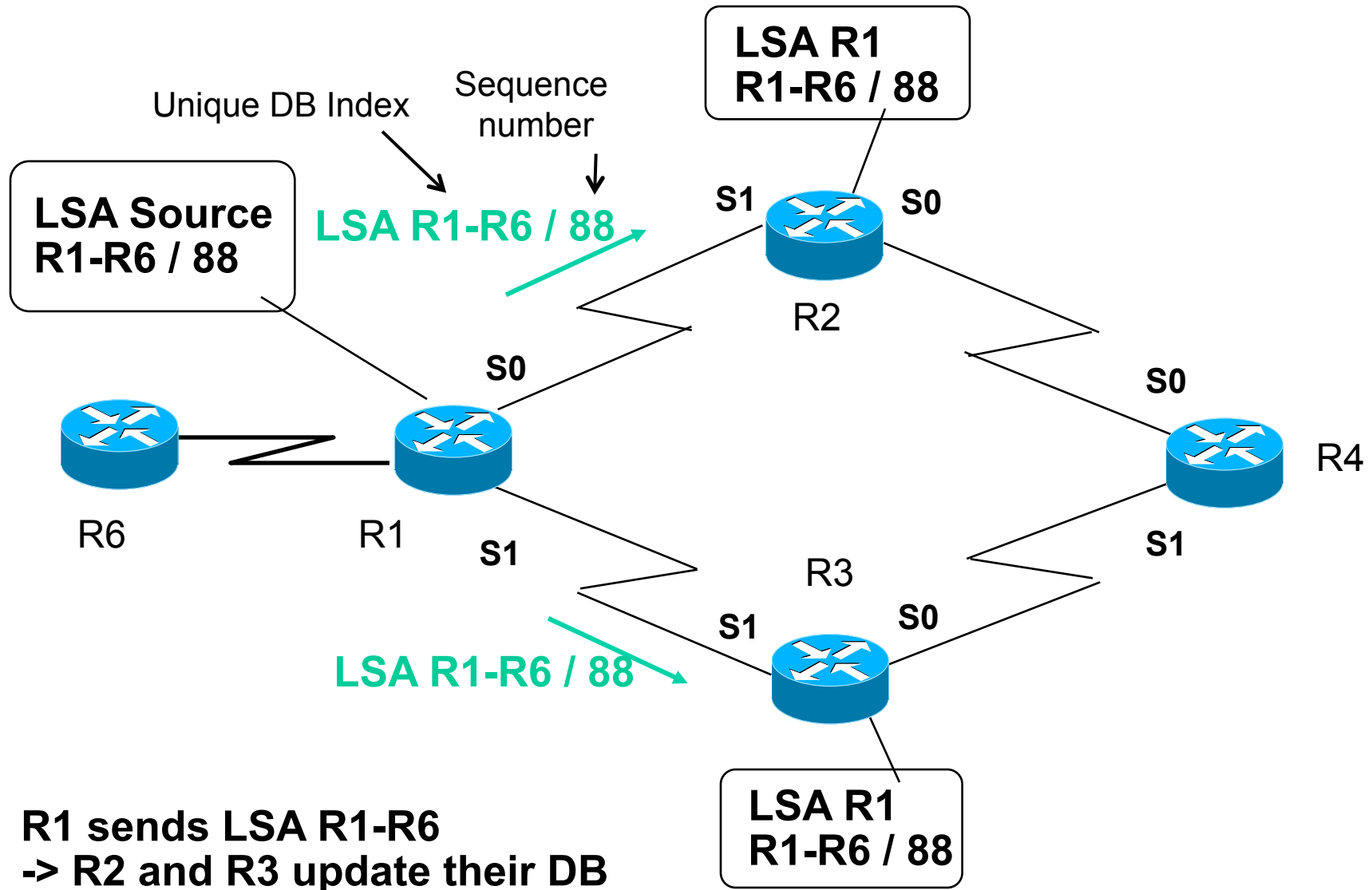
- In order to stop flooding, each LSA carries a sequence number
- Only increased if LSA has changed
 - So each router can check if a particular LSA had already been forwarded
 - To avoid LSA storms
- 32 bit number



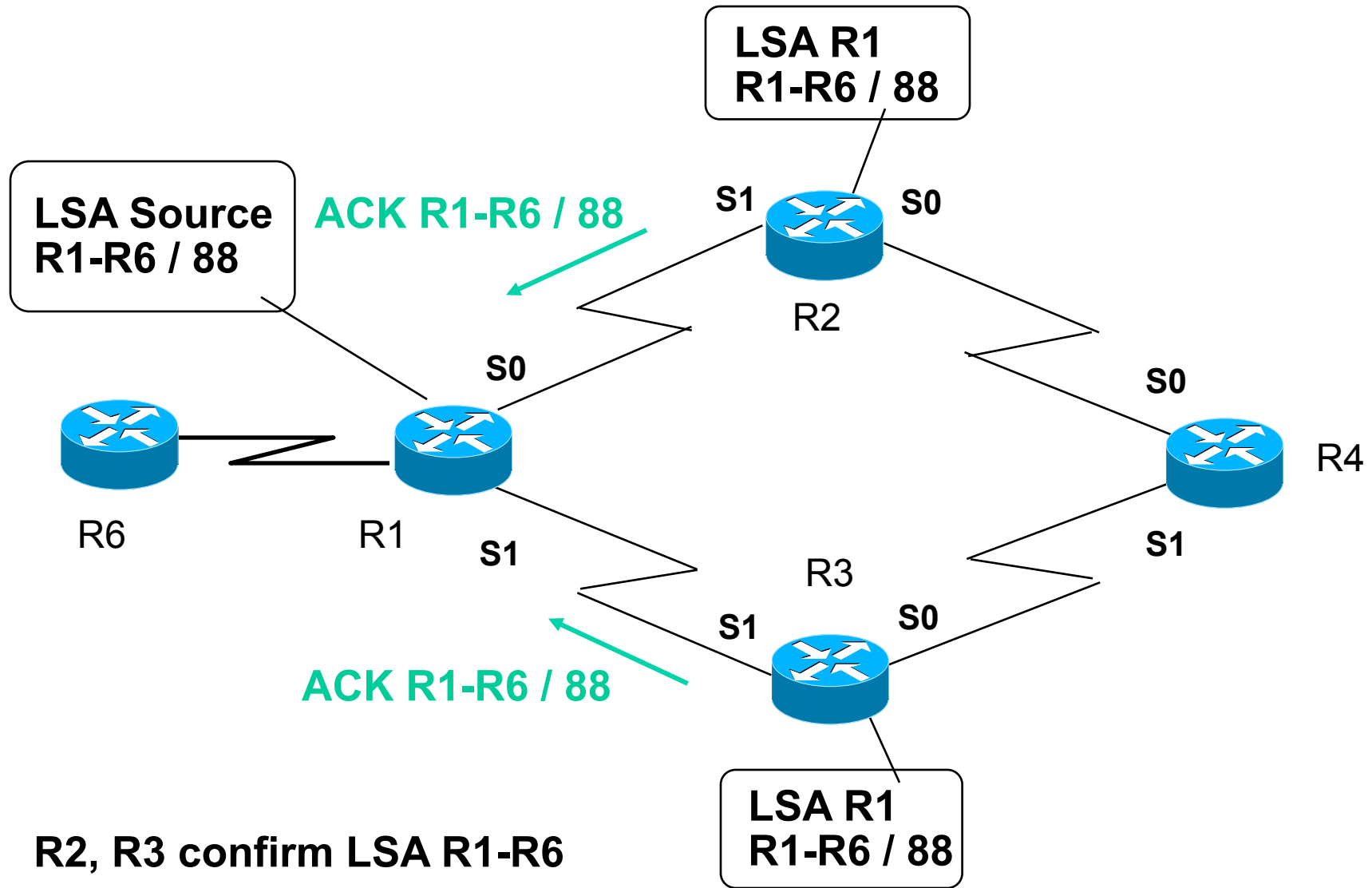
LSA Broadcast Mechanism (2)

- **LSA must be safely distributed to all routers within an area (domain)**
 - Consistency of the topology-database depends on it
 - Every LS-Update is acknowledged explicitly (using LS-ACK) by the neighbor router
 - If a LS-ACK stays out, the LS-Update is repeated (timeout)
 - If the LS-ACK fails after several trials, the adjacency-relation (the link state between the routers) is cleared

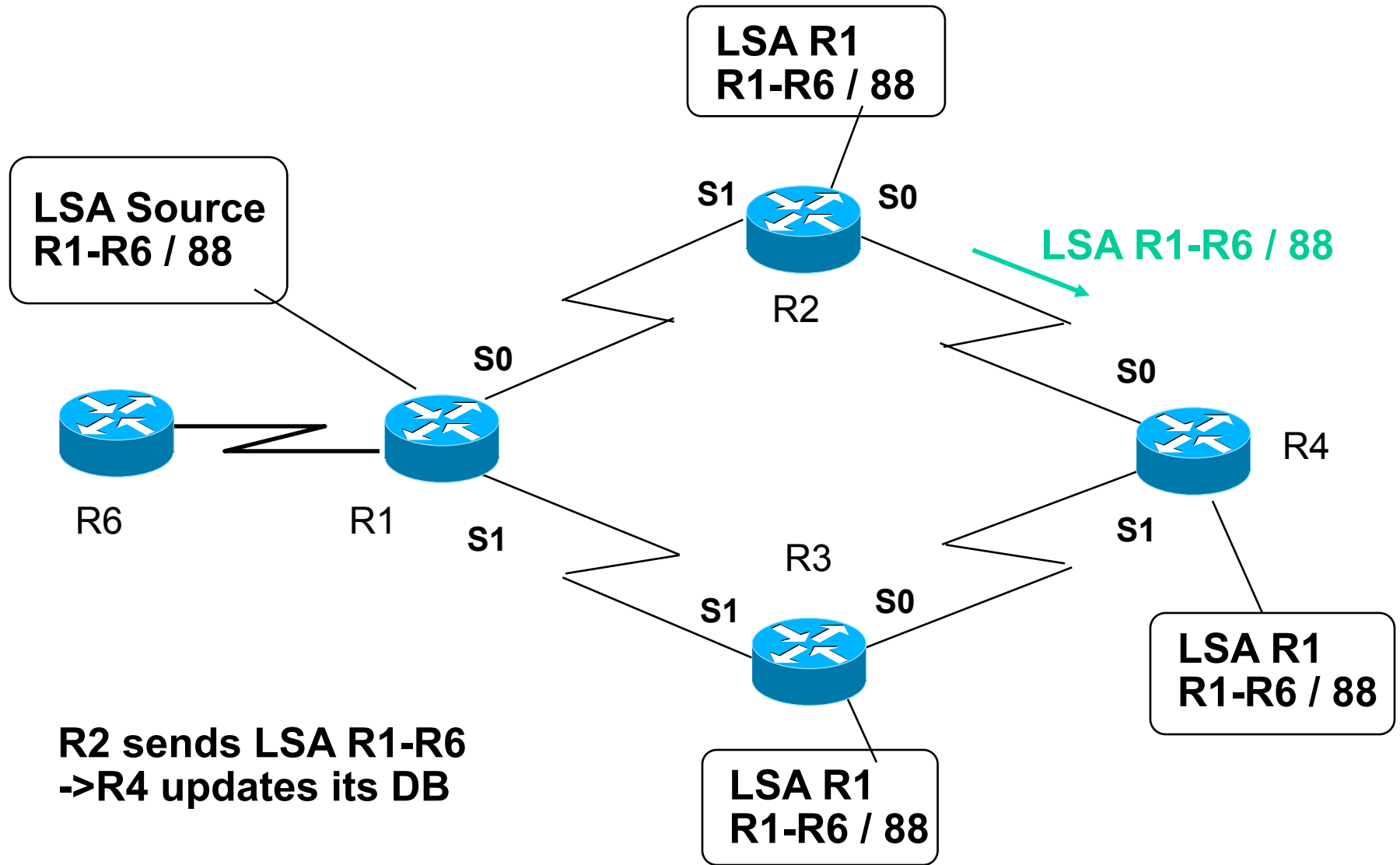
LSA Broadcast Example (1)



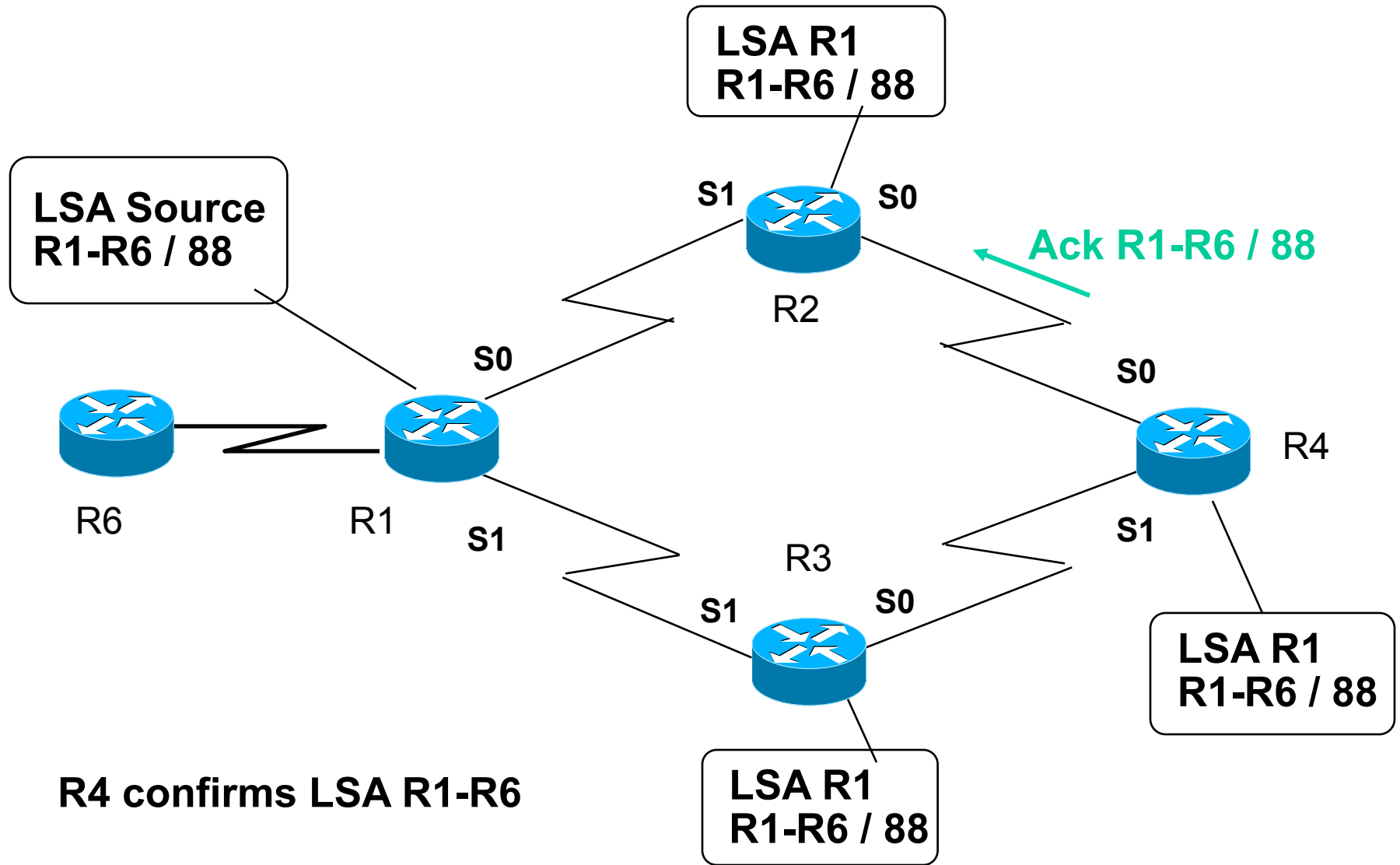
LSA Broadcast Example (2)



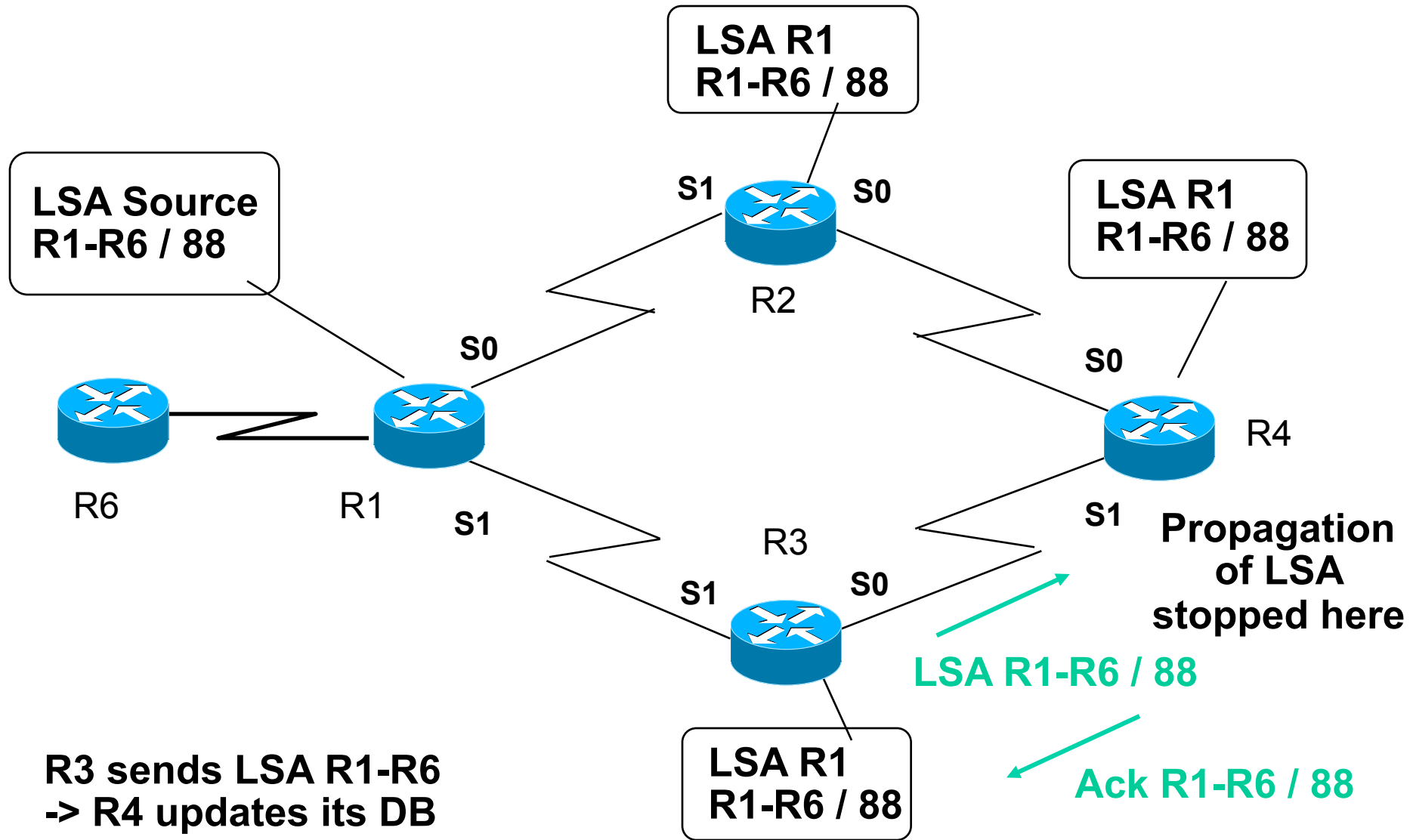
LSA Broadcast Example (3)



LSA Broadcast Example (4)



LSA Broadcast Example (5)



LSA Usage

- **Additionally, link states are repeated every 30 minutes to refresh the databases**
 - Link states – if not refreshed - become obsolete after 60 minutes and are removed from the databases
- **Reasons:**
 - Automatic correction of unnoticed topology-mistakes (e.g. happened during distribution or some router internal failures in the memory)
 - Combining two separated parts of an OSPF area (here OSPF also assures database consistency without intervention of an administrator)

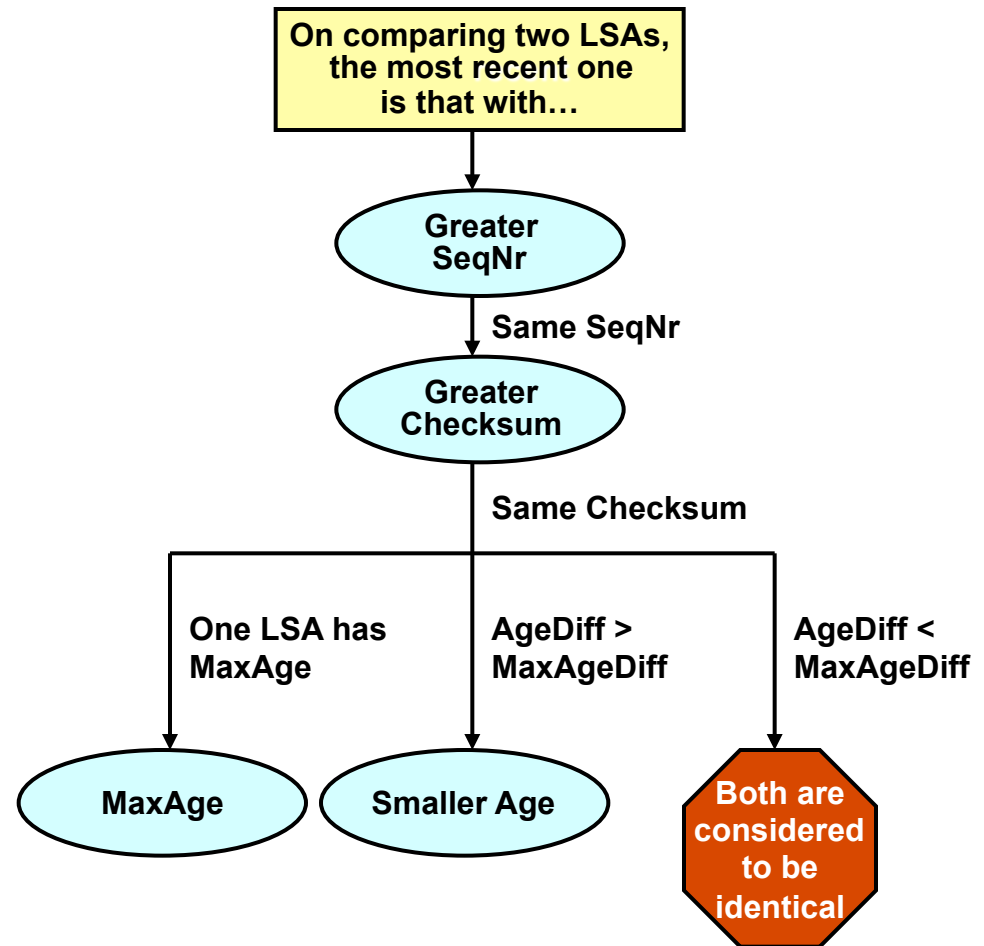
How are LSA unique?

- **Each router as a node in the graph (link state topology database)**
 - Is identified by a unique Router-ID
 - Note: automatically selected on Cisco routers
 - Either numerically highest IP address of all loopback interfaces
 - Or if no loopback interfaces then highest IP address of physical interfaces
- **Every link and hence LS between two routers**
 - Can be identified by the combination of the corresponding Router-IDs
 - Note:
 - If there are several parallel physical links between two routers the Port-ID will act as tie-breaker

Detailed Flooding Decisions

FYI

- **LSA is identified by its**
 - LS type
 - Link State ID
 - Advertising Router
- **The most recent one of two instances of the same LSA is determined by:**
 - LS sequence number
 - LS checksum
 - LS age
- **MaxAgeDiff (15 min) as tolerance value**



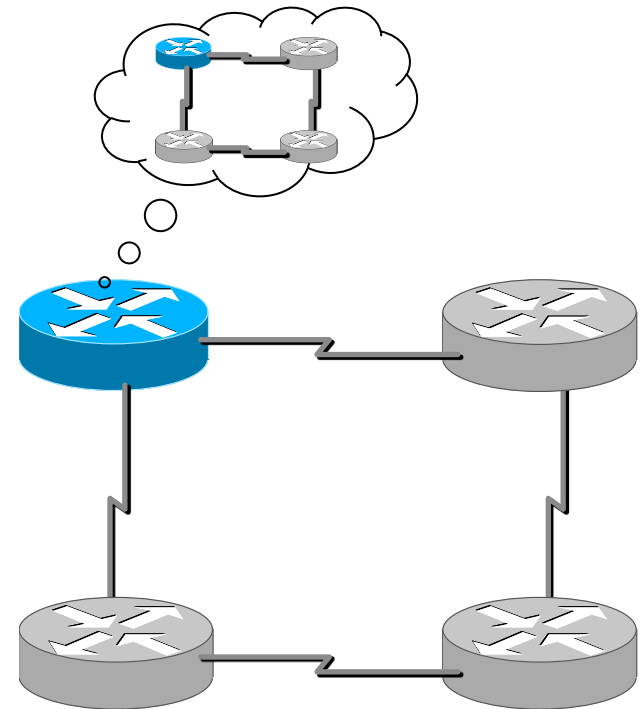
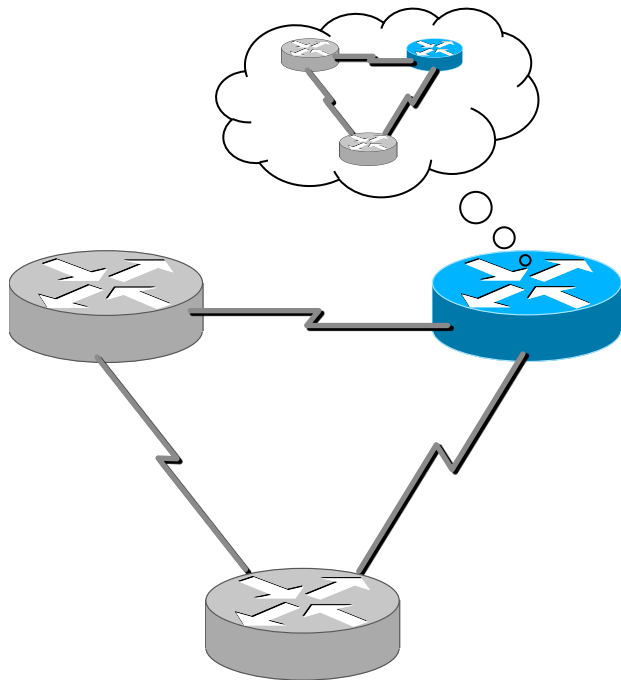
- **Originating router sets LS age = 0 seconds**
- **Increased during flooding by InfTransDelay by every router**
- **Also increased while stored in database**
- **Age is never incremented past MaxAge (60 min)**
- **LSAs having MaxAge:**
 - Are not used in routing table calculation anymore
 - Are reflooded immediately
 - Are always considered as most recent
 - Thus quickly flushed from routing domain
- **Responsible router maintains LSRefreshTime (30 min) to refresh LSAs periodically**

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
 - Introduction
 - The Dijkstra Algorithm
 - Communication Procedures
 - LSA Broadcast Handling
 - Split Area
 - Broadcast Networks
 - Area Principles
 - Stub Areas
 - Route Summarization
 - Virtual Link
 - Summary
 - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

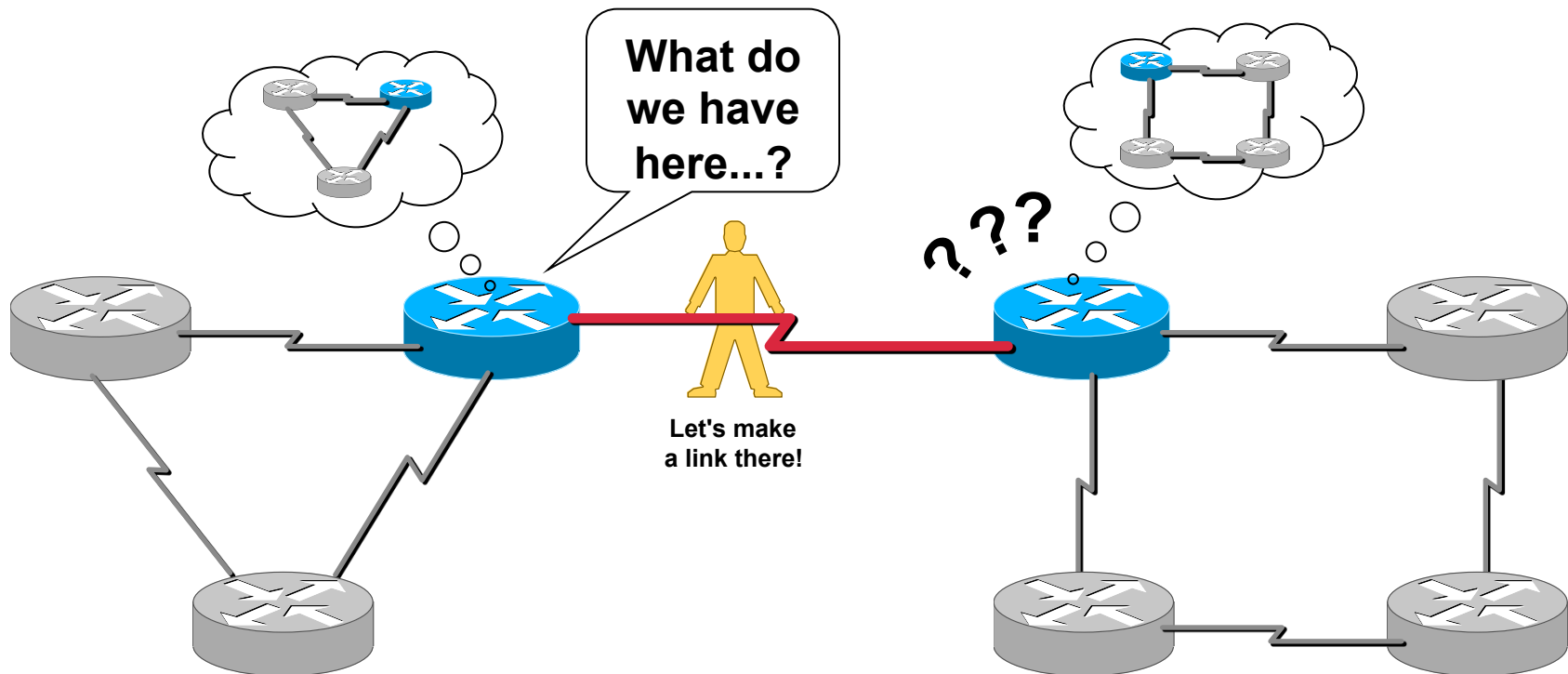
Basic Principle (1)

- Consider two routers, lucky integrated in their own networks...



Basic Principle (2)

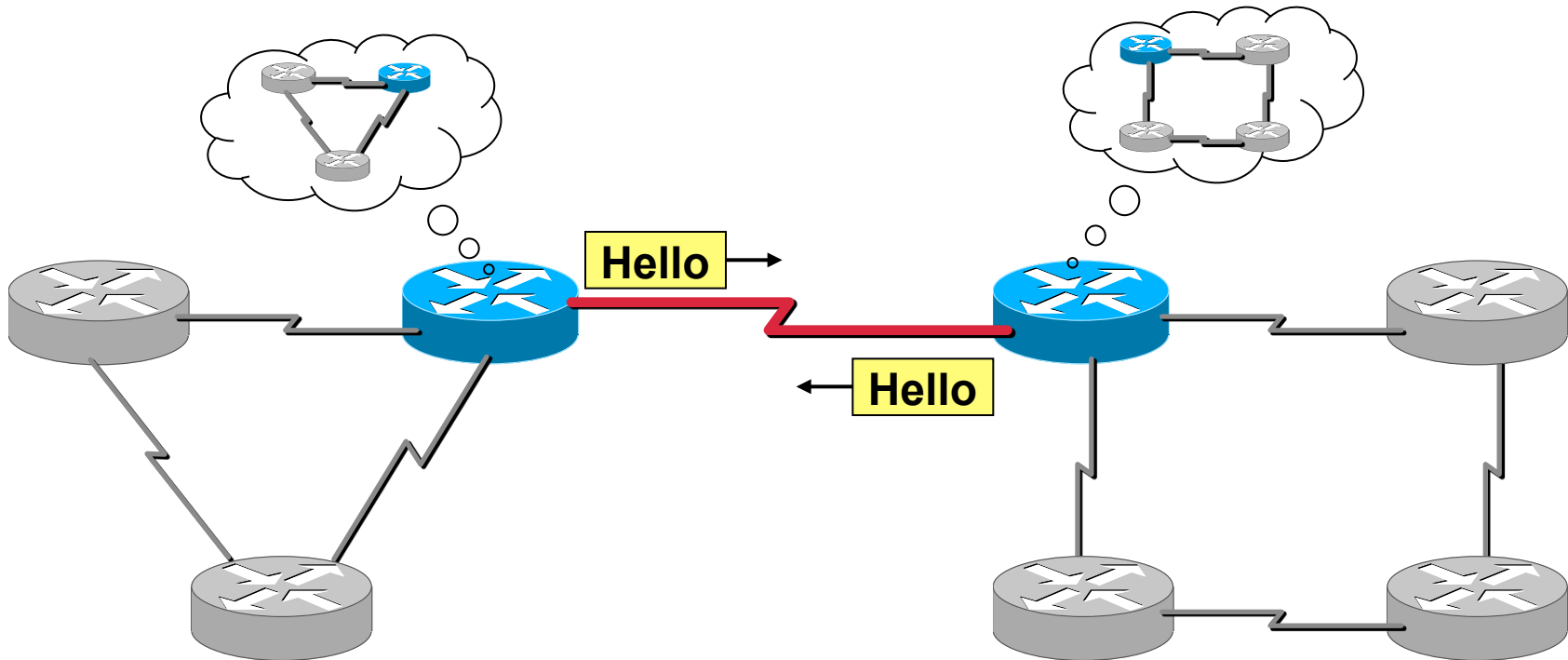
- Suddenly, some brave administrator connects them via a serial cable...
- Both interfaces are still in the "Down state"



Basic Principle (3)

- **Init state:**

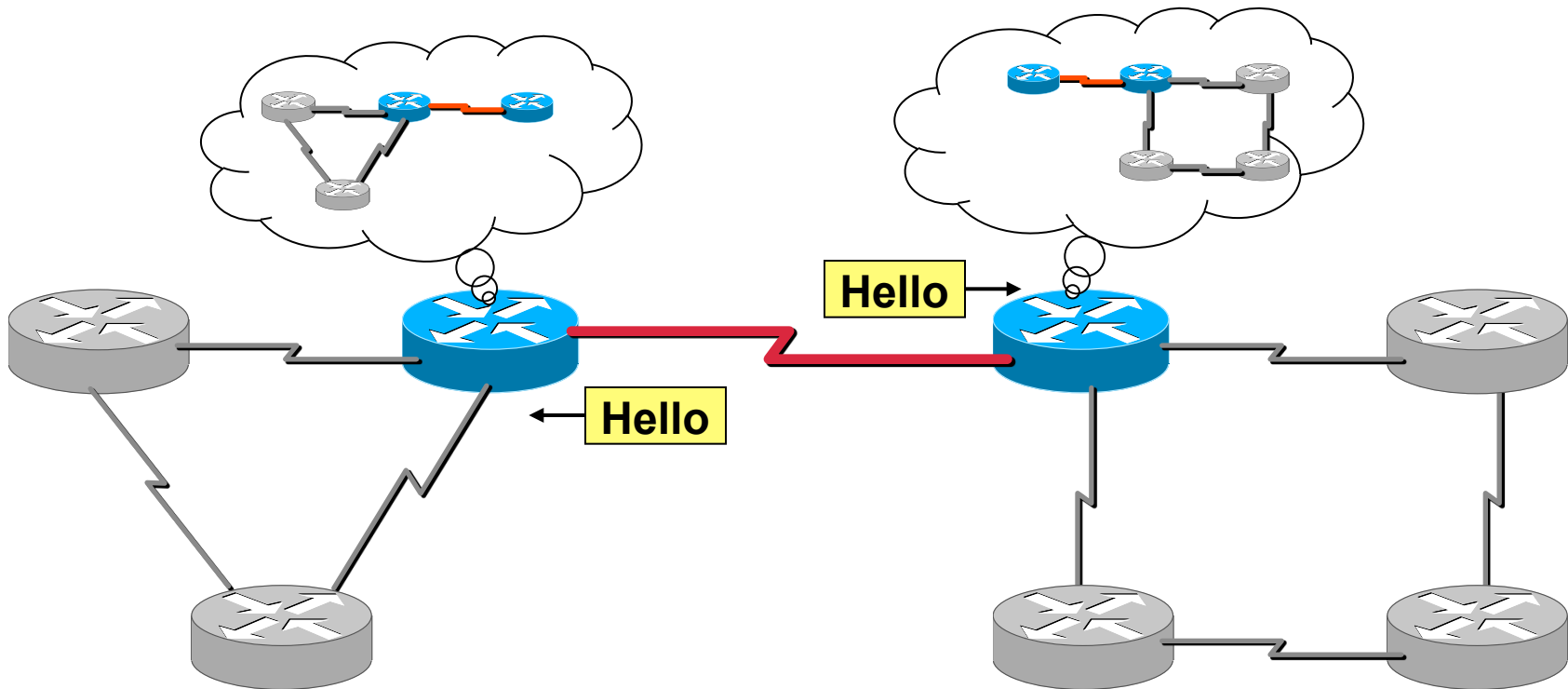
- Friendly as routers are, they welcome each other using the "Hello protocol"...



Basic Principle (4)

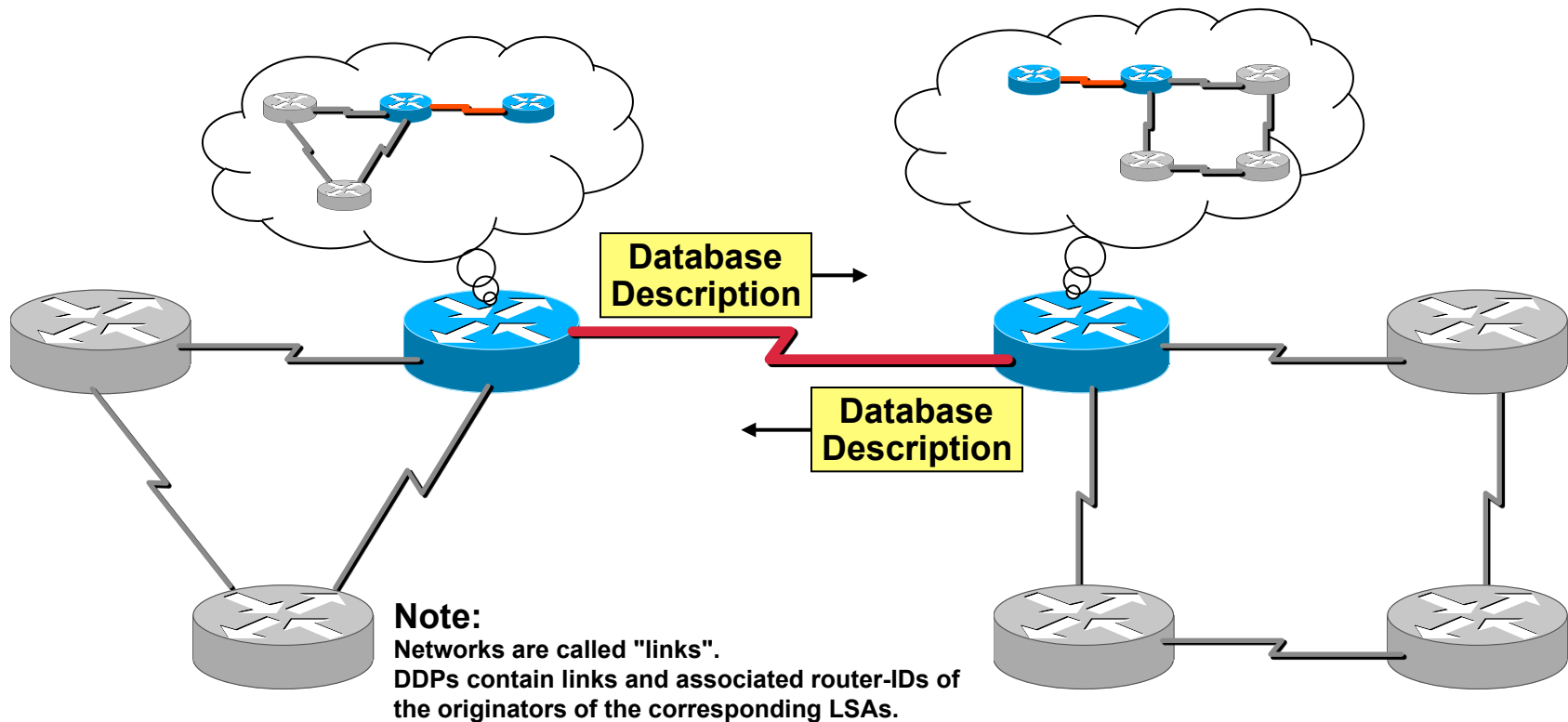
- **Two-way state:**

- Each Hello packet contains a list of all neighbors (IDs)
- Even the two routers themselves are now listed (=> 2-way state condition)
- Both routers are going to establish the new link in their database...



Basic Principle (5)

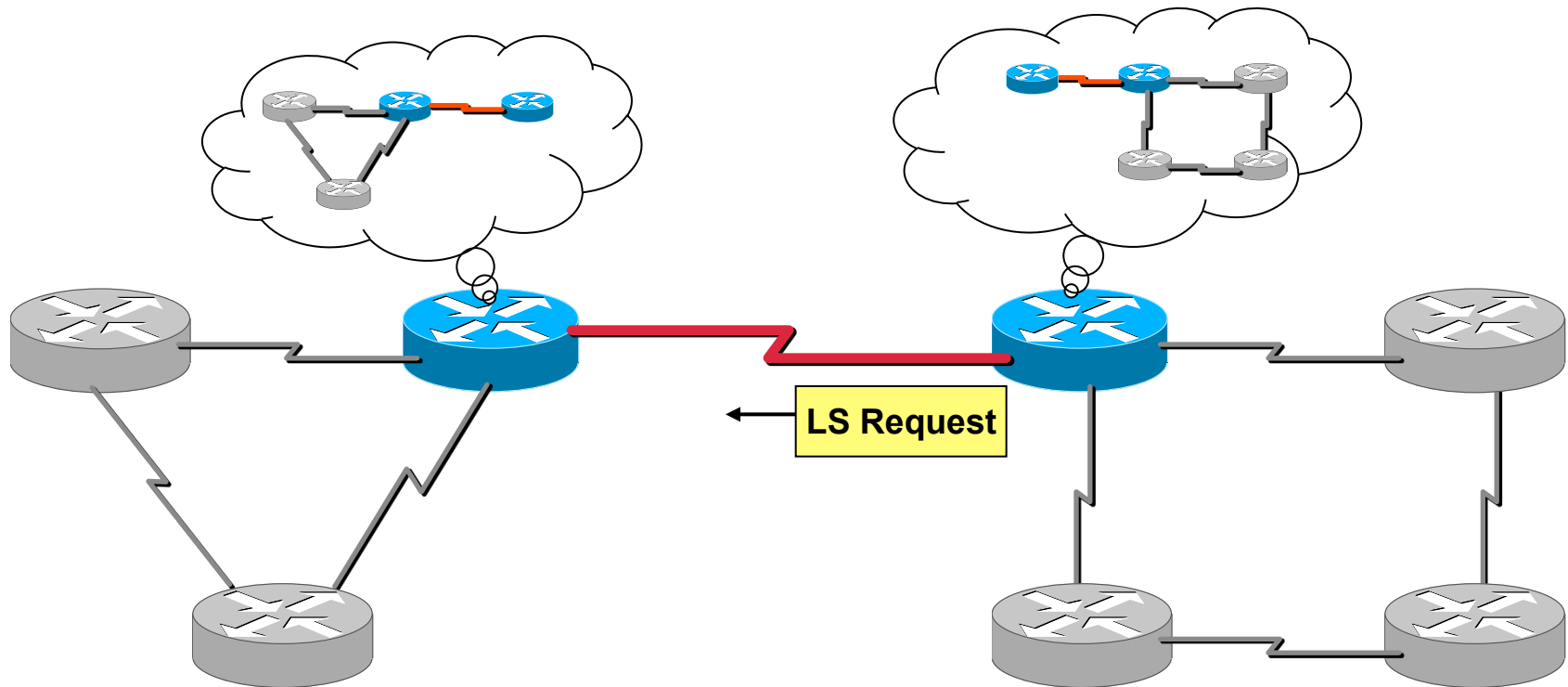
- **Exstart state:**
 - Determination of master (highest IP address) and slave
 - Needed for loading state later
- **Exchange state:**
 - Both routers start to offer a short version of their own roadmap, using "Database Description Packets" (DDPs)
 - DDPs contain partial LSAs, which summarize the links of every router in the neighbor's topology table.



Basic Principle (6)

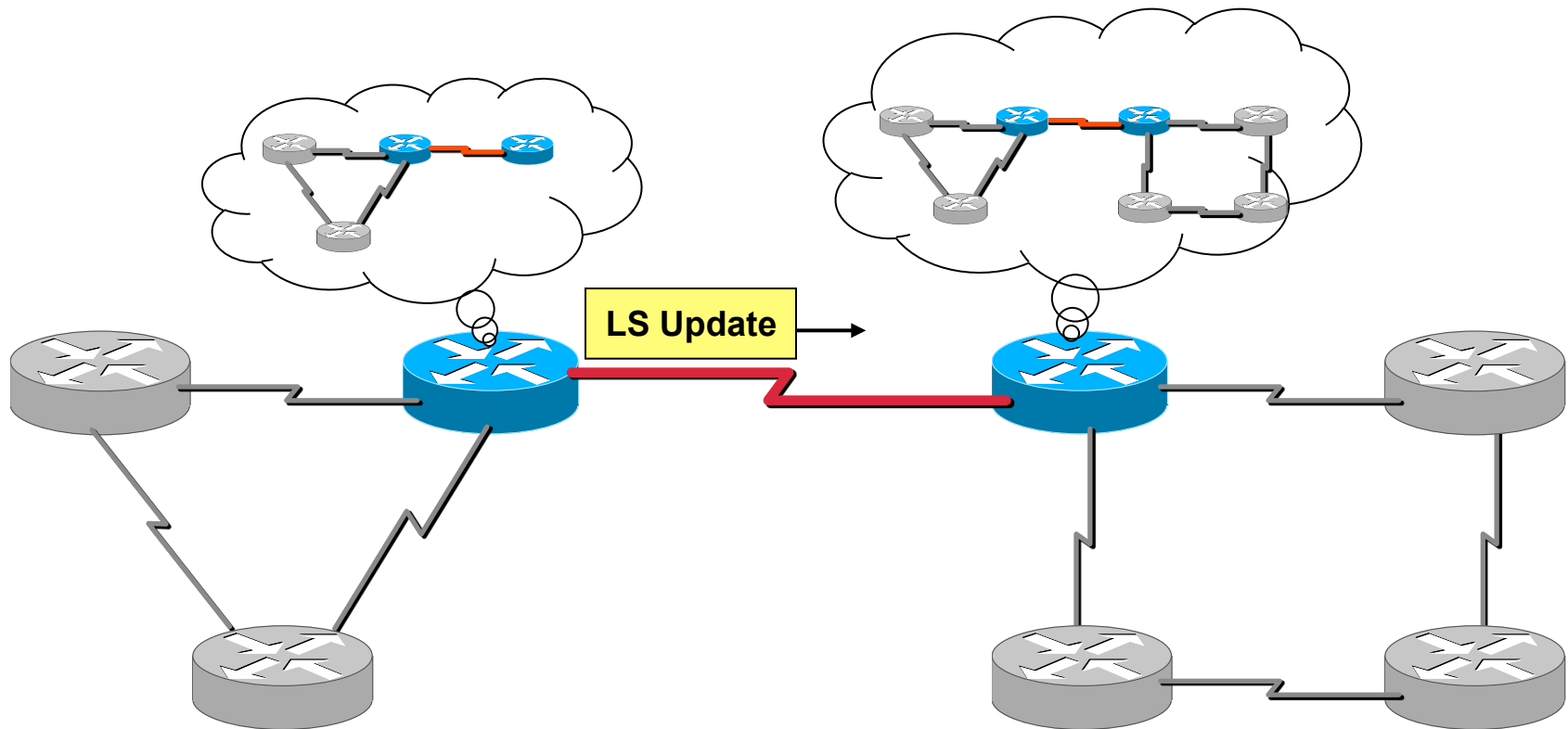
- **Loading State:**

- One router (here the right one) recognizes some missing links and asks for detailed information using a "Link State Request" (LSR) packet...



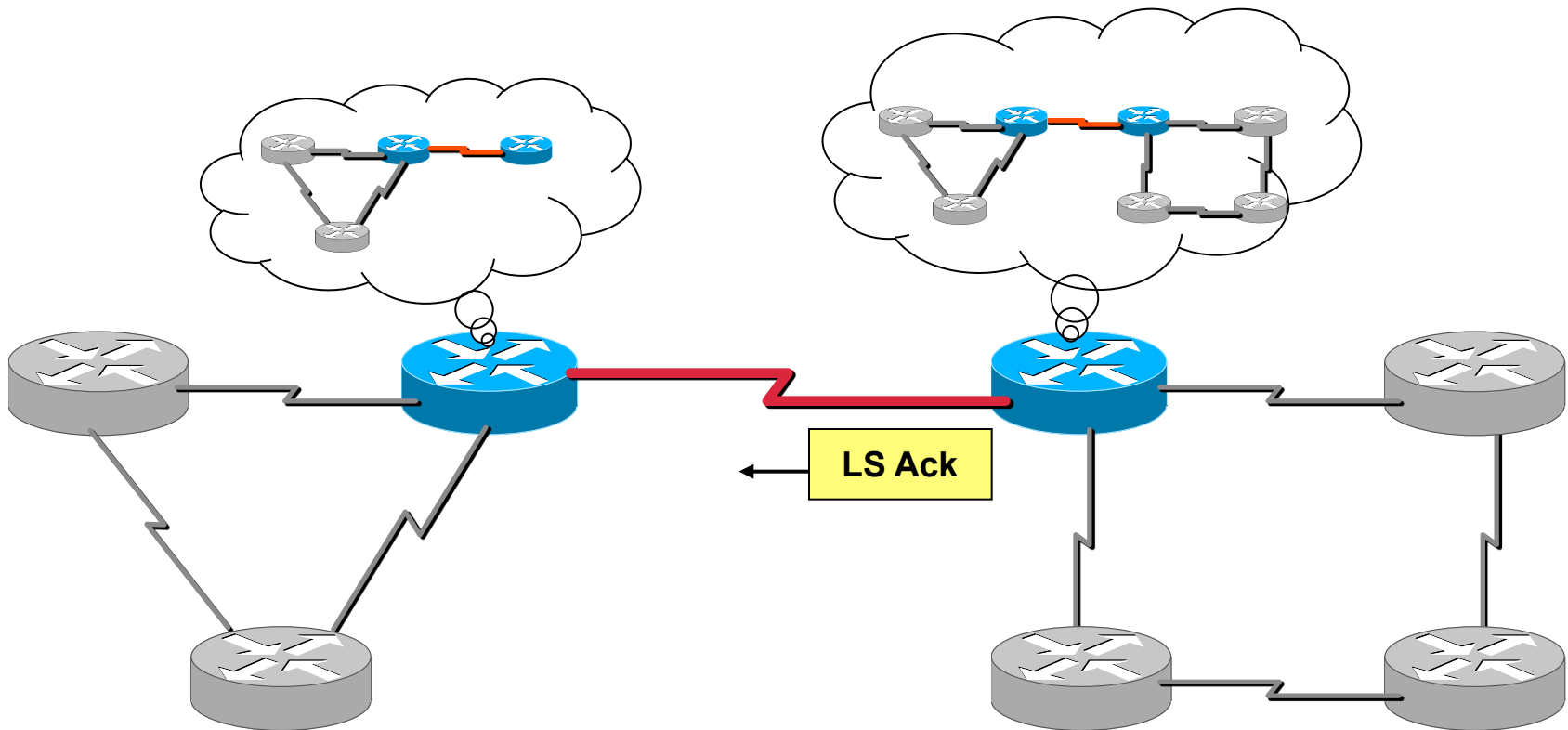
Basic Principle (7)

- The left router replies immediately with the requested link information, using a "Link State Update" (LSU) packet ...



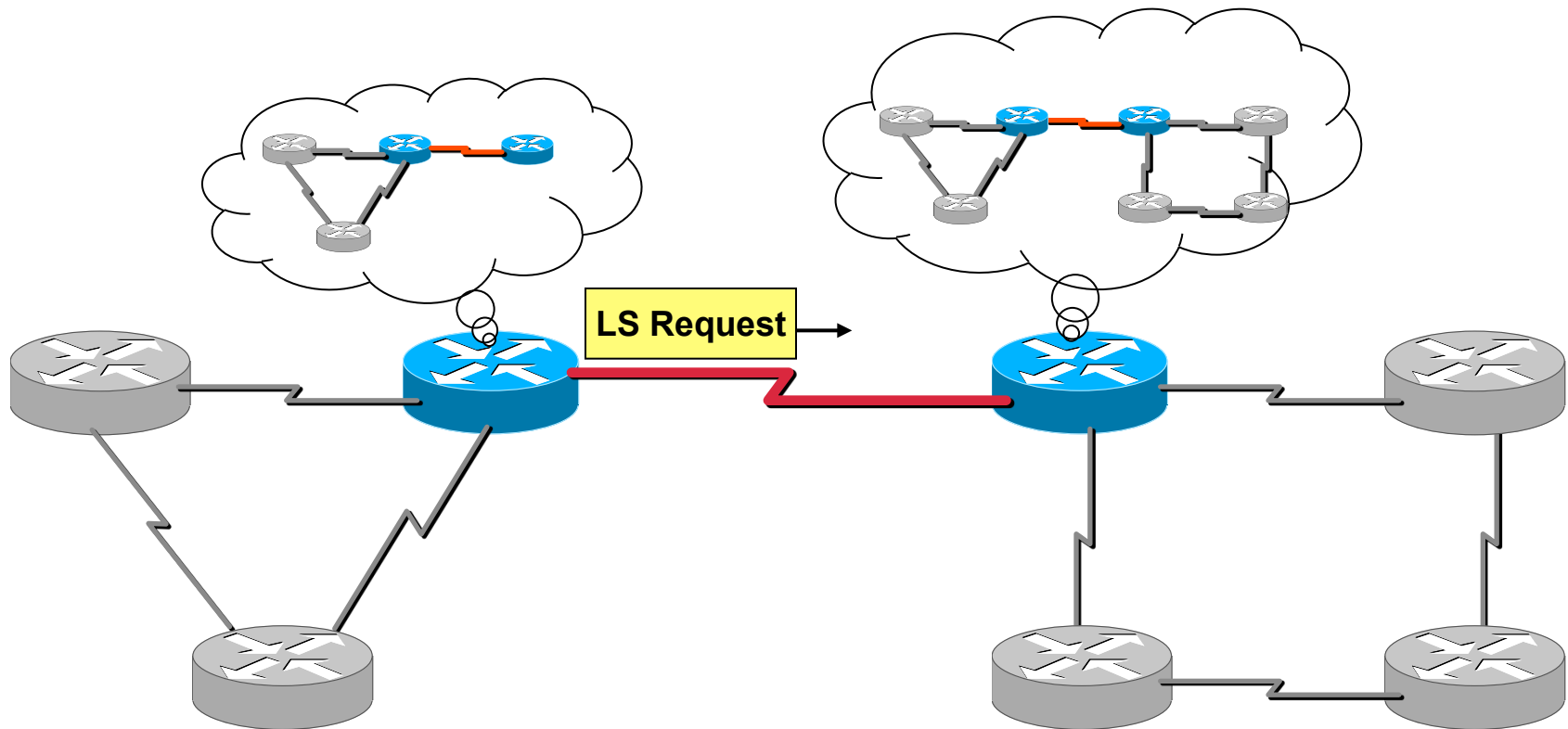
Basic Principle (8)

- The right router is very thankful, and returns a "Link State Acknowledgement"...



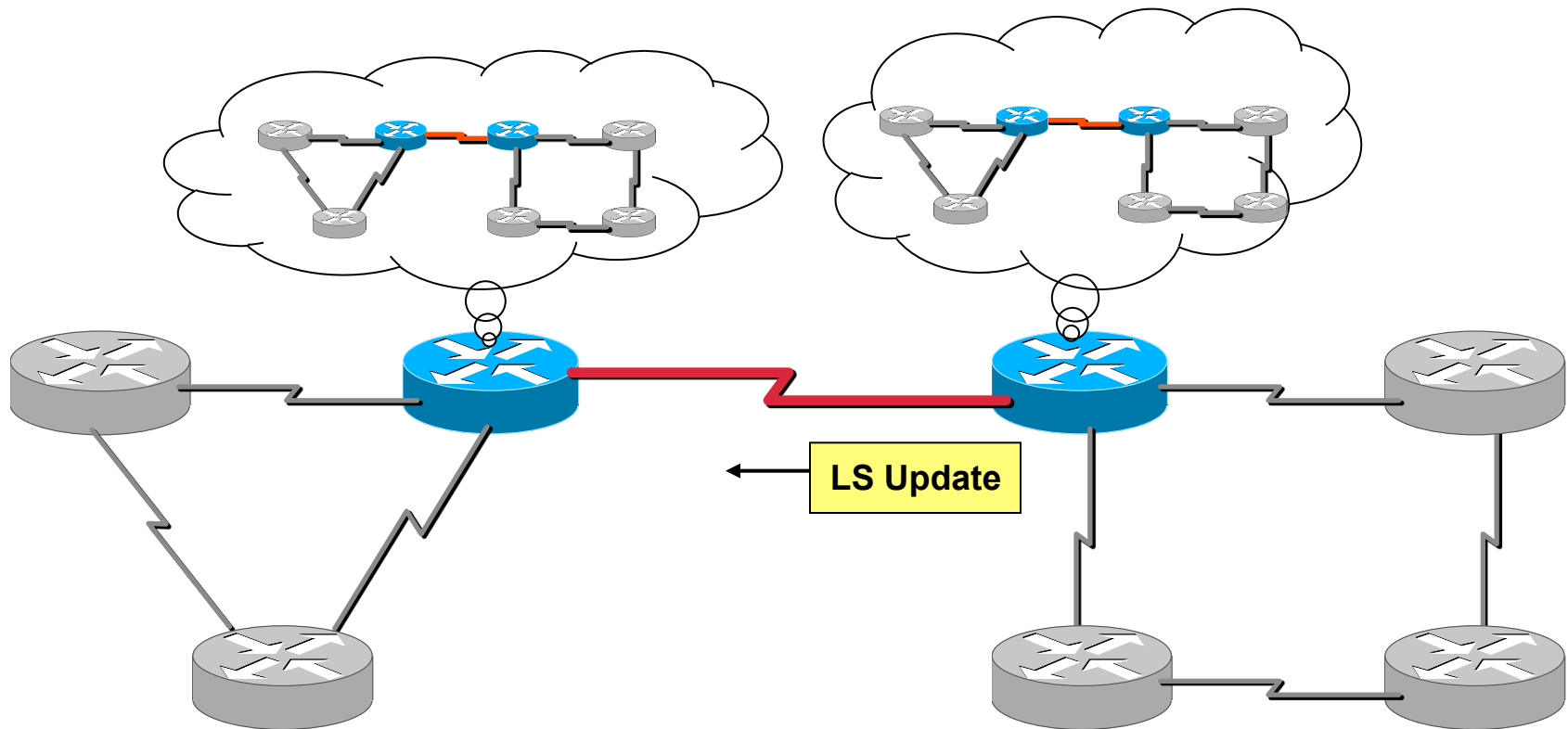
Basic Principle (9)

- Then the left router recognizes some unknown links and asks for further details...



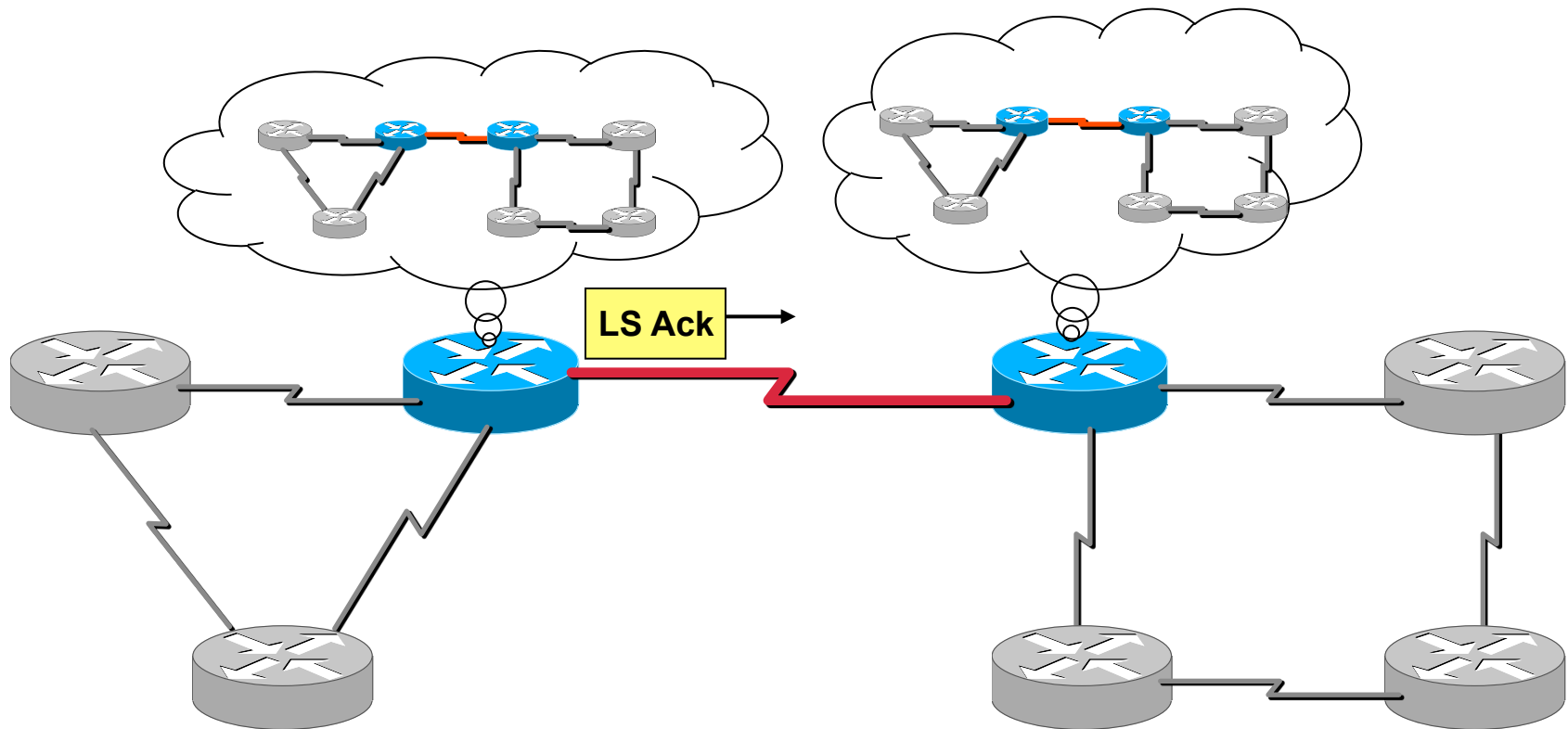
Basic Principle (10)

- The right router sends detailed information for the requested unknown links...



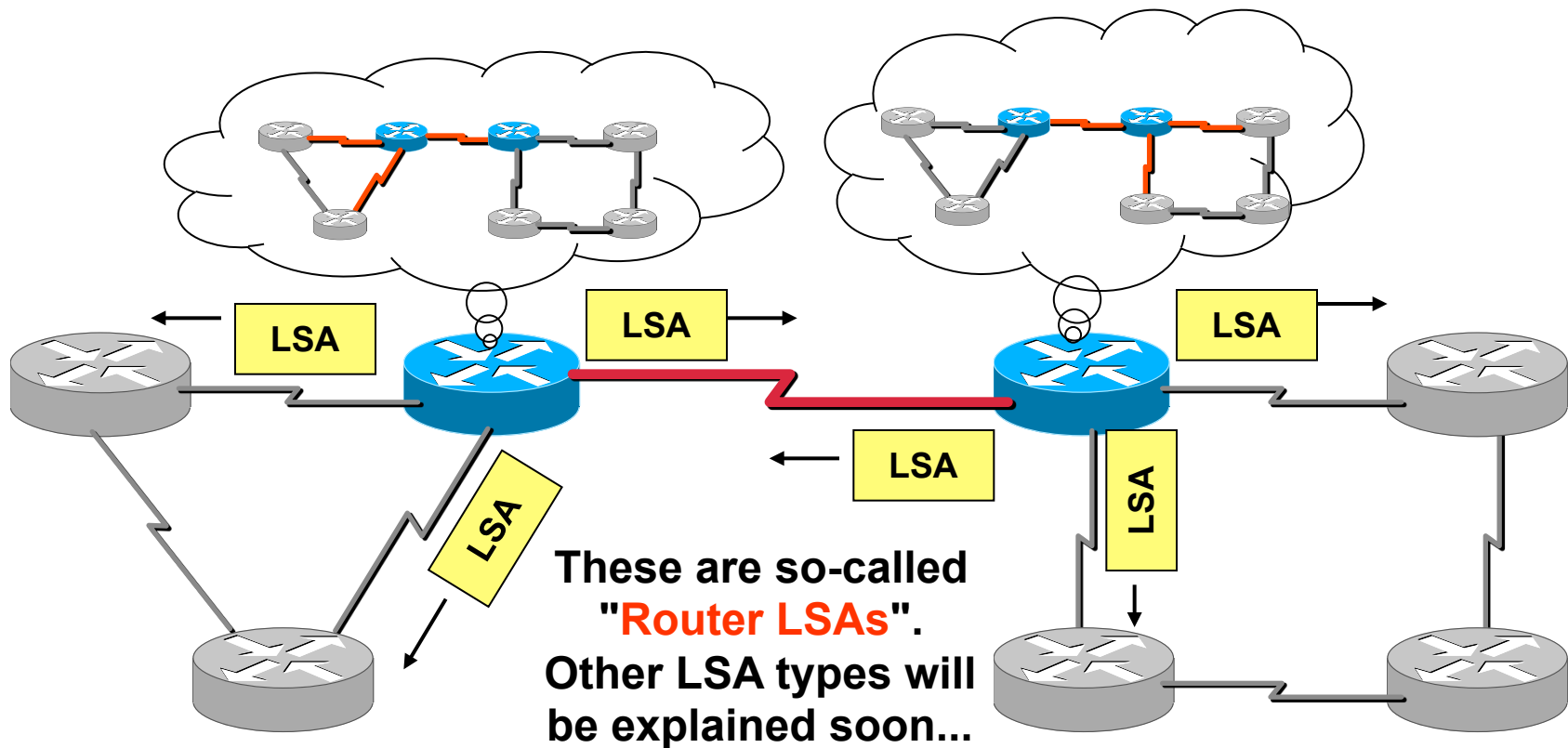
Basic Principle (11)

- The left router replies with a link state acknowledgement – **a new adjacency has been established...**
 - Neighbors are "fully adjacent" and reached the "full state"



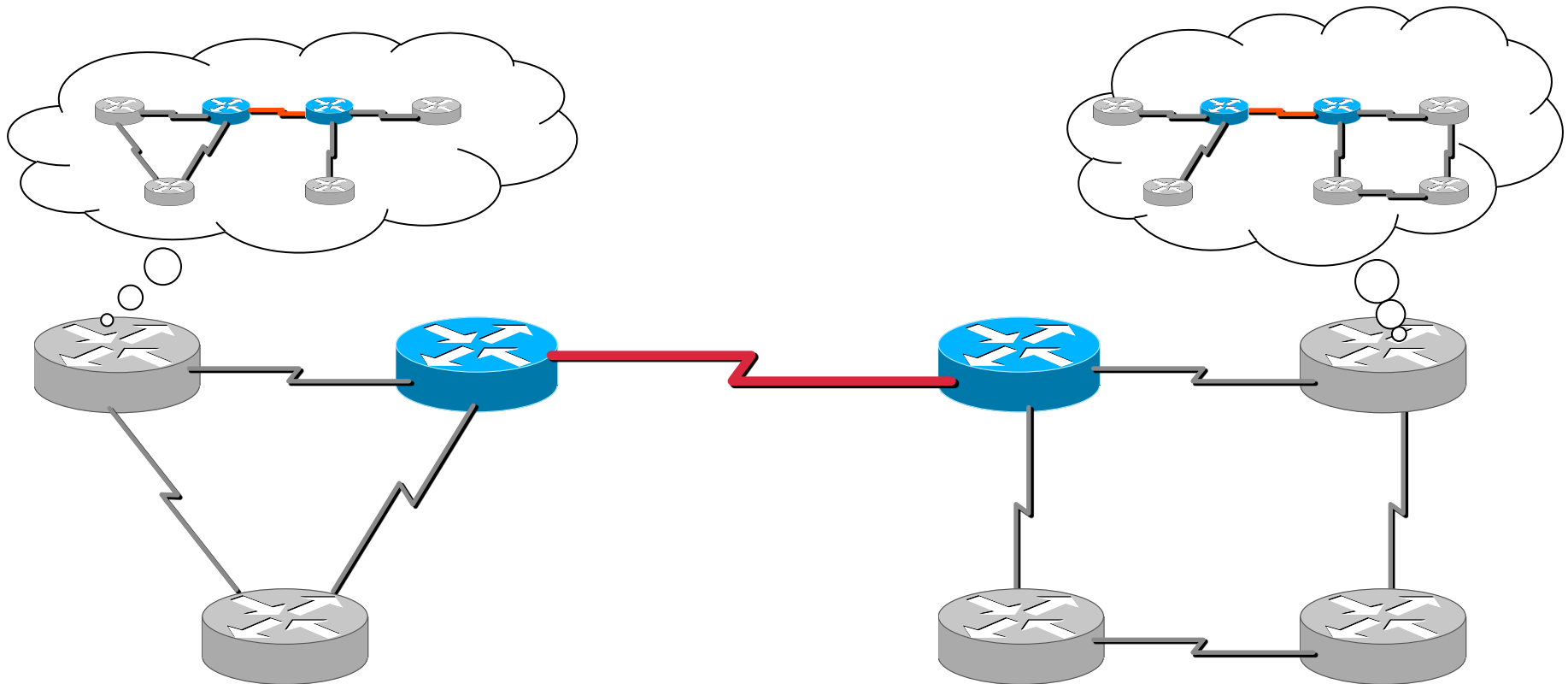
Basic Principle (12)

- Both routers tell all other routers about all local adjacencies by flooding link state advertisements (LSAs)
- Both routers now see their own IDs listed in the periodically sent Hello packets



Database Inconsistency

- When connecting two networks, LSA flooding only distributes information of the **local** links of the involved neighbors (!)



Healing Inconsistency (1)

- **Every router sends its LSAs every 30 minutes (!)**
 - Heals but long time of routing table / topology table inconsistency when combining a former split area of a OSPF domain
- **Triggering database synchronization between any two routers in the network**
 - In order to avoid long time of inconsistency
 - So whenever a router is informed by a Router-LSA about some changes in the network this router additionally will do a database synchronization with the router from which the Router-LSA was received
 - Database description packets will help to reduce traffic to the necessary minimum

Healing Inconsistency (2)

- **Optionally flash updates configured**
 - Upon receiving an LSA a router not only forwards this LSA but also immediately sends its own LSAs
 - Cisco default (can be turned off)

- **Golden OSPF design rule:**
 - Avoid splitting of an area in an OSPF environment by avoiding any single point of failures
 - Hence most parts of an area should be connected redundantly to each other

Finally: Convergence!

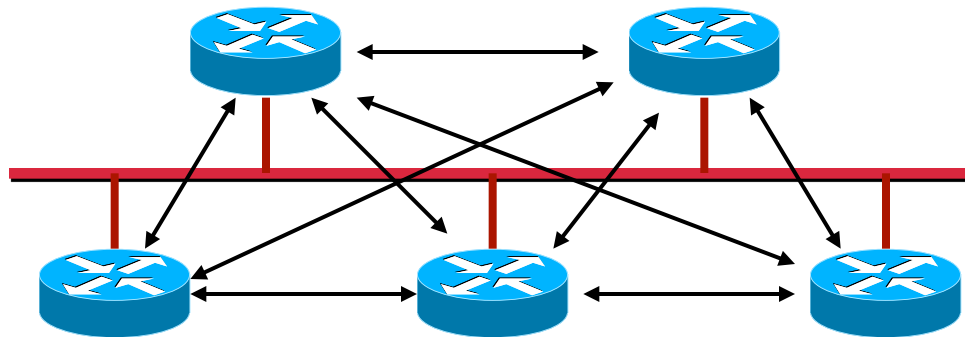
- **When LSAs are flooded, OSPF is quiet (at least for 30 minutes)**
- **Only Hello's are sent out on every interface to check adjacencies**
 - Topology changes are quickly detected
 - Default Hello interval: **10 seconds (LAN, 60 sec WAN)**
 - Hellos are terminated by neighbors

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
 - Introduction
 - The Dijkstra Algorithm
 - Communication Procedures
 - LSA Broadcast Handling
 - Split Area
 - Broadcast Networks
 - Area Principles
 - Stub Areas
 - Route Summarization
 - Virtual Link
 - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

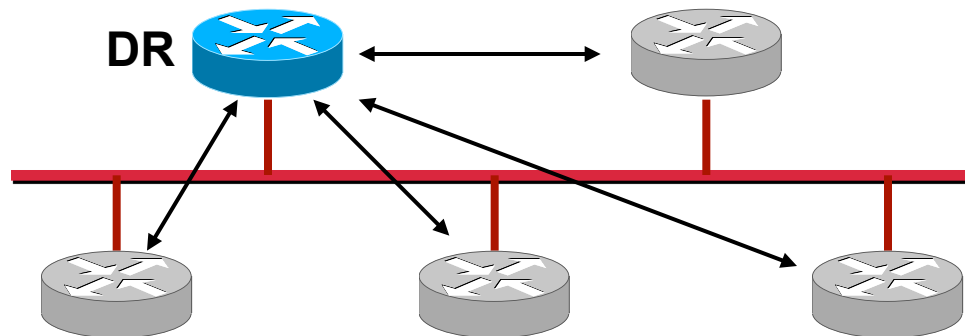
Broadcast Multi-Access Media (1)

- When several OSPF routers have access to the same Ethernet segment they would create $n(n-1)/2$ adjacencies
- Furthermore, SPF algorithm requires to represent a fully meshed network as **tree**



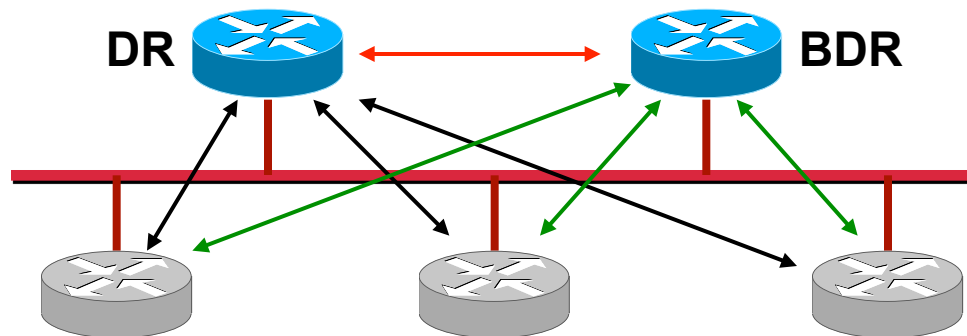
Broadcast Multi-Access Media (2)

- **Solution: Elect one "Designated Router" (DR)** to represent the whole LAN segment
 - Election uses the Hello protocol
- **DR sends Network LSA**
 - List of all local routers
 - Ensures that every router on the link has the same topology database
 - Also contains subnet mask (!)
- **Each other router establishes an adjacency only to the DR**
 - Using "All DR" multicast address 224.0.0.6



Broadcast Multi-Access Media (3)

- Only the DR will send LSAs to the rest of the network
- For backup purposes also a **Backup DR** is elected (**BDR**)
 - All routers also establish adjacencies to the BDR
 - BDR itself also establishes adjacency to DR



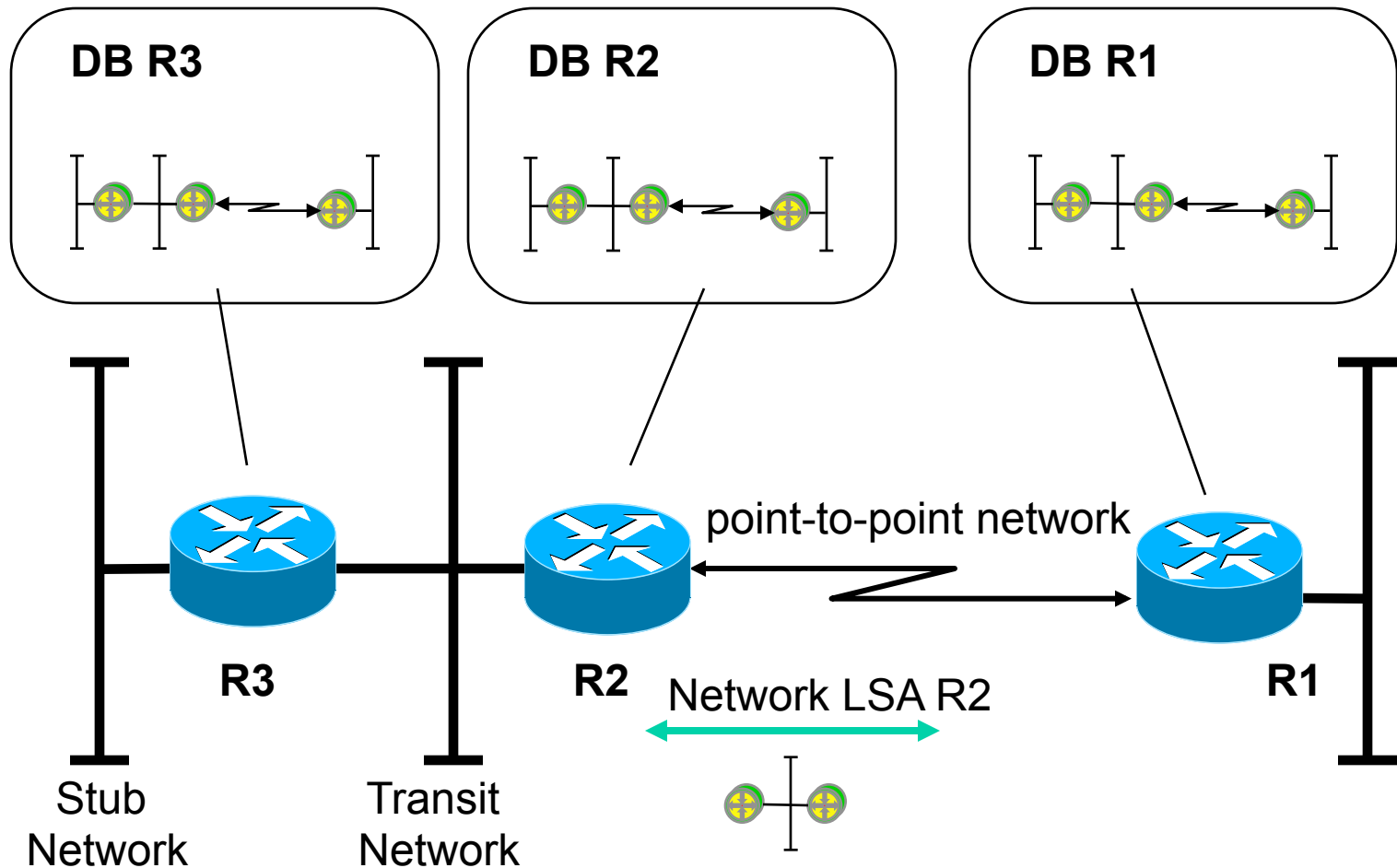
DR/BDR Election Process

- **Election process starts if no DR/BDR listed in the hello packets during the init state (i. e. when two routers begin to establish an adjacency)**
 - Note: if already one DR/BDR chosen, any new router in the LAN would not change anything!
 - Therefore, the power-on order of routers is critical !!!
- **Always configure loopback interface in order to "name" your routers**
 - Loopback interface never goes down
 - Ensures stability
 - Simple to manage

DR, Router LSA, Network LSA

- **Designated Router (DR) is responsible**
 - For maintaining neighbourhood relationship via virtual point-to-point links using the already known mechanism
 - DB-Description, LS-Request LS-Update, LS-Acknowledgement, Hello, etc.
- **Router-LSA implicitly describes**
 - These virtual point-to-point links by specifying such a network as transit-network
 - Remark: Stub-network is a LAN network where no OSPF router is behind
- **To inform all other routers of domain about such a special topology situation**
 - DR is additionally responsible for emitting Network LSAs
- **Network LSA describes**
 - Which routers are members of the corresponding broadcast network

OSPF Network LSA R2



Designated Router R2 notifies other nodes about the multi-access network using Network-LSA (transport mechanism are LS-Update packets hop-by-hop)

Details: OSPF Multicast Usage

- **OSPF uses dedicated IP multicast addresses for exchanging routing messages**
 - 224.0.0.5 (“All OSPF Routers”)
 - 224.0.0.6 (“All Designated Routers”)
- **224.0.0.5 is used as destination address**
 - By all routers for Hello-messages
 - DR and BR determination at start-up
 - link state supervision
 - By DR router for messages towards all non-DR routers
 - LS-Update, LS-Acknowledgement
- **224.0.0.6 is used as destination address**
 - By all non-DR routers for messages towards the DR
 - LS-Update, LS-Request, LS-Acknowledgement and database description messages

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
 - Introduction
 - The Dijkstra Algorithm
 - Communication Procedures
 - LSA Broadcast Handling
 - Split Area
 - Broadcast Networks
 - Area Principles
 - Stub Areas
 - Route Summarization
 - Virtual Link
 - Summary
 - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

OSPF Domain / OSPF Area

- **OSPF domain can be divided in multiple OSPF areas**
 - To improve performance
 - To decouple network parts from each other
- **Performance improvement**
 - By restricting Router-LSA and Network-LSA to the originating area
 - Note: receiving a Router-LSA will cause the SPF algorithm to be performed
- **Decoupling is actually done**
 - By route summarization enabled through the usage of classless routing and careful IP address plan

OSPF Domain / OSPF Area

- **Every area got its own topology database**
 - Which is unknown to other areas
 - Area specific routing information stays inside this area
- **On topology changes**
 - Routing traffic causing Dijkstra's algorithm to be performed stays inside the area where the change appears
 - Route summarization reduces routing traffic drastically
- **OSPF areas are labelled with area-IDs**
 - Unique within the OSPF domain
 - Written in IP address like format or just as number
- **An OSPF domain contains**
 - At least one single area or several areas

OSPF Area Border Router

- **OSPF areas are connected by special routers**
 - Area Border Router (ABR)
- **ABR**
 - Maintains a topology database for each area he is connected to
- **All OSPF areas must be connected over a special area**
 - Backbone Area
 - Area-ID = 0.0.0.0 or area-ID = 0
 - If there is only one area in the OSPF domain this OSPF area will be the backbone area

OSPF Backbone Area

- **Non-backbone areas must not be connected directly**
 - Connection allowed only via Backbone Area
- **This OSPF rule forces**
 - A star-like topology of areas with the backbone area in the centre
- **ABRs**
 - Are connected to the backbone area by direct physical links in normal cases
 - Exception with virtual link technique if direct physical link can not be provided
 - A virtual link can be used to "tunnel" the routing traffic between an isolated area and the backbone area through another area

- **OSPF provides three types of routing:**
 - Intra-area routing:
 - Inside of an area (using Level 1 Router; Internal Router IR)
 - Router Link LSA (LSA type1)
 - Network Link LSA (LSA type2)
 - Note: Backbone Router is a Backbone Area Internal Router

 - Inter-area routing:
 - Between areas over a Backbone Area (using Area Border)
 - Summary Link LSA (LSA type3 and type4)
 - Type 3 to announce networks
 - Type 4 to announce IP address of ASBRs

- **OSPF provides three types of routing (cont.):**
 - Exterior routing:
 - Paths to external destinations (other AS) are configured statically or imported with EGP or BGP using Autonomous Systems Boundary Routers (ASBRs)
 - AS External Summary LSA (LSA type5) to announce external networks

Area Border Router

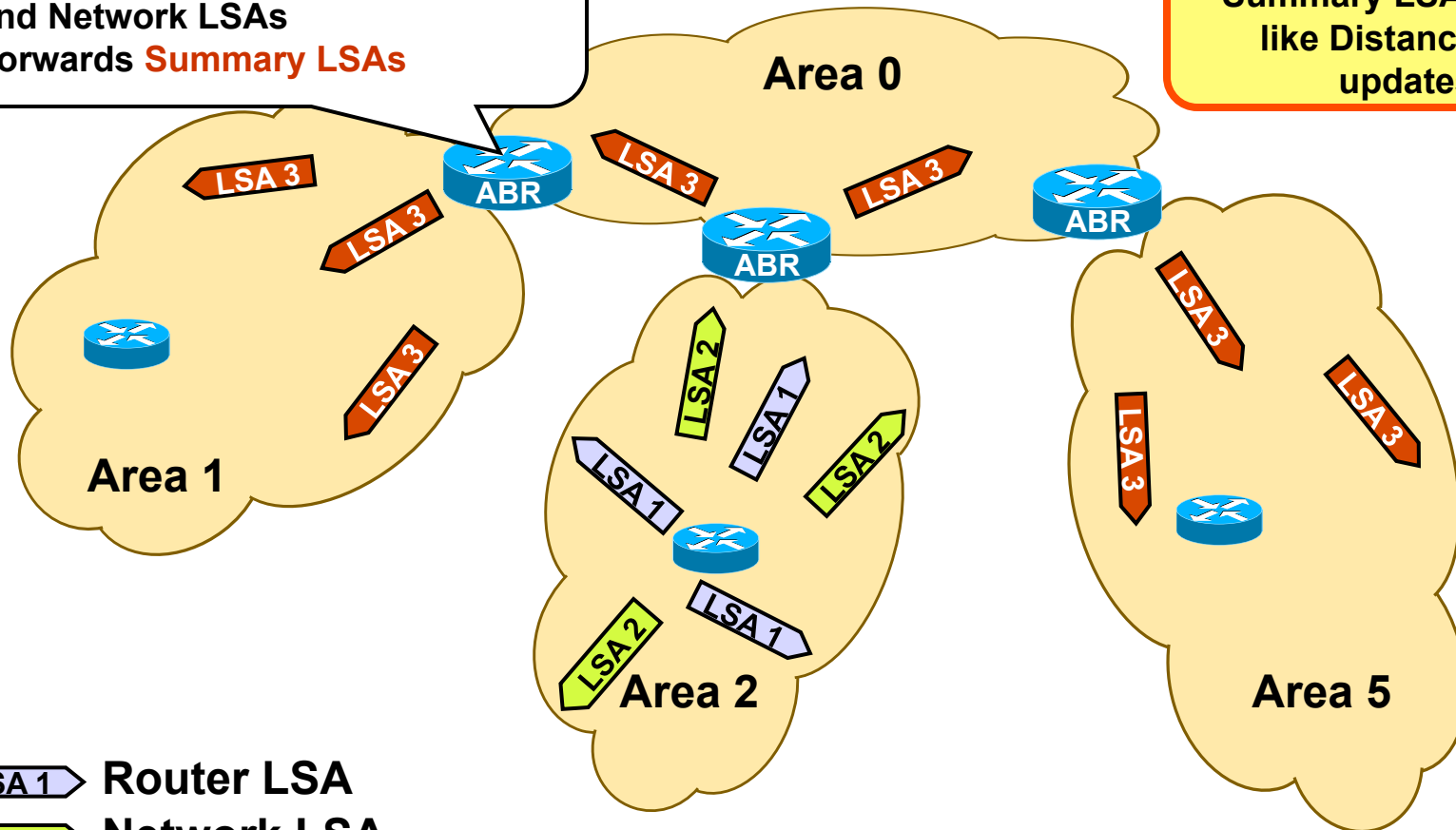
- **Area Border Router maintains two topology maps**
 - One for its area
 - One for the Backbone Area
- **Area Border Router exports the routes of its area to the Backbone Area**
 - Collects all topology information of its area and sends Summary LSAs to the Backbone Area
- **Area Border Router imports all routes of other areas (received from the backbone area) in its own area**
 - This is done again using Summary LSAs

ABR

Area Border Router (ABR):

Terminates Router LSAs
and Network LSAs
Forwards **Summary LSAs**

Note:
Summary LSAs behaves
like Distance Vector
updates !!!

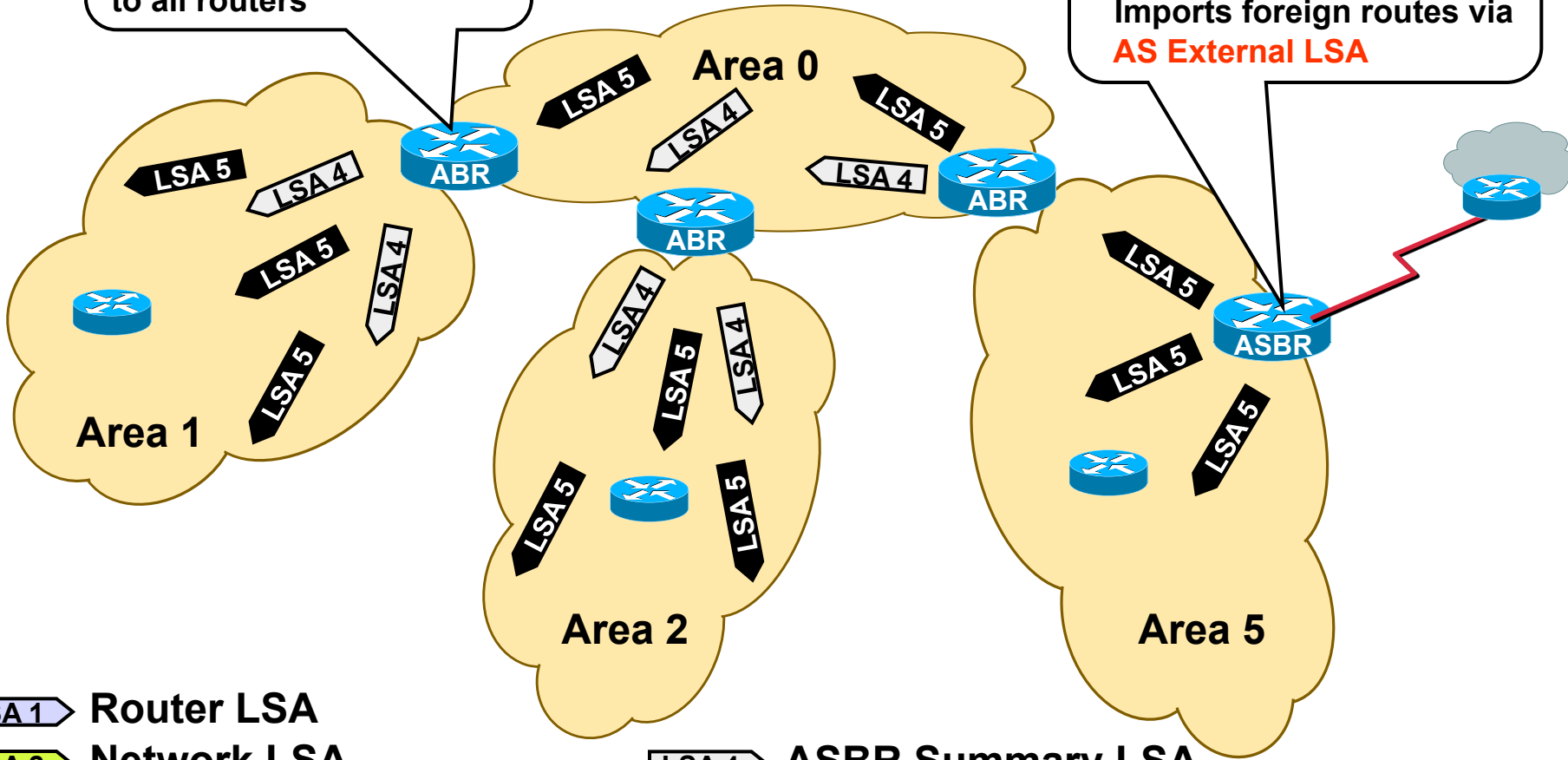


- LSA 1 Router LSA
- LSA 2 Network LSA
- LSA 3 Summary LSA

ASBR

When an ABR receives an AS External LSA it emits **ASBR Summary LSAs** to all routers

Autonomous System Border Router (ASBR)
Imports foreign routes via **AS External LSA**



- LSA 1 Router LSA
- LSA 2 Network LSA
- LSA 3 Summary LSA

- LSA 4 ASBR Summary LSA
- LSA 5 AS External LSA

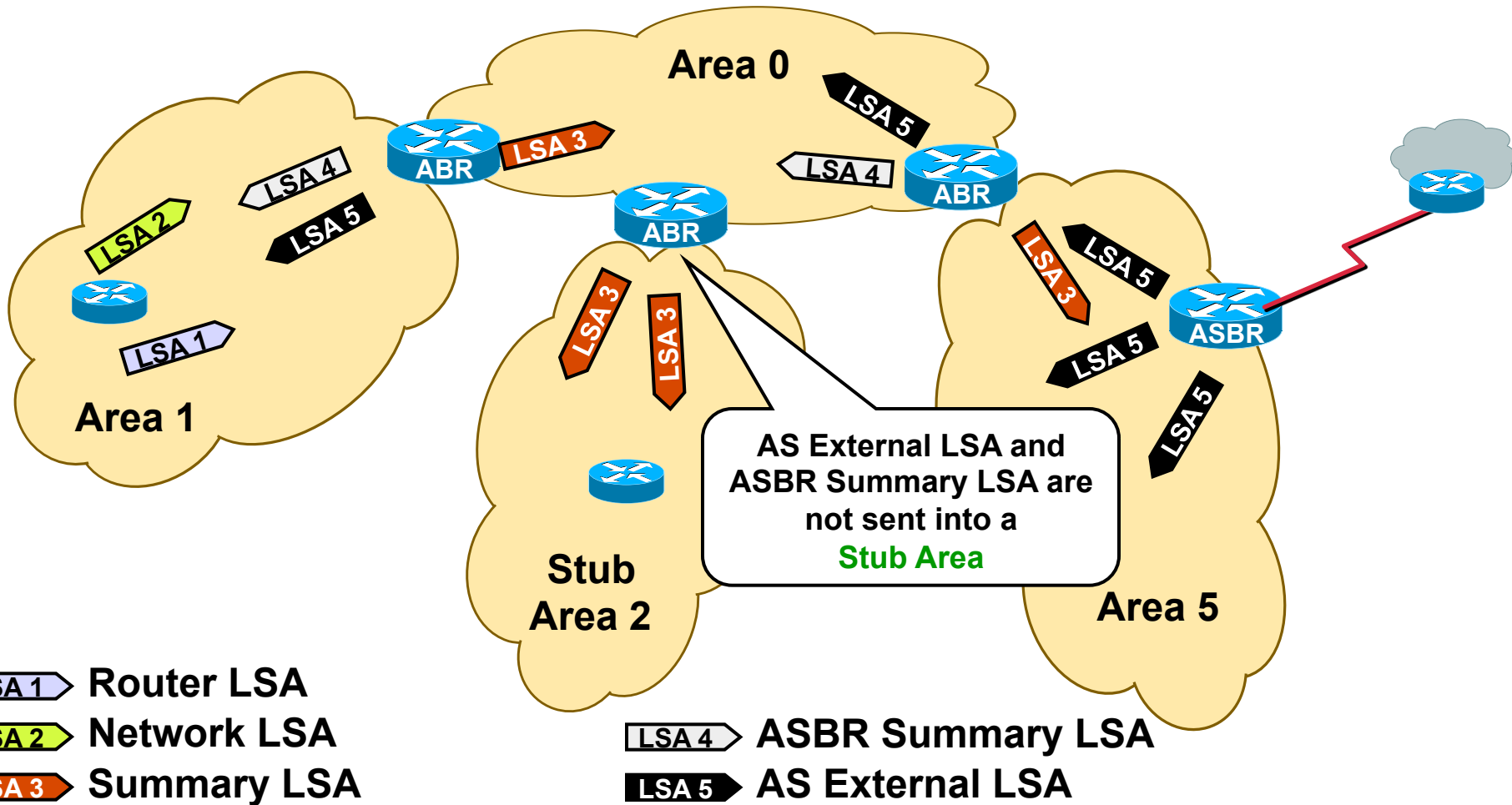
Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
 - Introduction
 - The Dijkstra Algorithm
 - Communication Procedures
 - LSA Broadcast Handling
 - Split Area
 - Broadcast Networks
 - Area Principles
 - Stub Areas
 - Route Summarization
 - Virtual Link
 - Summary
 - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

OSPF Stub Areas

- **Normally, every internal router gets information about all networks**
 - Internal and external NET-IDs
- **OSPF allows definition of Stub Areas**
 - To minimize memory requirements of internal routers of non-backbone areas for external networks
 - Only the Area Border Router of a particular area knows all external destinations
 - Internal routers only get a default route entry (to this Area Border Router)
 - Any traffic that do not stay inside the OSPF domain (external networks) is forwarded to the Area Border Router

Stub Area



OSPF Totally Stubby Areas

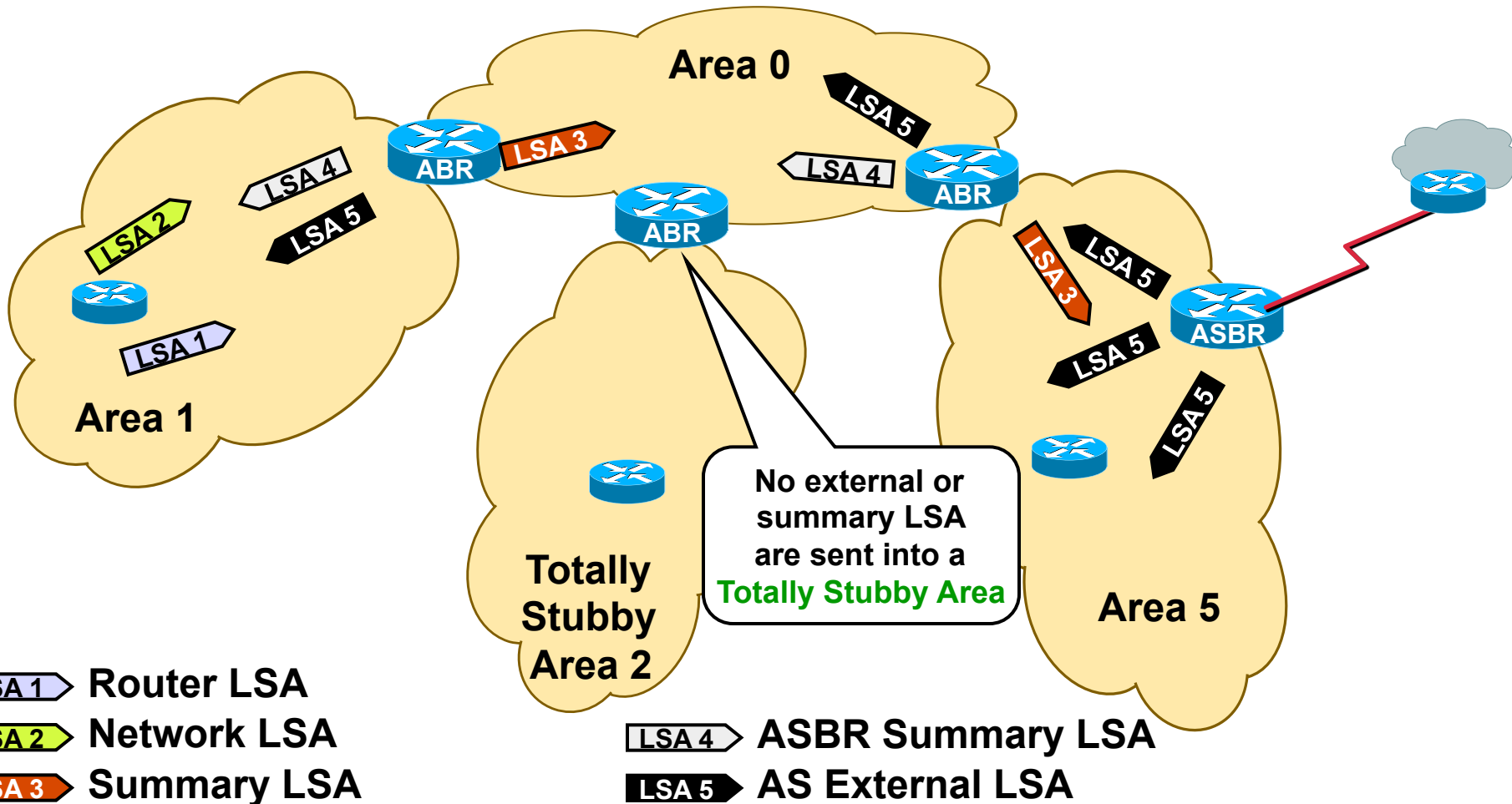
FYI

- **Cisco allows definition of Totally Stubby Areas**
 - Internal routers follow default route also for networks of other areas (no Summary-LSA)
 - That means for internal networks of other areas
- **In such an area**
 - ASBRs are forbidden
- **But if an ASBR should be located in such as totally stubby area**
 - NSSA (Not So Stubby Area) functionality can be used using LSA type 7 updates.

Totally Stubby Area

FYI

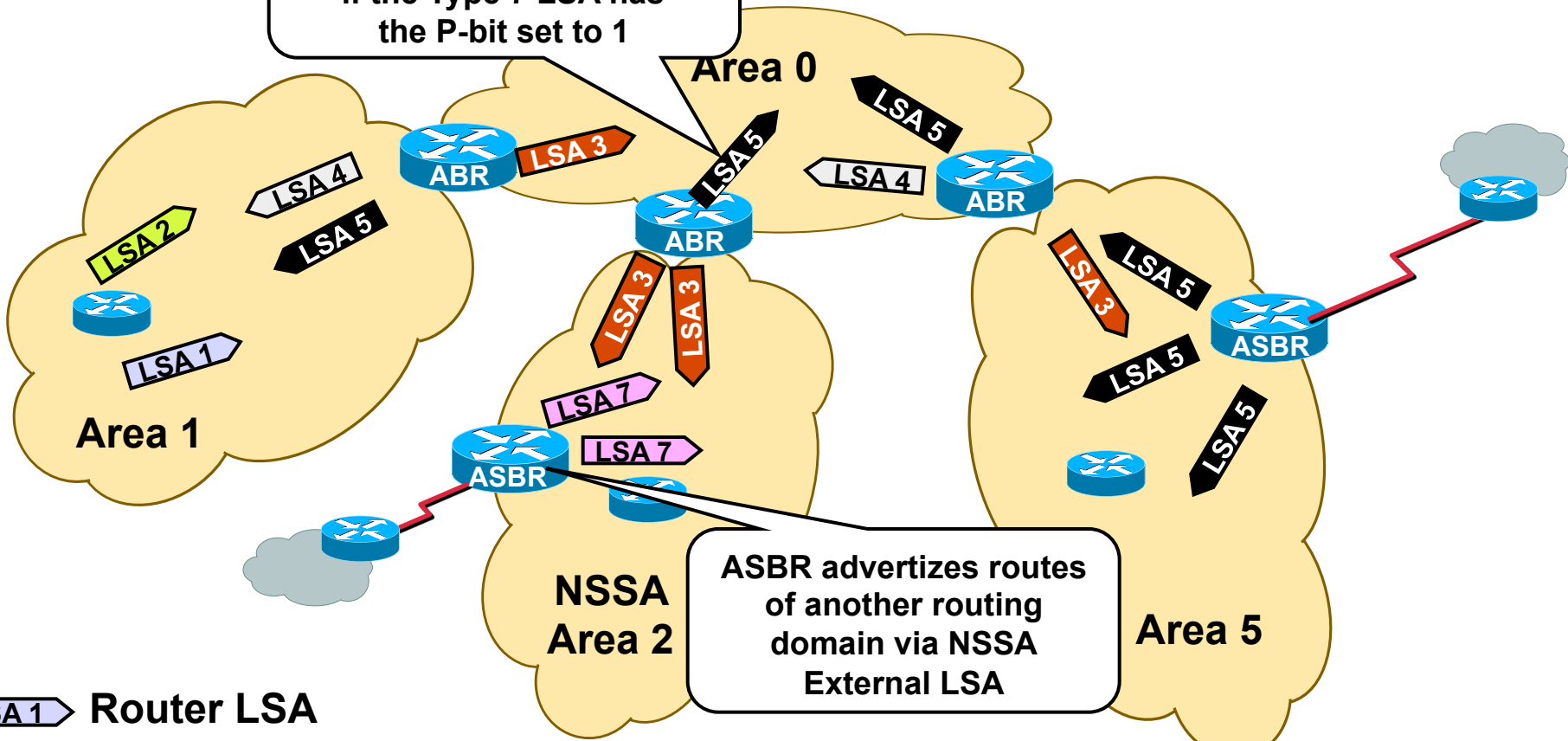
Cisco Specific



Not So Stubby Area (NSSA)

FYI

ABR will **translate** the Type 7 LSA into a Type 5 LSA only if the Type 7 LSA has the P-bit set to 1



- LSA 1 Router LSA
- LSA 2 Network LSA
- LSA 3 Network Summary LSA

- LSA 4 ASBR Summary LSA
- LSA 5 AS External LSA
- LSA 7 NSSA External LSA

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
 - Introduction
 - The Dijkstra Algorithm
 - Communication Procedures
 - LSA Broadcast Handling
 - Split Area
 - Broadcast Networks
 - Area Principles
 - Stub Areas
 - Route Summarization
 - Virtual Link
 - Summary
 - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

Summary LSA and Route Summarization

- **Summary LSA is generated by Area Border Router to inform**
 - Routers inside its area about costs of networks from outside (message direction: Backbone Area -> Area)
--> import of net-IDs
 - Routers outside its area about costs of its internal networks (message direction: Area -> Backbone Area)
--> export of net-IDs
- **Additionally Summary Link LSA can be used for Route Summarization**
 - Several net-IDs can be summarized to a single net-ID using an appropriate subnet-mask

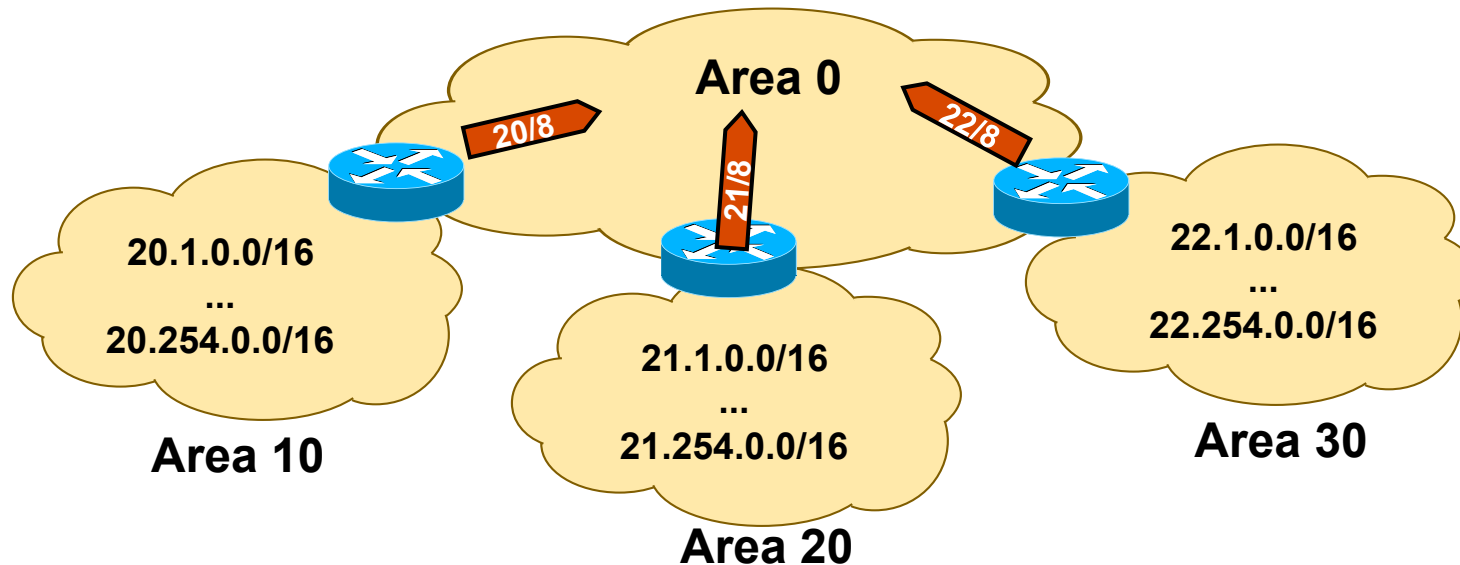
- **Route Summarization can be configured manually for Area Border Routers**
 - To minimize number of routing table entries
 - To provide decoupling of OSPF areas
- **Basically, an OSPF domain allows combining any IP-address with any arbitrary subnet masks**
 - Classless Routing
- **No automatic Route Summarization at the IP address class boundary (A,B or C) like RIPv1**
 - Note: RIPv1 implements Classful Routing

- **Summarization can occur at any place of the IP-address**
- **For instance, many class C addresses can be summarized to one single address (with a prefix)**
 - E.g. class C addresses 201.1.0.0 to 201.1.255.0 (subnet-mask 255.255.255.0) can be summarized by a single entry 201.1.0.0 with subnet-mask 255.255.0.0
 - Note1: when summarizing several networks, only the lowest costs of all these networks are reported (RFC 1583)
 - Note2: when summarizing several networks, only the highest costs of all these networks are reported (RFC 2328)

- **OSPF Route Summarization demands**
 - A clever assignment of IP-addresses and areas to enable Route Summarization
- **Hence OSPF not only forces a star shaped area topology but also demands for a sound IP-address design**
- **Note:**
 - It is still possible to use arbitrary subnet masks and arbitrary addresses anywhere in the network because of classless routing
 - In conflict cases "Longest Match Routing Rule" is applied
 - But this means a bad network design

Example Summarization

- Efficient OSPF address design requires hierarchical addressing
- Address plan should support summarization at ABRs

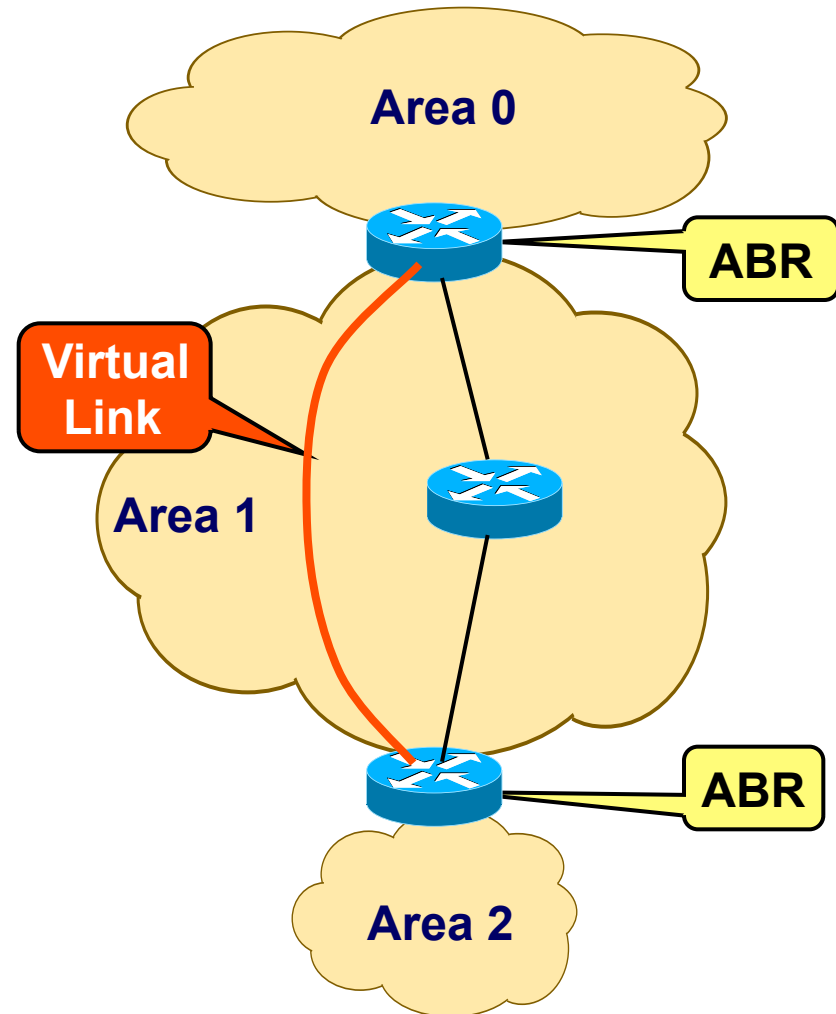


Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
 - Introduction
 - The Dijkstra Algorithm
 - Communication Procedures
 - LSA Broadcast Handling
 - Split Area
 - Broadcast Networks
 - Area Principles
 - Stub Areas
 - Route Summarization
 - Virtual Link **FYI**
 - Summary
 - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

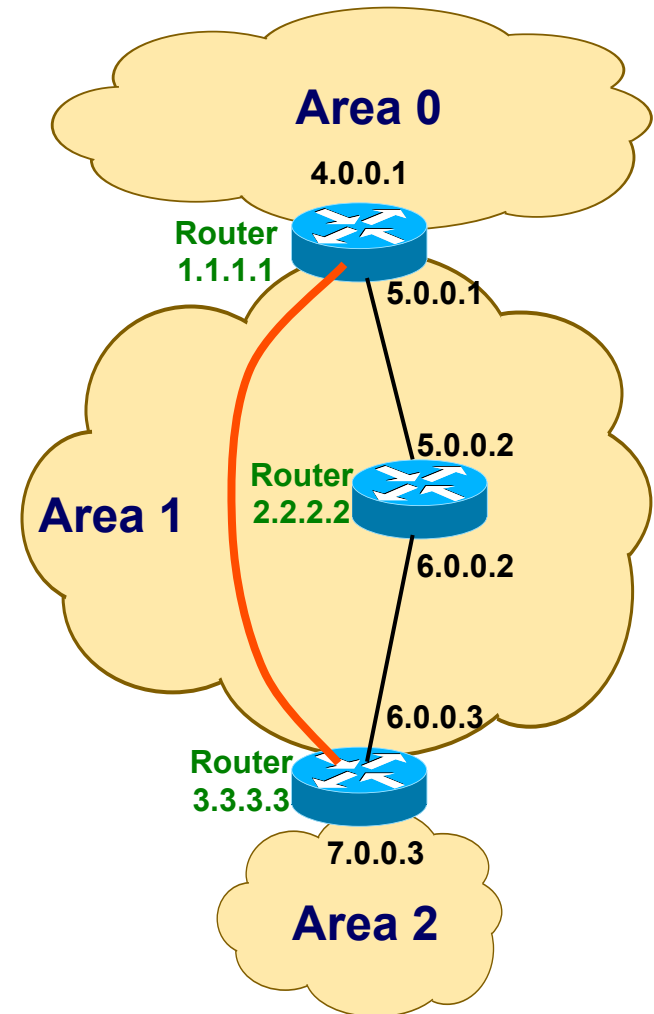
Virtual Links

- Another way to connect to area 0 using a point-to-point tunnel
- Transit area must have full routing information
 - Must *not* be stub area
- **Bad Design!**



Virtual Link Example

- **Now router 3.3.3.3 has an interface in area 0**
- **Thus router 3.3.3.3 becomes an ABR**
 - Generates summary LSA for network 7.0.0.0/8 into area 1 and area 0
 - Also summary LSAs in area 2 for all the information it learned from areas 0 and 1



Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
 - Introduction
 - The Dijkstra Algorithm
 - Communication Procedures
 - LSA Broadcast Handling
 - Split Area
 - Broadcast Networks
 - Area Principles
 - Stub Areas
 - Route Summarization
 - Virtual Link
 - Summary
 - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

Distance-Vector **versus** Link-State

- Distance-Vector:
 - Every router notifies directly connected routers about all reachable routes
 - Using broadcast messages
 - Maintains its routing table according to information from neighbor routers
- Link-State:
 - Every router notifies all routers about the state of his directly connected links
 - Using flooding mechanism (LSA)
 - Calculates optimal paths whenever a new LSA is received

OSPF Benefits 1

- **Network load is significantly smaller than that of distance vector protocols**
 - Short hello messages between adjacent routers versus periodical emission of the whole routing table
- **Even update messages after topology modifications are smaller than the routing table of distance vector protocols**
 - LSAs only describe the local links for which a router is responsible -> incremental updates !!!
- **Massive network load**
 - Occurs only on combining large splitted network parts of an OSPF domain (many database synchronizations)

OSPF Benefits 2

- **SPF-techniques take advantages from several features:**
 - Every router maintains a complete topology-map of the entire network and calculates independently its desired paths (actually based on the original LSA message)
 - This local ability for route calculation grants a fast convergence
 - LSA is not modified by intermediate routers across the network
 - The size of LSAs depends on the number of direct links of a router to other routers and not on the number of subnets!

OSPF Benefits 3

- **During router configuration, every physical port is assigned a cost value**
 - Per ToS (Type of Service)
 - Each ToS can be assigned a separate topology map (8 possible combinations)
 - IP's ToS field may be examined for packet forwarding
 - Note: OSPF ToS support disappeared in RFC 2328
- **Determination of the best path for a specific ToS is based on the summary costs along the paths**
 - RIP uses hop count only
- **Equal costs automatically enables load balancing between these paths**

OSPF Benefits 4

- **Subnet masks of variable length can be attached to routes (in contrast to RIPv1)**
- **External routes are marked (tagged) explicitly to be differentiated from internal routes**
- **OSPF messages can be authenticated to grant secure update information**
- **OSPF routing messages use IP-multicast addresses: lower processing effort**
- **Point-to-point connections do not need own IP-address**
 - In theory more economic use of address space is possible
 - But for practical reasons regarding network management also on point-to-point connections usage of IP addresses are recommended

OSPF in Large Networks

- **OSPF area concept can be used**
 - A two level hierarchy is used to decrease
 - CPU time for SPF calculations
 - Memory requirement for storing topology database
 - One backbone area
 - Several non-backbone areas
 - Non-backbone area can be connected by area border router to backbone area only
 - Summarization possible at area border routers
 - Route aggregation to reduce size of routing tables
 - Summarization means that some net-IDs can be summarized as one net-ID only

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
 - Introduction
 - The Dijkstra Algorithm
 - Communication Procedures
 - LSA Broadcast Handling
 - Split Area
 - Broadcast Networks
 - Area Principles
 - Stub Areas
 - Route Summarization
 - Virtual Link
 - Summary
 - OSPF Header Details **FYI**
- **Introduction to Internet Routing (BGP, CIDR)**

Router LSA – Type 1

- **Router ID (Highest IP address)**
- **Number of Links**
- **Link Descriptions**
 - Link type (P2P, Stub, ...)
 - Neighboring router ID
 - Router interface address
 - ToS (typically not supported today)
 - Metrics

Network LSA – Type 2

- **DR's IP address**
- **One Subnet mask for this broadcast segment**
- **List of Router-IDs of all routers in the broadcast segment**

Network Summary LSA – Type 3

- Originated by **ABRs** only
- Each LSA Type 3 contains a number of
 - Destination networks + Subnet masks
 - Metric for each destination network
- This is basically a distance-vector routing information (!)

ASBR Summary LSA – Type 4

- Originated by **ABRs**
- Advertise routes to ASBRs
- Nearly identical to Type 3
 - Except destination is ASBR not a network
- Each LSA Type 4 contains
 - Router IDs of ASBRs
 - Mask 0.0.0.0 (host route)
 - Metric

AS External LSA – Type 5

- **Originated by ASBRs**
 - External type 1
 - External type 2 (default)
- **Advertises**
 - External routes
 - Default route
- **Contains**
 - External Net-ID + Mask
 - Metric
 - Next hop (external, not ASBR)

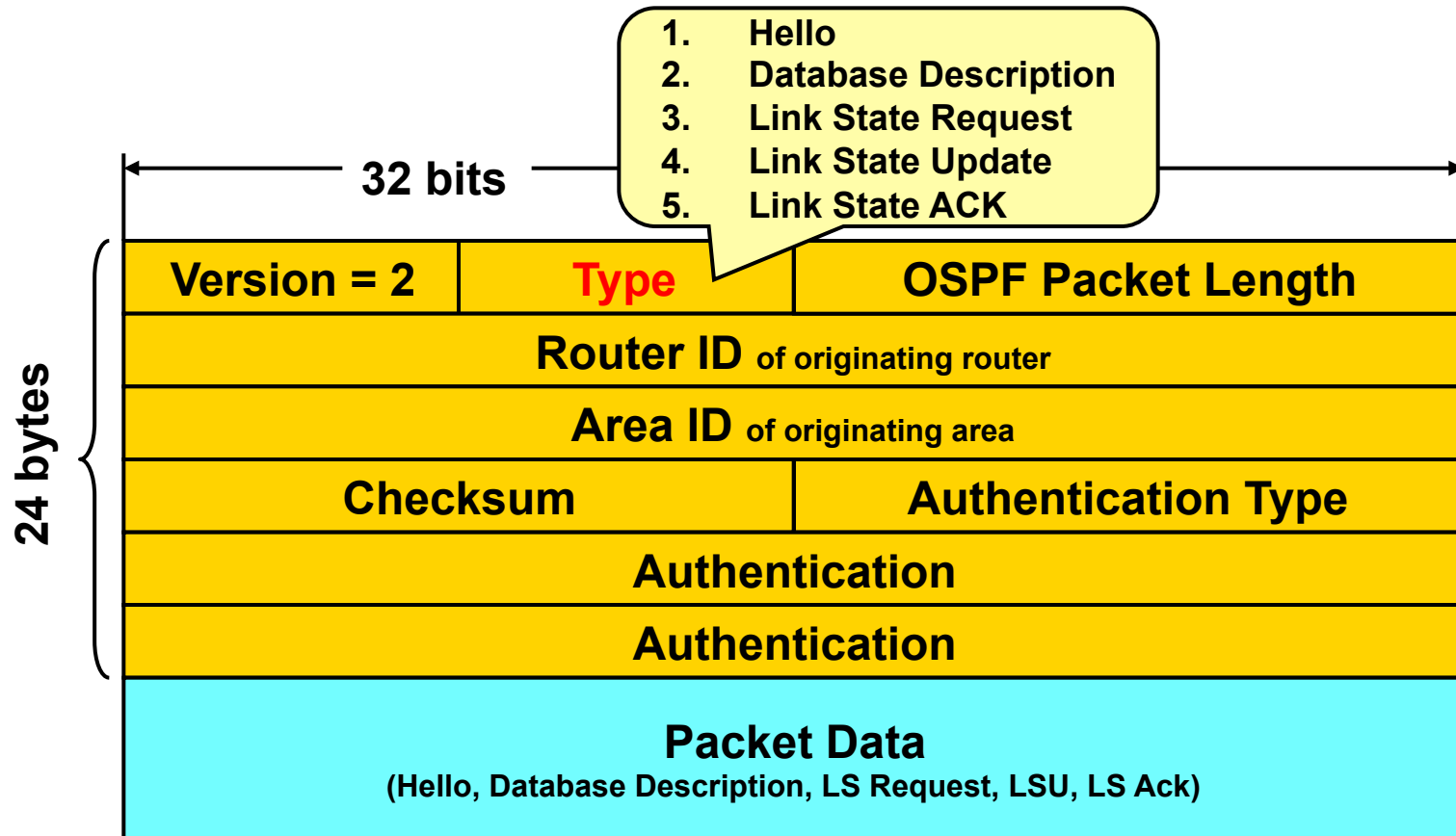
NSSA External LSA – Type 7

- **Originated by ASBRs within NSSAs**
- **Almost identical to Type 5**
 - But only flooded within NSSA
- **RFC 1587**

Other LSAs

- **Group Membership LSA (6)**
 - For MOSPF
- **External Attribute LSA (8)**
 - Alternative to IBGP
 - Should transport BGP information within an OSPF domain
 - Not yet implemented, no RFC yet (?)
- **Opaque LSA (9)**
 - Application specific information
 - Link local scope
- **Opaque LSA (10)**
 - Application specific information
 - Area-local scope
- **Opaque LSA (11)**
 - Application specific information
 - AS scope

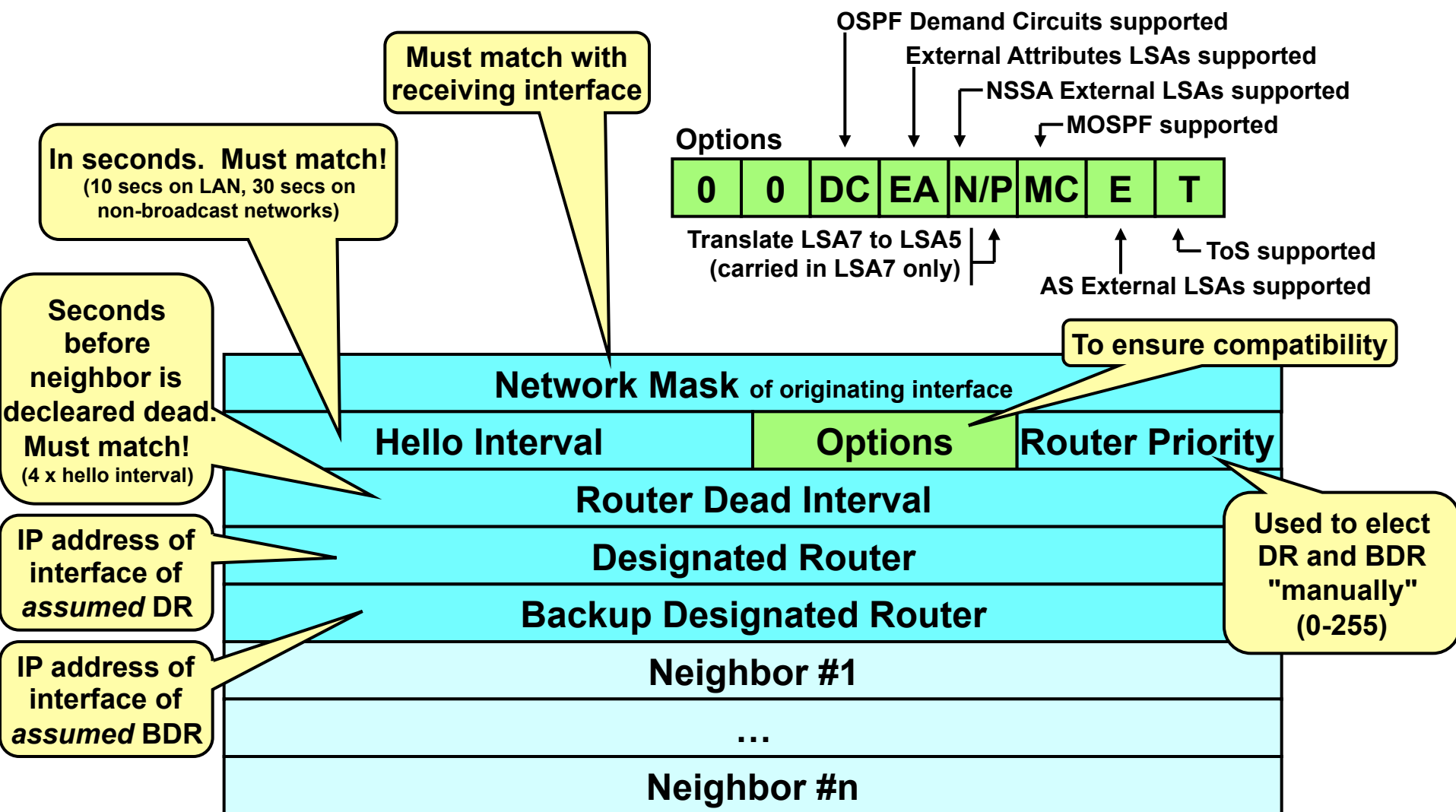
General OSPF Packet Structure



- Carried directly in IP (protocol number 89)
- **All OSPF packets begin with a 24-byte OSPF packet header**

Hello Packet

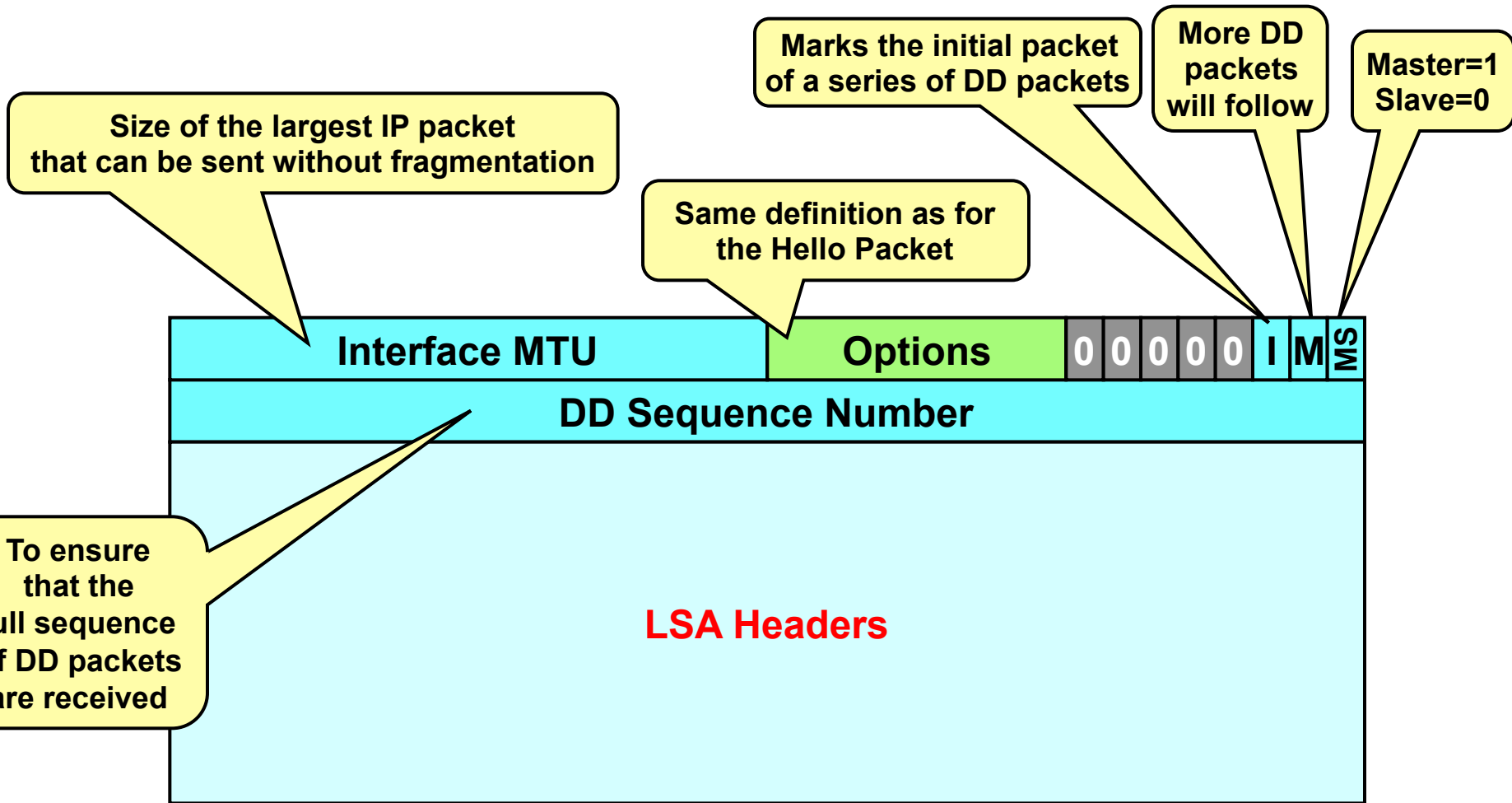
Type 1



Database Description Packet

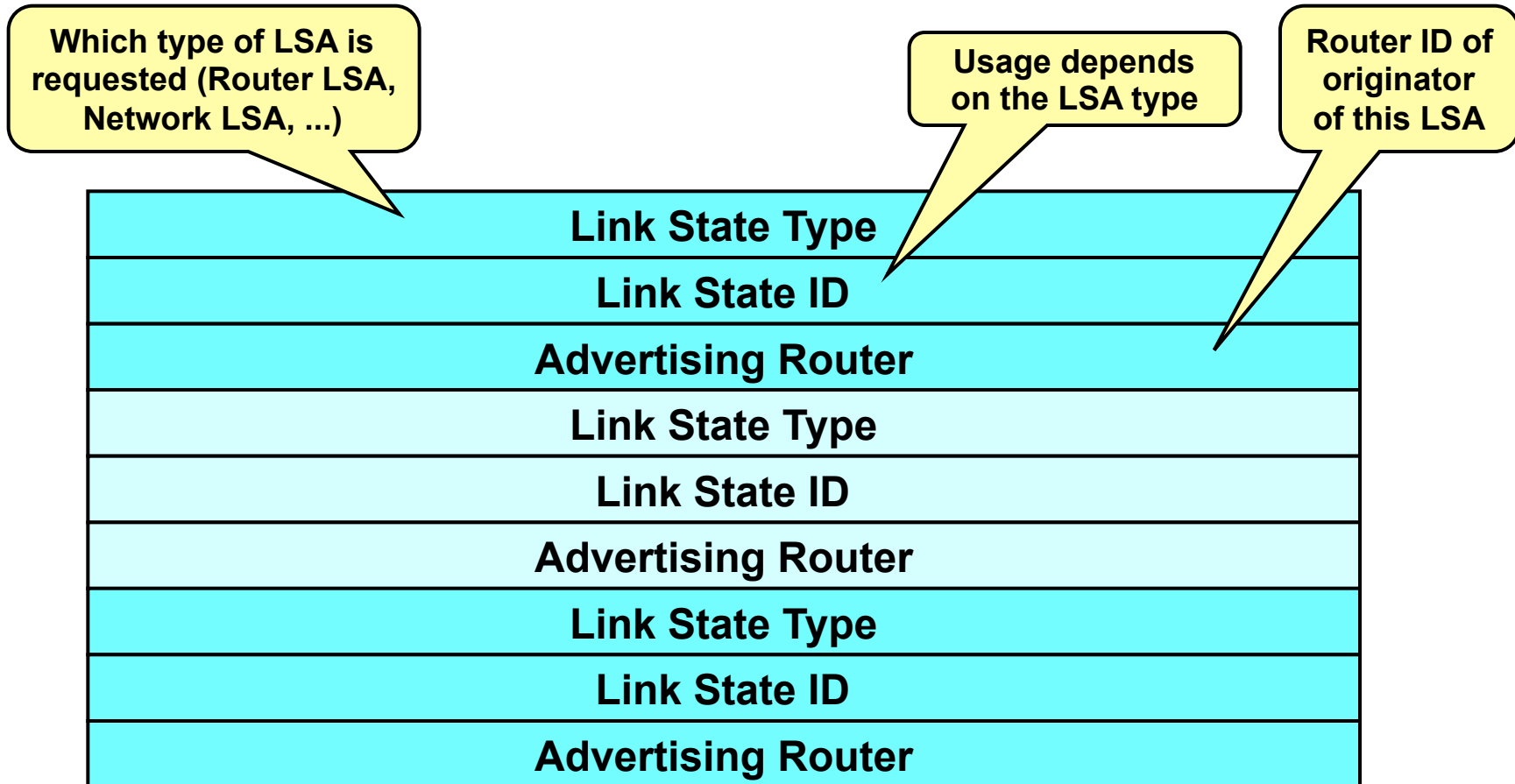
Type 2

Also called "DDP"



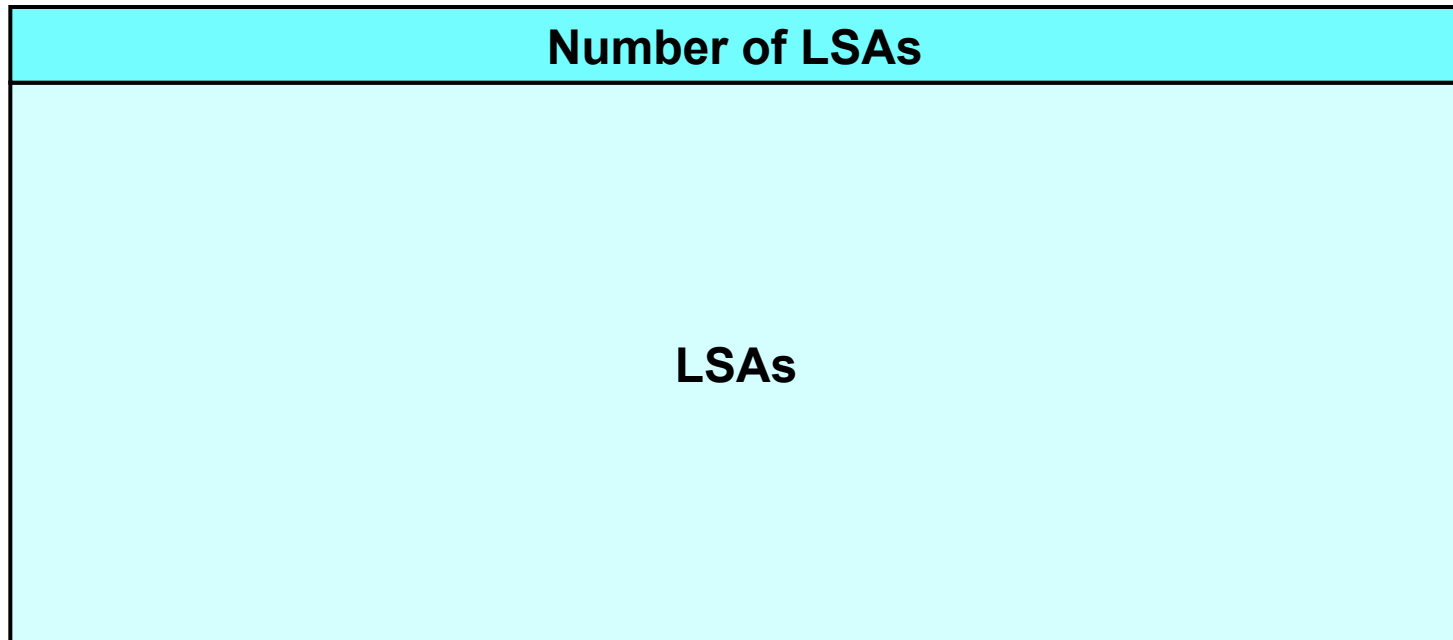
Link State Request Packet

Type 3



Link State Update Packet

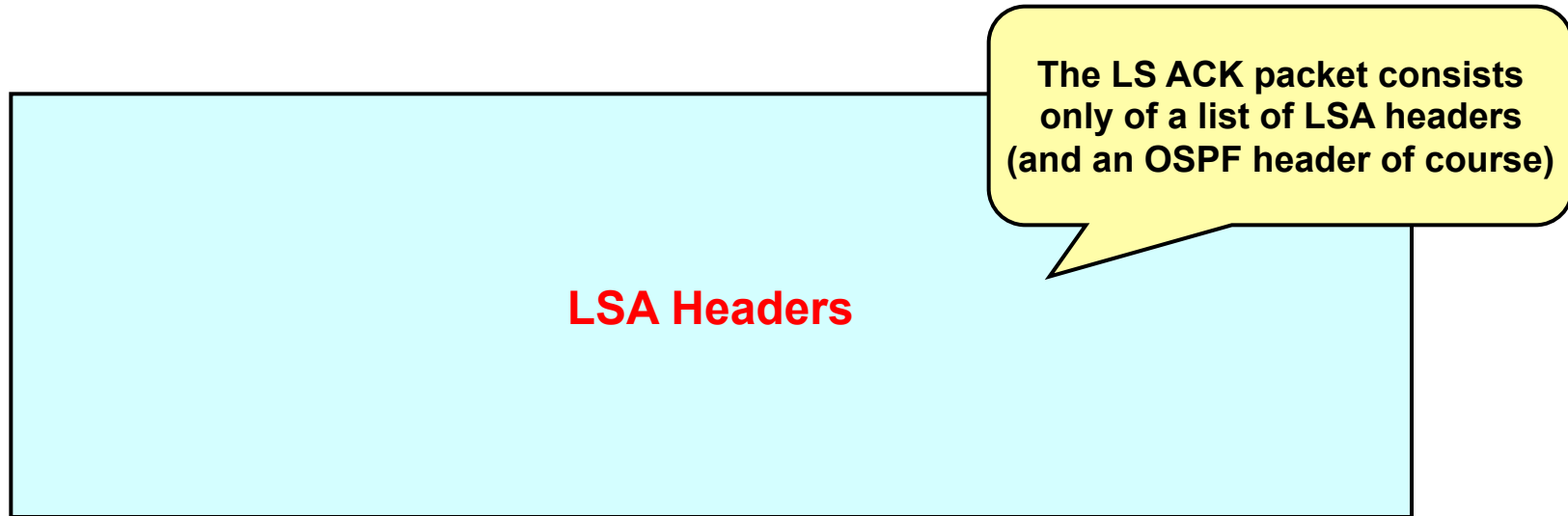
Type 4



- **LSUs contain one or more LSAs (limited by MTU)**
- **Used for flooding and response to LS requests**
- **LSUs are carried hop-by-hop**

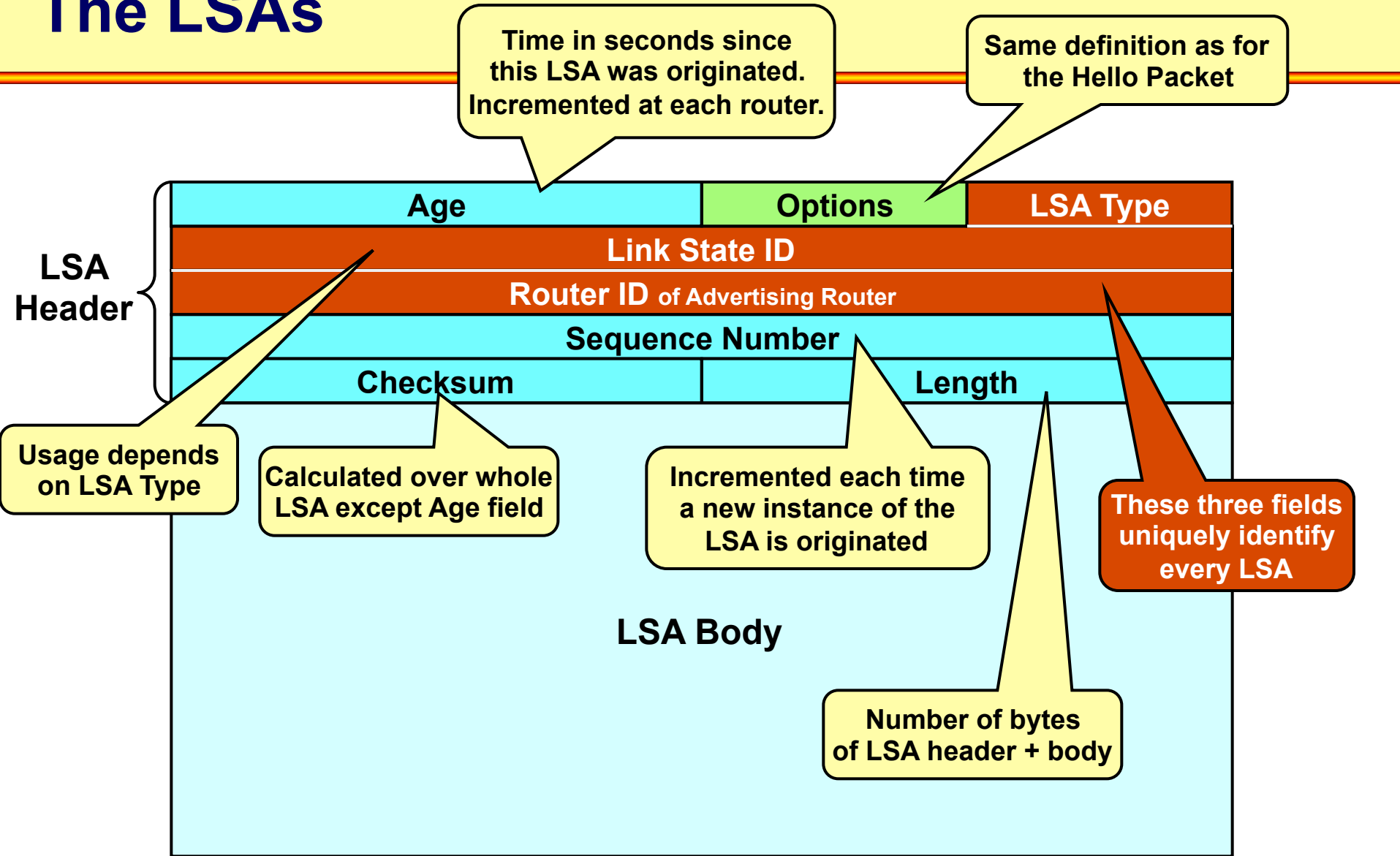
Link State ACK Packet

Type 5

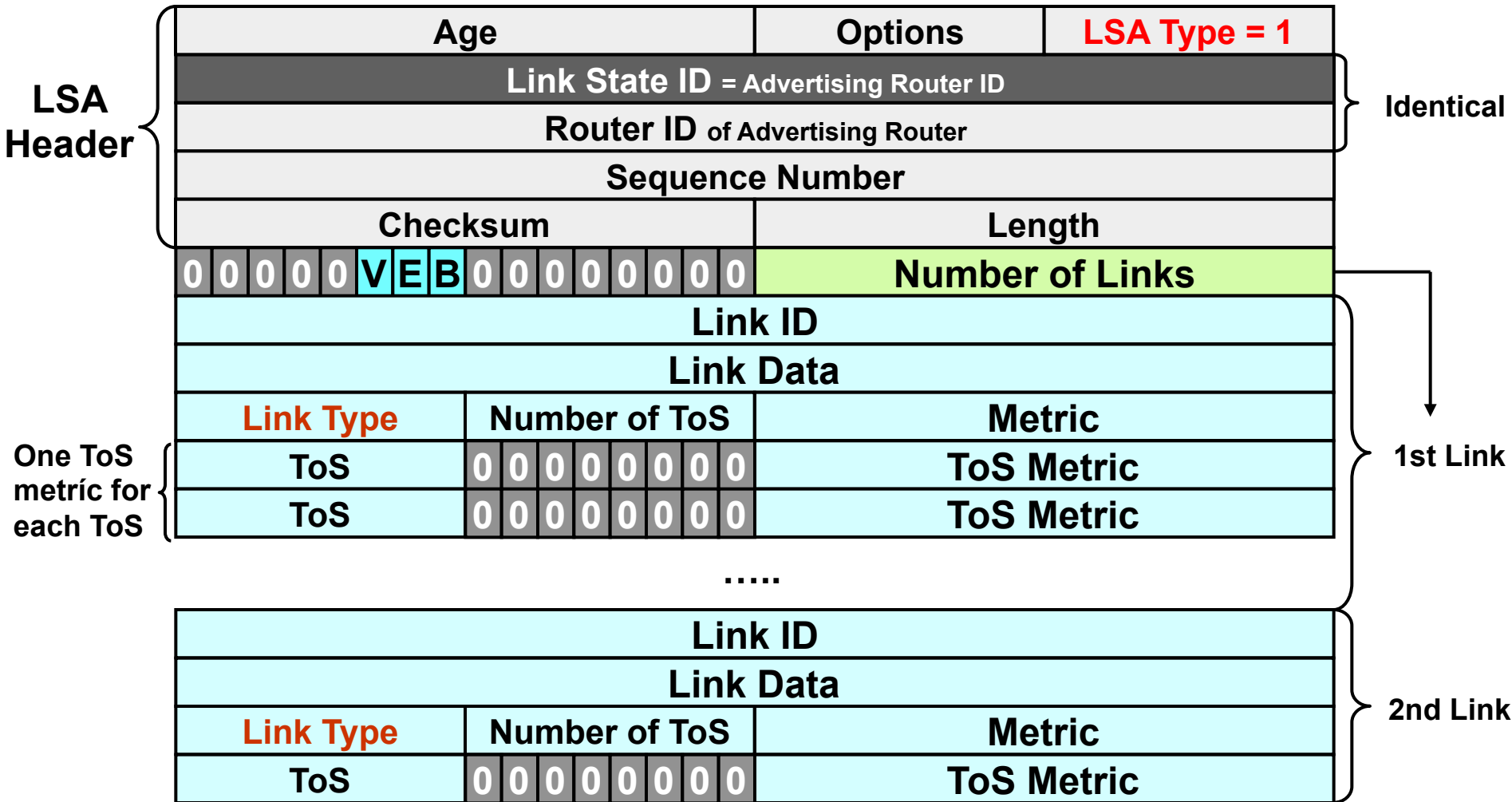


- Each LSA received must be **explicitly** acknowledged → reliable flooding!
- Acknowledged LSA is identified by **LSA header**
- Single Link State ACK packet can acknowledge multiple LSAs

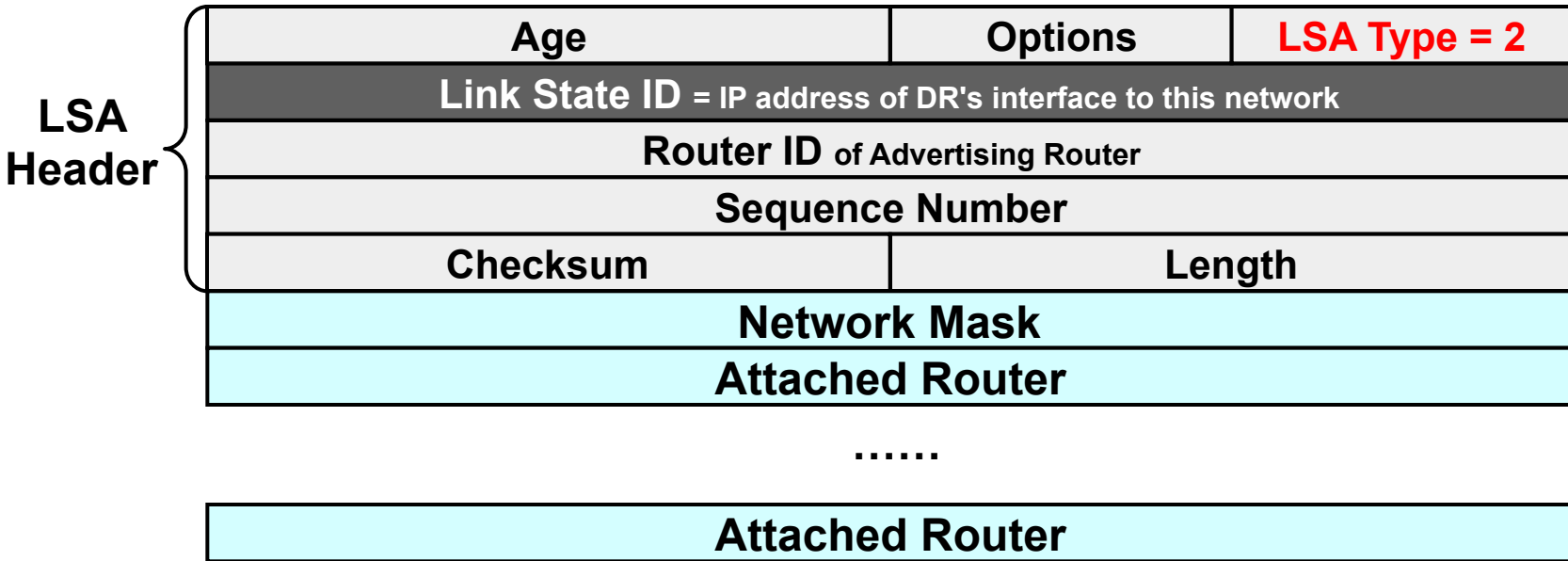
The LSAs



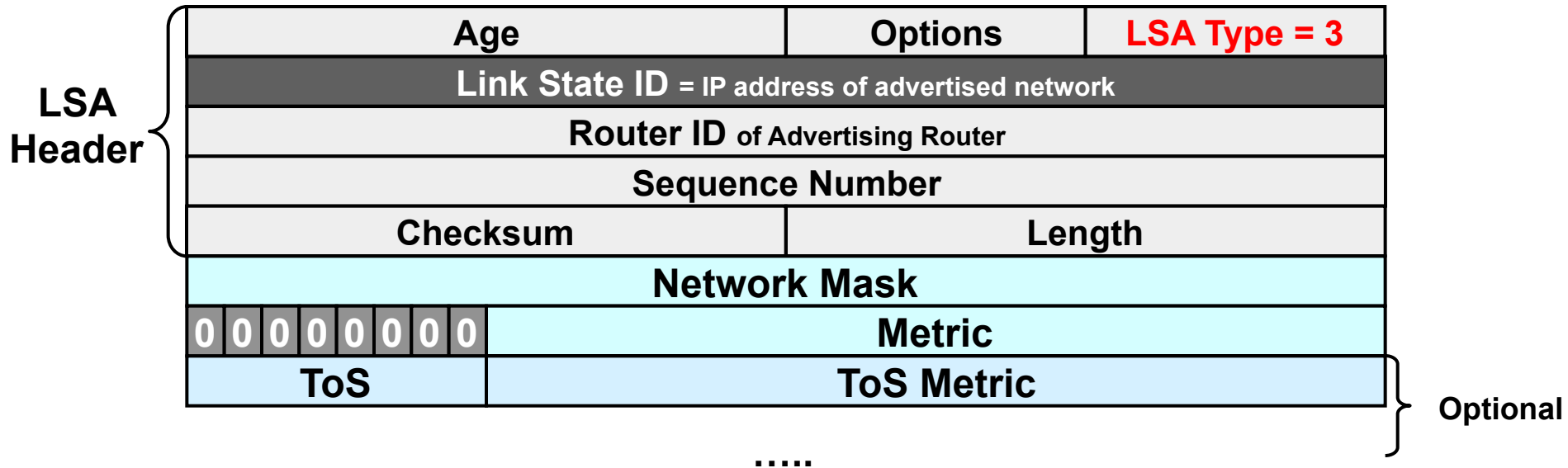
Router LSA



Network LSA

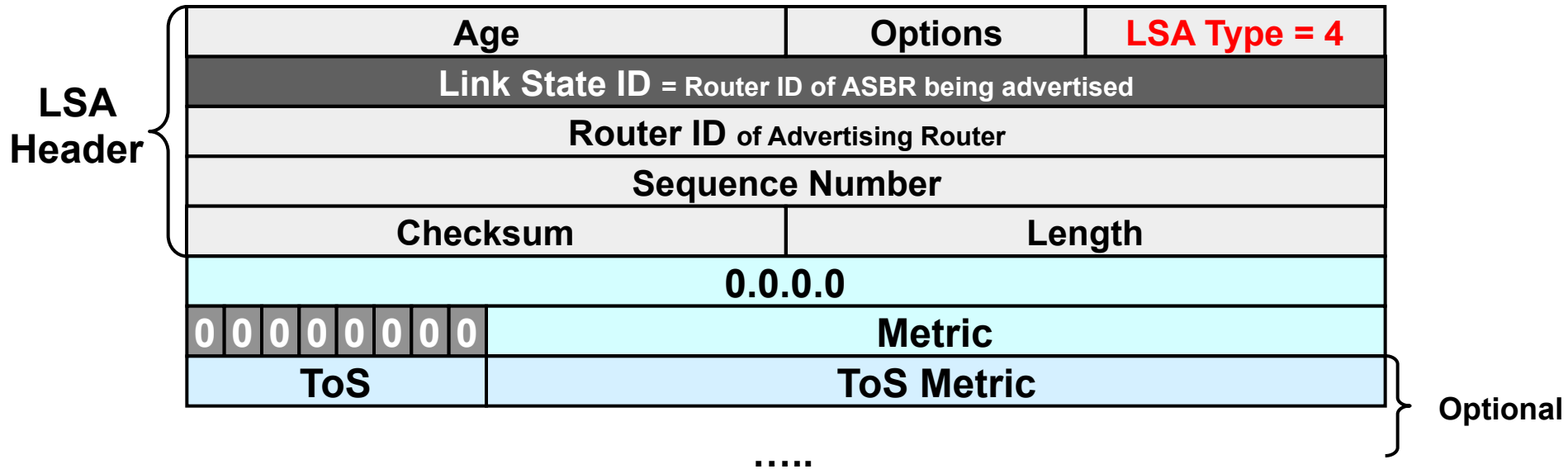


Network Summary LSA



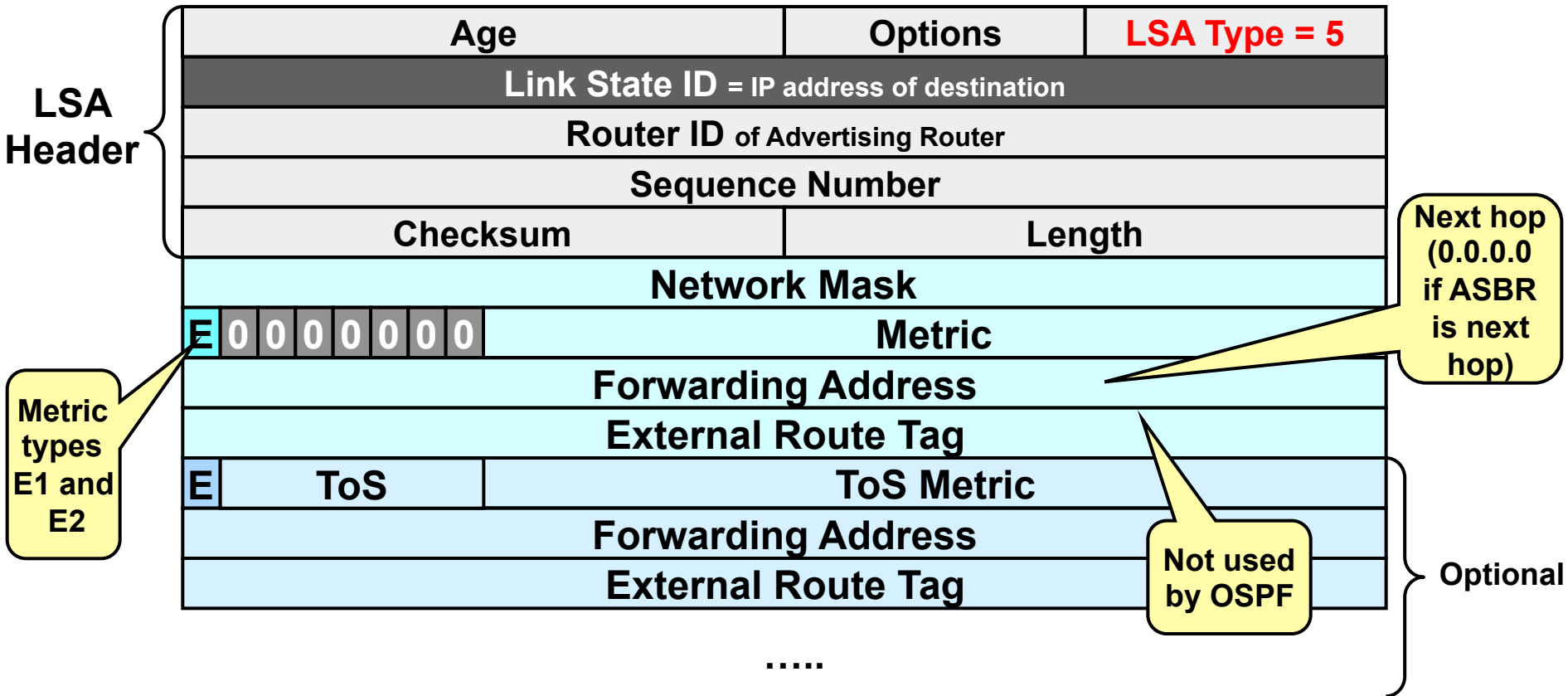
- If a **default route** is advertised, both the Link State ID and the Network Mask fields will be 0.0.0.0
- Also used for route summarization
- Note: Cisco only supports ToS=0

ASBR Summary LSA



- Note: Cisco only supports ToS=0

Autonomous System External LSA



- When describing a default route, both the Link State ID and the Network Mask are set to 0.0.0.0.

NSSA External LSA

- **Same structure as AS External LSA**
- **Forwarding address is**
 - Next hop address for the network between NSSA and adjacent AS, if this network is advertised as internal route
 - Router ID of NSSA-ASBR otherwise

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**
 - Introduction
 - BGP Basics
 - BGP Attributes
 - BGP Special Topics
 - CIDR

Routing in Small Networks

- **In small networks**

- Distance vector or link state protocols like RIP or OSPF can be used for dynamic routing
- It is possible that every router of the network knows about all destinations
 - All destination networks will appear in the routing tables
- Routing decisions are based on technical parameters
 - E.g. hop count, link bandwidth, link delay, interface costs
- It is sufficient that routing relies only on technical parameters
 - Small networks will be administered by a single authority
 - Non-technical parameter like traffic contracts have no importance

Routing in Large Networks

- **With increasing network size limitations of these protocols can be recognized**
 - Some limitations for example
 - Maximum hop count (RIP)
 - Time to transmit routing tables (RIP) on low speed links
 - CPU time for SPF calculation (OSPF)
 - Memory used for storing routing table (RIP, OSPF)
 - Memory used for storing topology database (OSPF)
 - Two level hierarchy centered around a core network (OSPF)
 - Route fluctuation caused by link instabilities (OSPF)
 - Routing based on non-technical criteria like financial contracts or legal rules is not possible

Routing in the Internet

- **Limitations prevent using routing protocols like RIP or OSPF for routing in the Internet**
 - Note: routing tables of Internet-core routers have about 415.000 net-ID entries (May 2012)
- **Routing in the Internet**
 - Is based on non-technical criteria like financial contracts or legal rules
 - Policy routing
 - Acceptable Use Policy (AUP) in parts of the Internet
 - Contracts between Internet Service Providers (ISP)

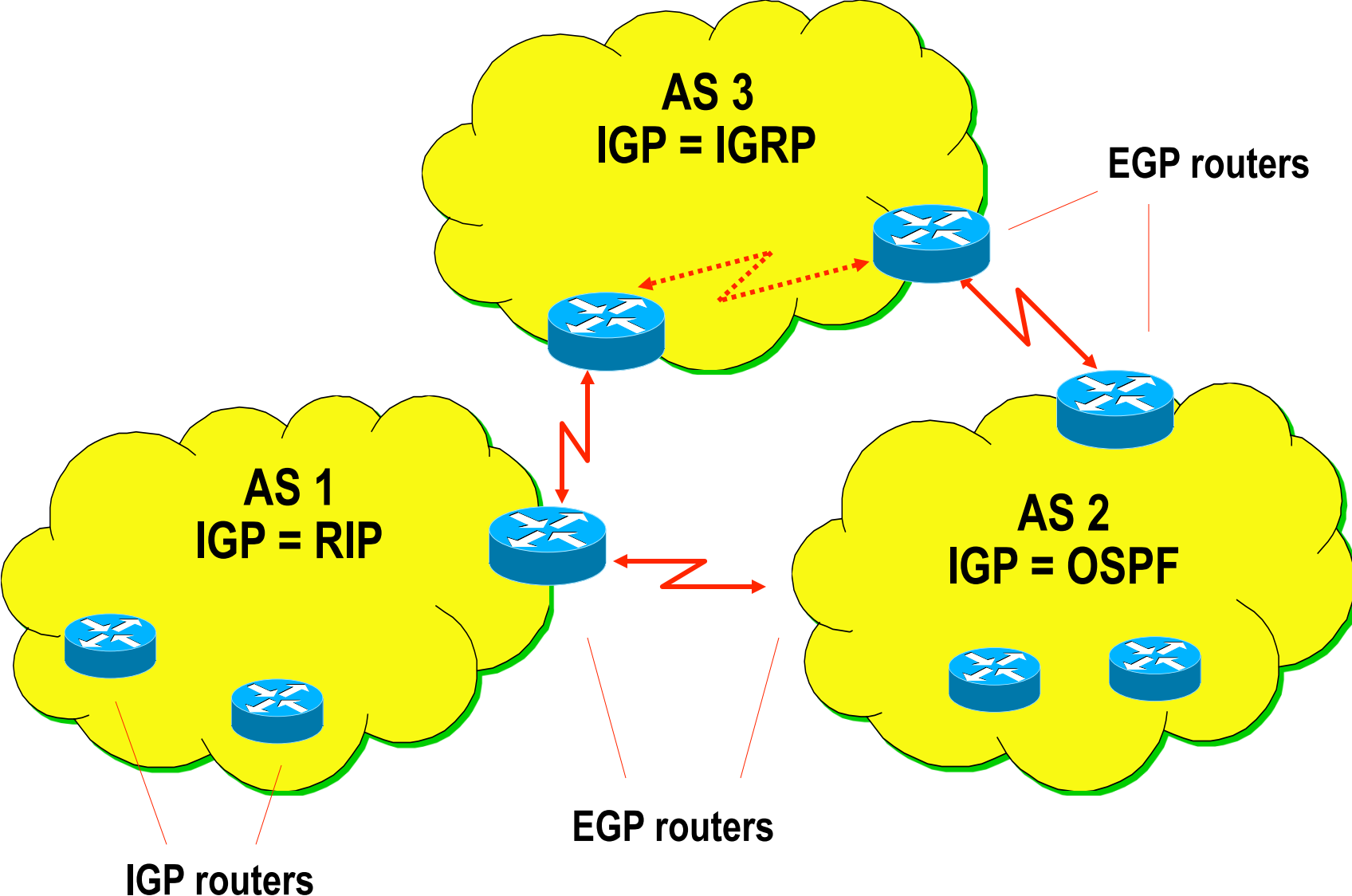
Routing Hierarchy, Autonomous Systems

- **Routing hierarchy is necessary for large networks**
 - To control expansion of routing tables
 - To provide a more structured view of the Internet
- **Routing hierarchy used in the Internet**
 - Based on concept of autonomous system (AS)
- **AS concept allows**
 - Segregation of routing domains into separate administrations
 - Note: routing domain is a set of networks and routers having a single routing policy running under a single administration

IGP, EGP

- **Within an AS one or more IGP protocols provide interior routing**
 - IGP - Interior Gateway Protocol
 - IGP examples
 - RIP, RIPv2, OSPF, IGRP, eIGRP, Integrated IS-IS
 - IGP router responsible for routing to internal destinations
- **Routing information between ASs is exchanged via EGP protocols**
 - EGP - Exterior Gateway Protocols
 - EGP router knows how to reach destination networks of other ASs
 - EGP examples
 - EGP-2, BGP-3, BGP-4

AS, IGP, EGP



AS Numbers

- **Hierarchy based on ASs allows forming of a large network**
 - By dividing it into smaller and more manageable units
 - Every unit may have its own set of rules and policies

- **AS are identified by a unique number**
 - Can be obtained like IP address from an Internet Registry
 - e.g. RIPE NCC (Reséaux IP Européens Network Coordination Center)

BGP-4 (1)

- **Border Gateway Protocol (BGP)**
 - Is the Exterior Gateway Protocol used in the Internet nowadays
 - Was developed to overcome limitations of EGP-2
 - RFC 1267 (BGP-3) older version
 - classful routing only
 - RFC 1771 (BGP-4) current version, DS
 - classless routing
 - Is based on relationship between neighboring BGP-routers
 - Peer to peer
 - Called BGP session or BGP connection

BGP-4 (2)

- **Border Gateway Protocol (cont.)**

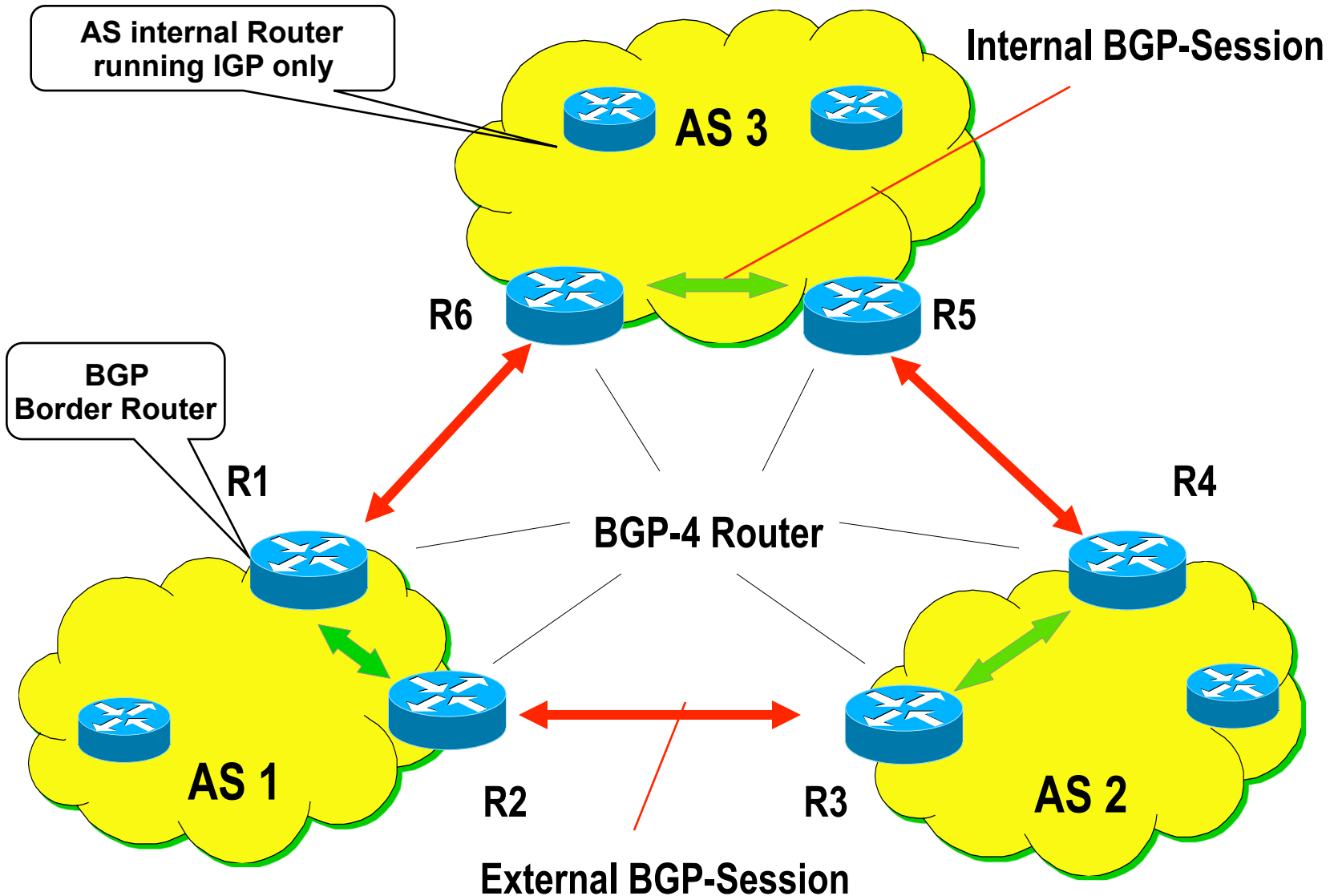
- Primary function

- Exchange of network reachability information with other autonomous systems via external BGP sessions
- But also within an autonomous system between BGP border routers via internal BGP sessions

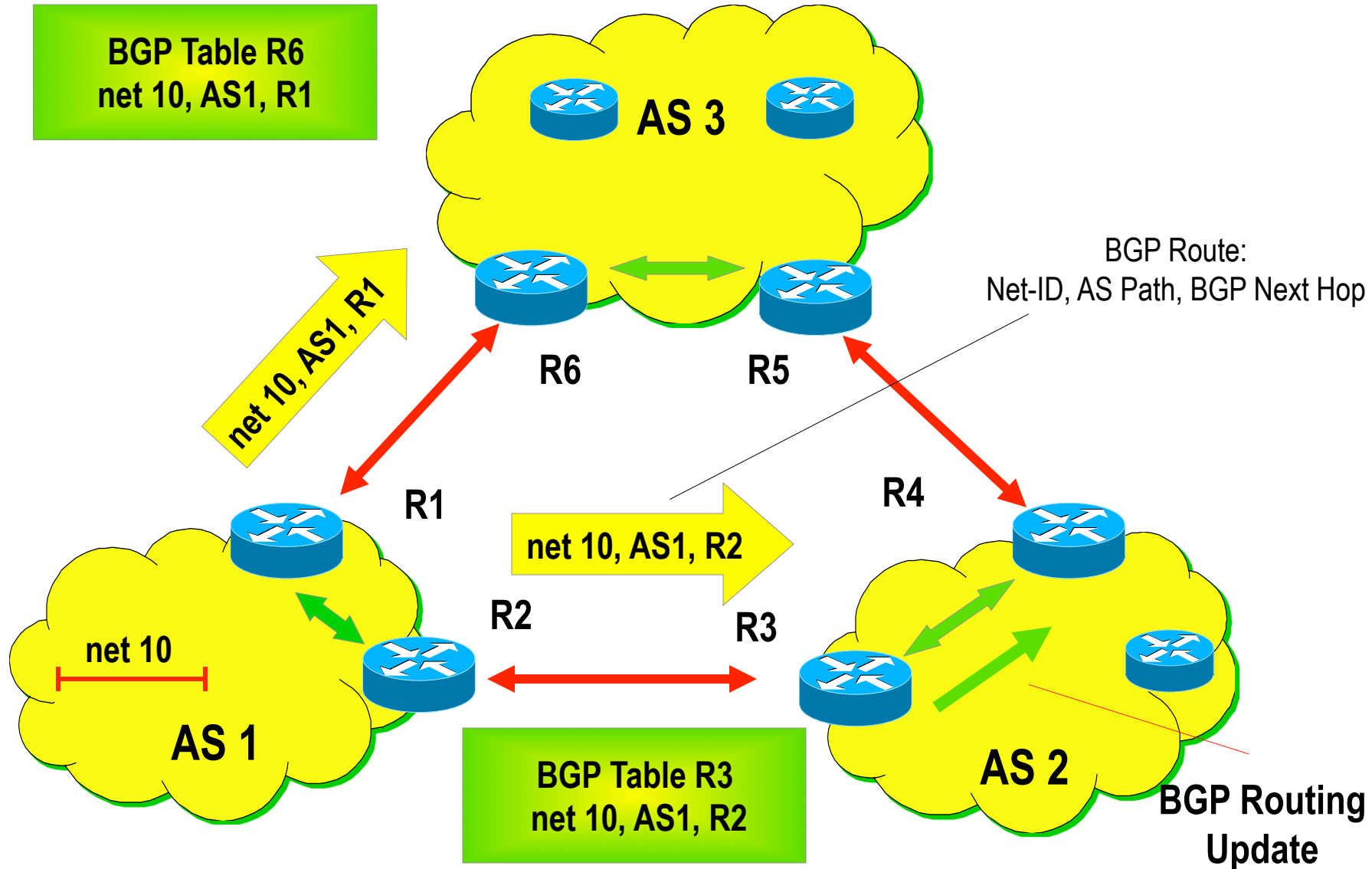
- BGP session runs on top of TCP

- Reliable transport connection
- Well known port 179
- TCP takes care of fragmentation, sequencing, acknowledgement and retransmission
- Hence these procedures need not be done by the BGP protocol itself

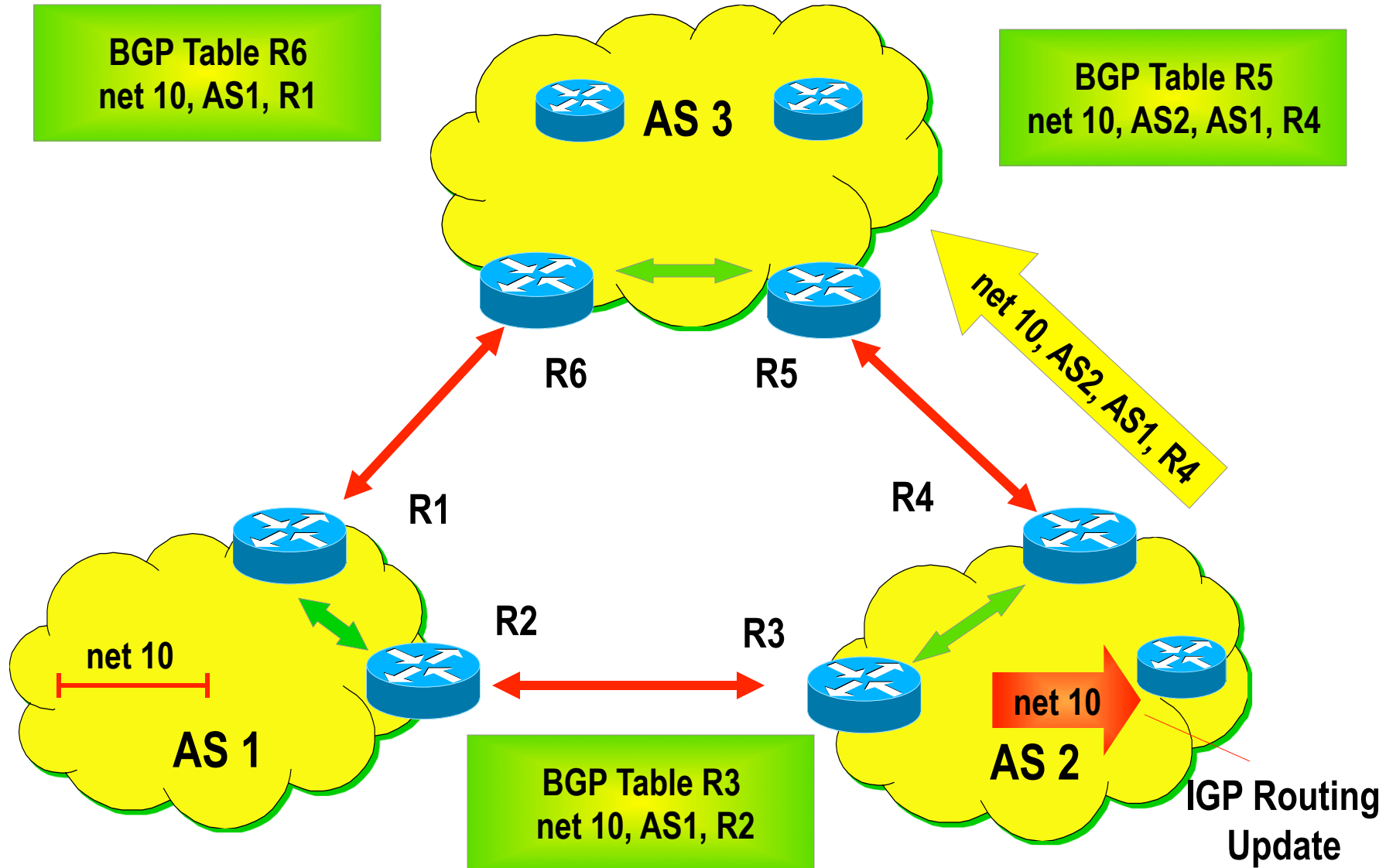
Basic Example (1)



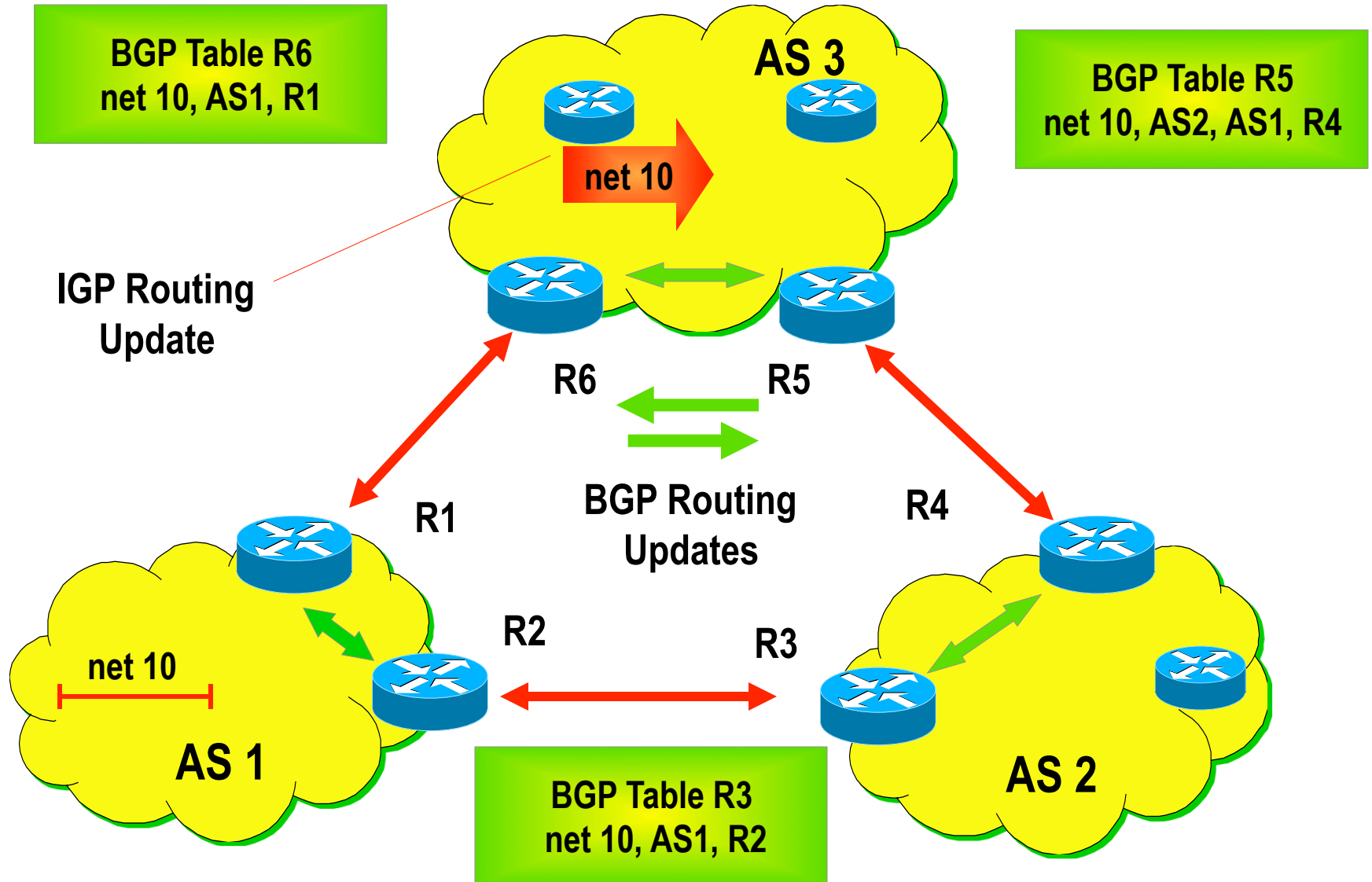
Basic Example (2)



Basic Example (3)



Basic Example (4)



BGP-4 Concepts (1)

- Reachability information exchanged between BGP routers carries a sequence of AS numbers
 - Indicates the path of ASs a route has traversed
- Path vector protocol
- This allows BGP to construct a graph of autonomous systems
 - Loop prevention
 - No restriction on the underlying topology
- The best path
 - Minimum number of AS hops
 - Note: criteria if no other BGP policies are applied
- Incremental update
 - After first full exchange of reachability information between BGP routers only changes are reported

BGP-4 Concepts (2)

- Description of reachability information by BGP attributes
 - For BGP routing
 - For establishing of routing policy between ASs
- BGP-4 advertises so called BGP routes
 - BGP route is unit of information that pairs a destination with the path attributes to that destination
 - AS Path is one among many other BGP attributes
- IP prefix and mask notation
 - Supports VLSM
 - Supports aggregation (CIDR) and supernetting
- Routes can be filtered using attributes, attributes can be manipulated
 - > Routing policy can be established

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**
 - Introduction
 - BGP Basics
 - BGP Attributes
 - BGP Special Topics
 - CIDR

Border Gateway Protocol (BGP)

- **BGP-3**

- Was classful
- Central AS needed (didn't scale well)
- Not further discussed here!
- RFC 1267

- **BGP-4**

- Classless
- Meshed AS topologies possible
- Used today – discussed in the following sections!!!
- RFC 1771

BGP-4 at a Glance

- **Carried within TCP**
 - Manually configured neighbor-routers
 - Therefore reliable transport (port 179)
- **Neighbor routers establish link-state**
 - Hello protocol (60 sec interval)
- **Incremental Updates upon topology changes**
 - New routes are updated
 - Lost routes are withdrawn
- **Each route is assigned a policy and an AS-Path leading to that network**
 - Using attributes

Path Vector Protocol

- **Metric: Number of AS-Hops**
- **All traversed ASs are carried in the AS-Path attribute**
 - BGP is a "Path Vector protocol"
 - Better than Distance Vector because of inherent topology information
 - No loops or count to infinity possible

BGP Database

- **BGP routers also maintain a BGP Database**
 - Roadmap information through path vectors
 - Attributes
- **Routing Table calculated from BGP Database**
- **CPU/Memory resources needed**

Some Interesting Numbers

- **Today's Internet BGP Backbone Routers are burdened**
 - About 415,000 routes (May 2012)
 - About 10,000 Autonomous Systems
- **Although excessive CIDR, NAT, and Default Routes**
- **Collapse expected**
 - Looking for new solutions

Basic Idea of BGP is Easy !

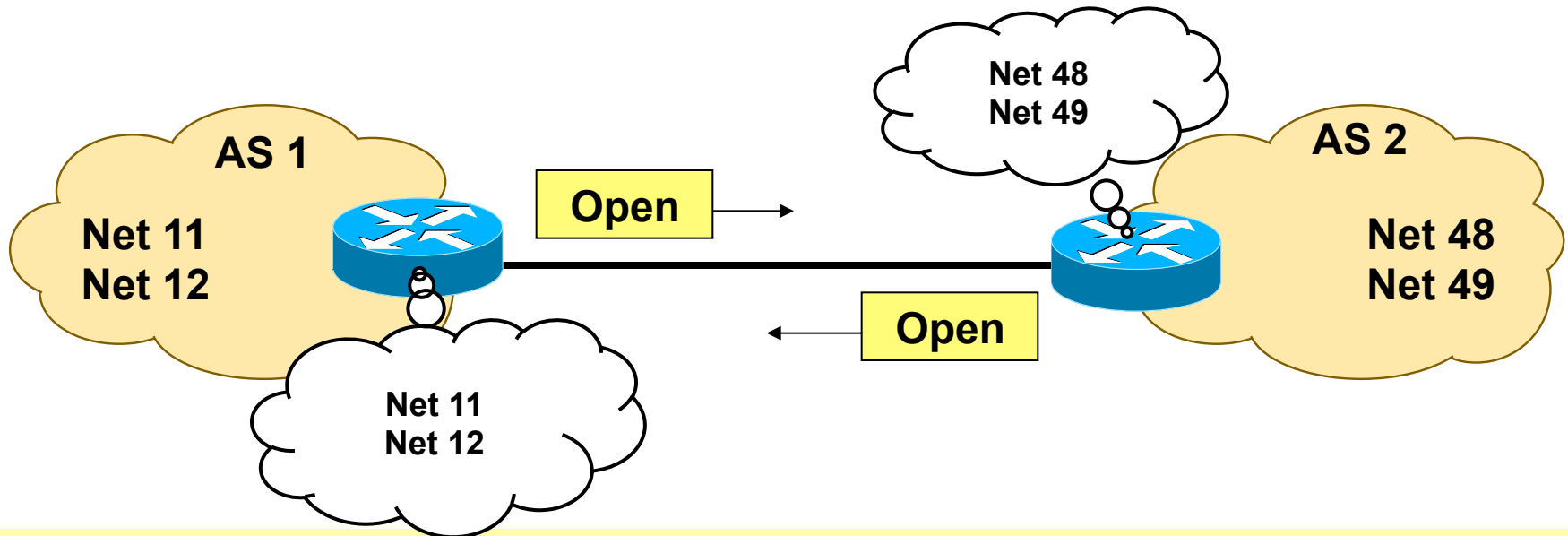
- 1) BGP notifies other Autonomous Systems about reachabilities of networks
- 2) Each single route has attributes associated to it
- 3) Routers can apply policies for each route based on these attributes (e.g. filtering routes)

BGP Limitations

- **Destination based routing**
 - No policies for source address
- **Hop-by-hop routing**
 - Leads to hop-by-hop policies
 - Connectionless nature of IP
 - Mitigated through
 - Community attribute
 - Peer groups

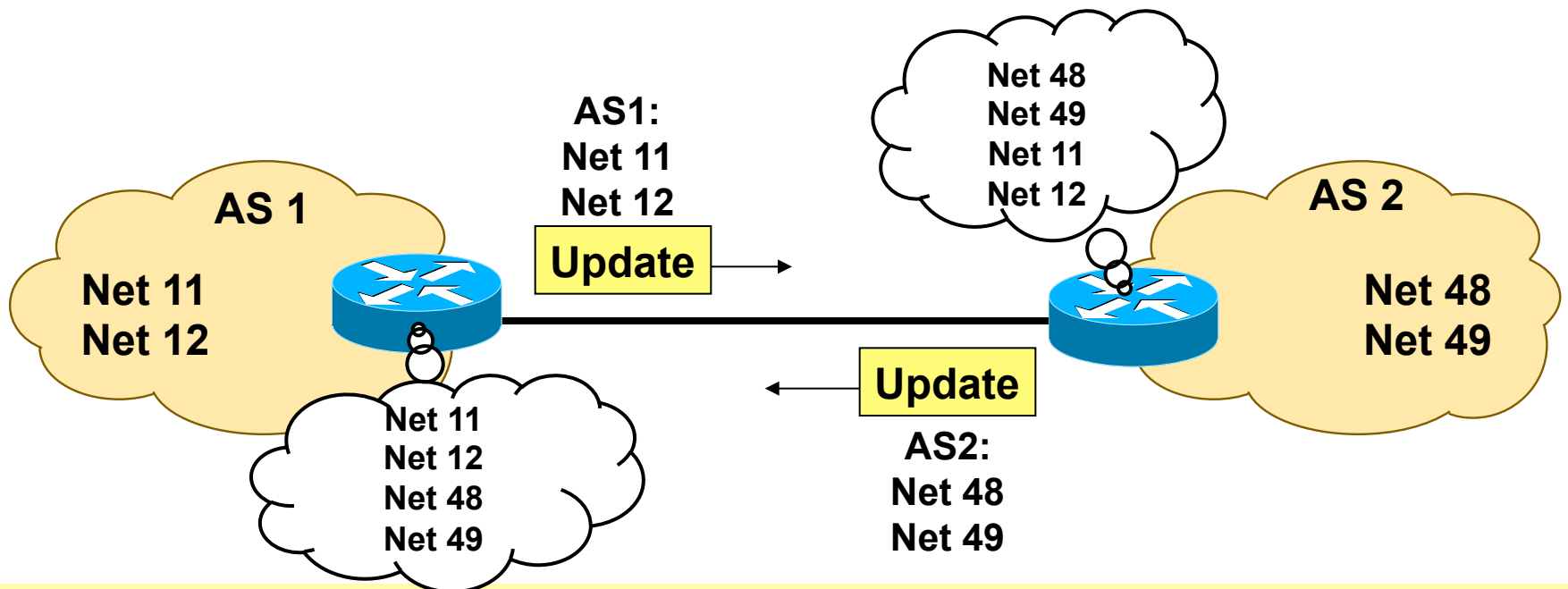
Neighborhood Establishment

- **Open Message**
 - BGP Version (4)
 - AS number
 - BGP Router-ID (IP address)
 - Hold Time
- **Problems are indicated with Notification message**



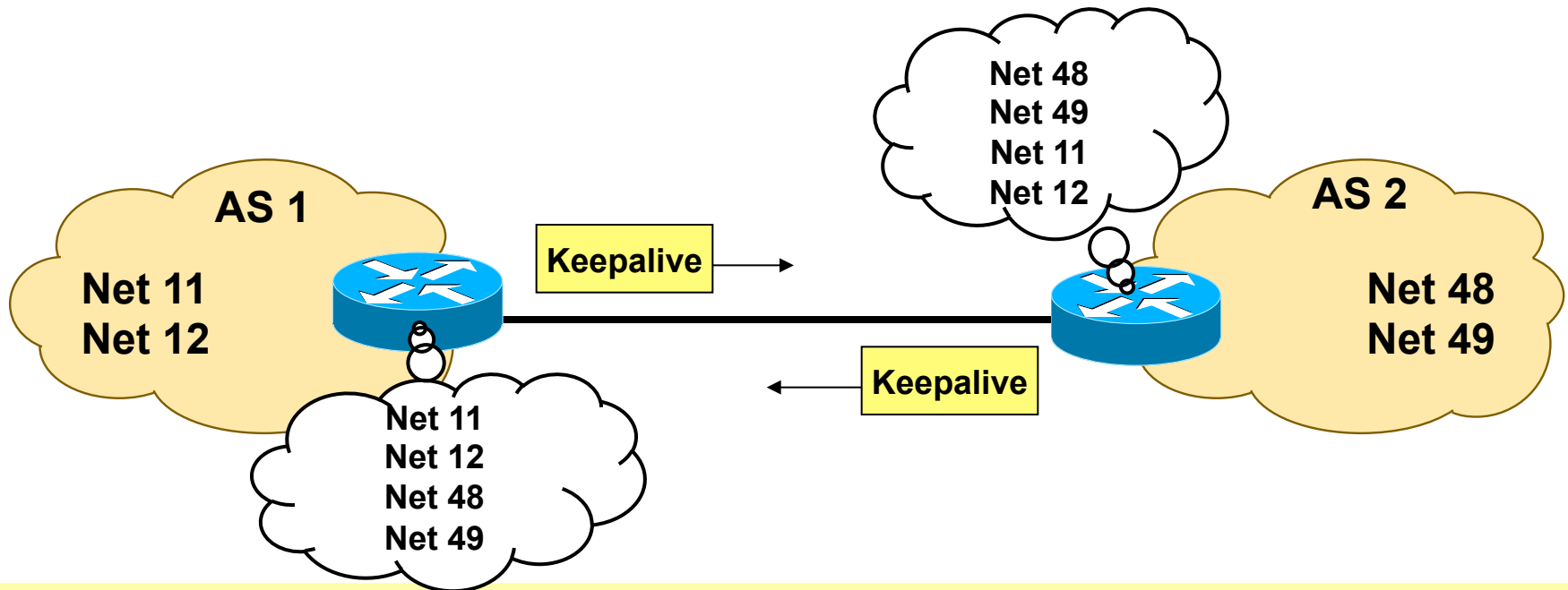
NLRI Update

- After open message, all known routes are exchanged using **update** messages
- Contains network layer reachability information (**NLRI**)
 - List of prefix and length



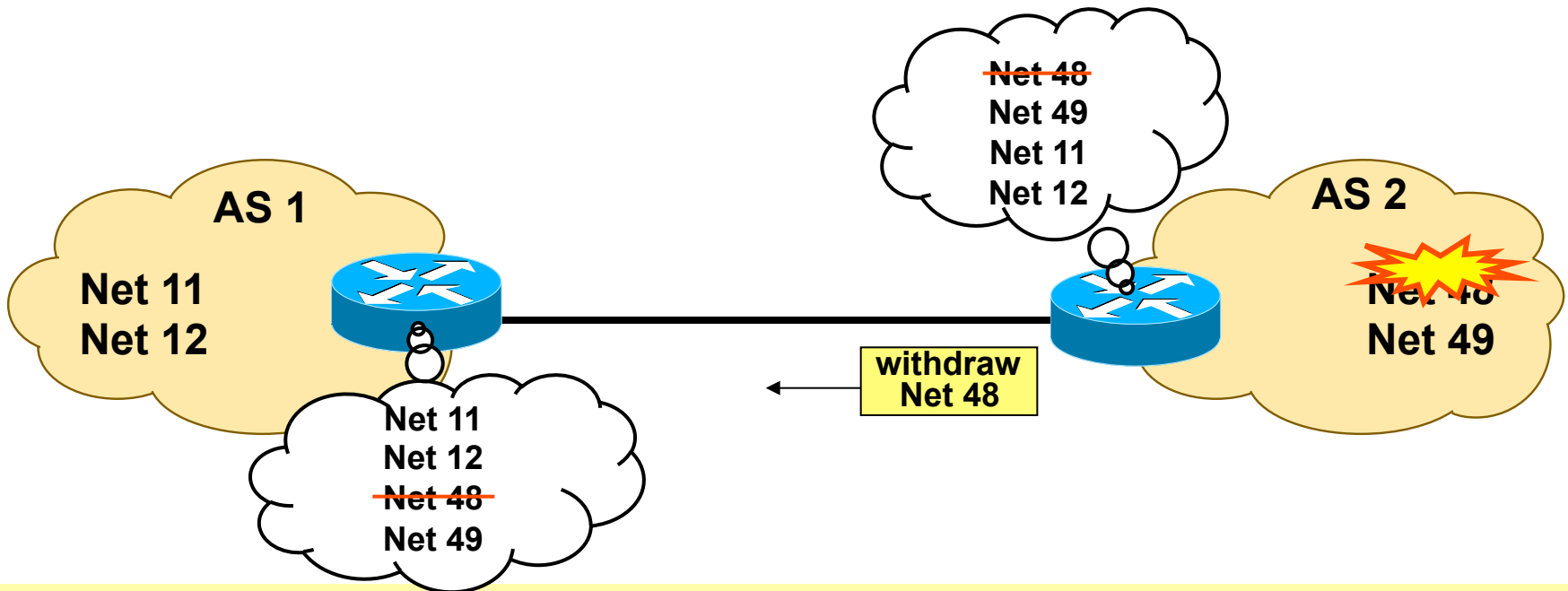
Steady State

- After Open/Update procedure, BGP is nearly **quiet** – No *periodic updates* !
- Only **keepalive** messages are sent
 - 19 Bytes
 - Per default every 60s



Topology Change:

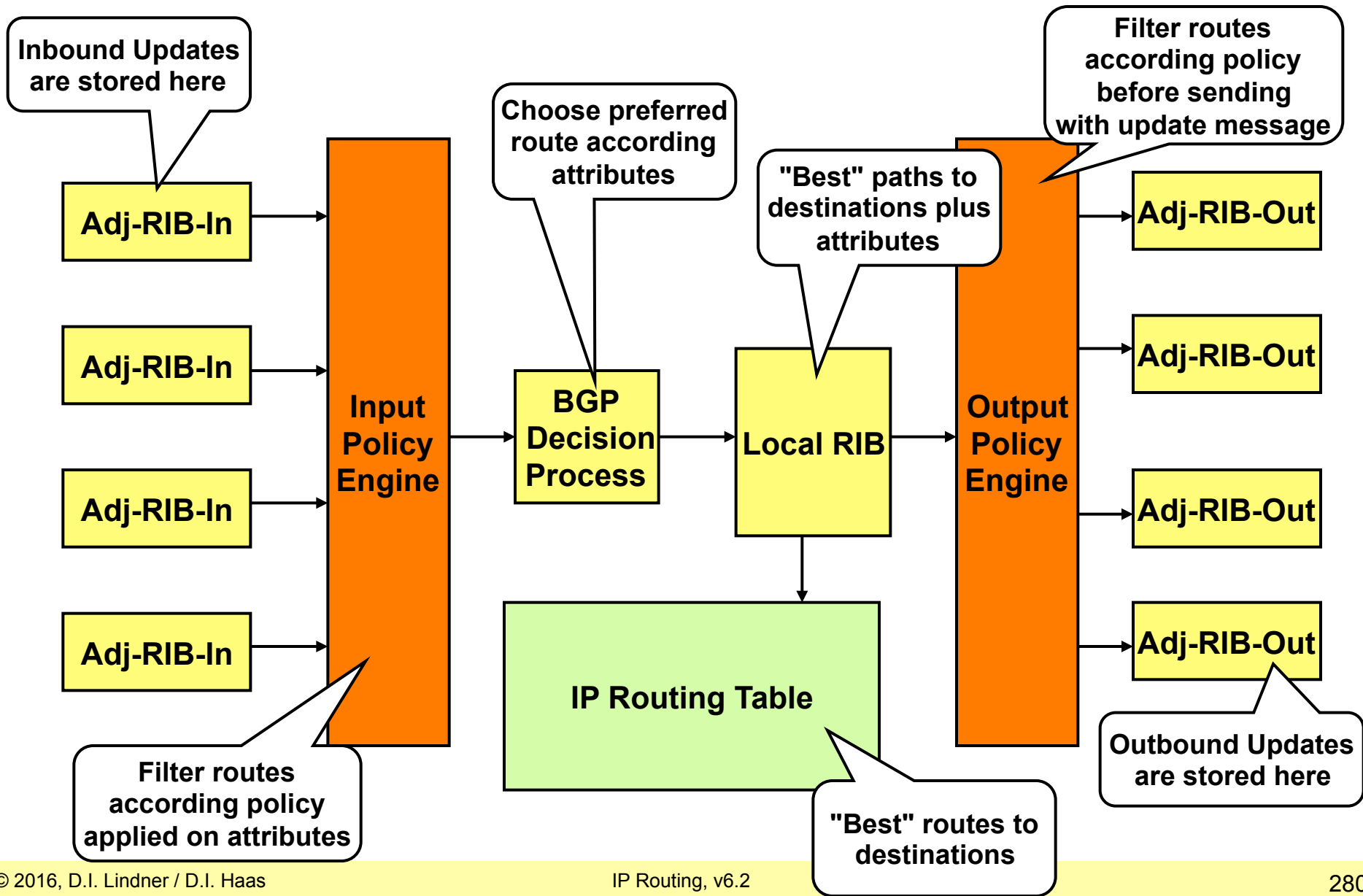
- **Incremental** Updates upon topology or attribute changes
- **Withdraw** message upon loss of network



RIB

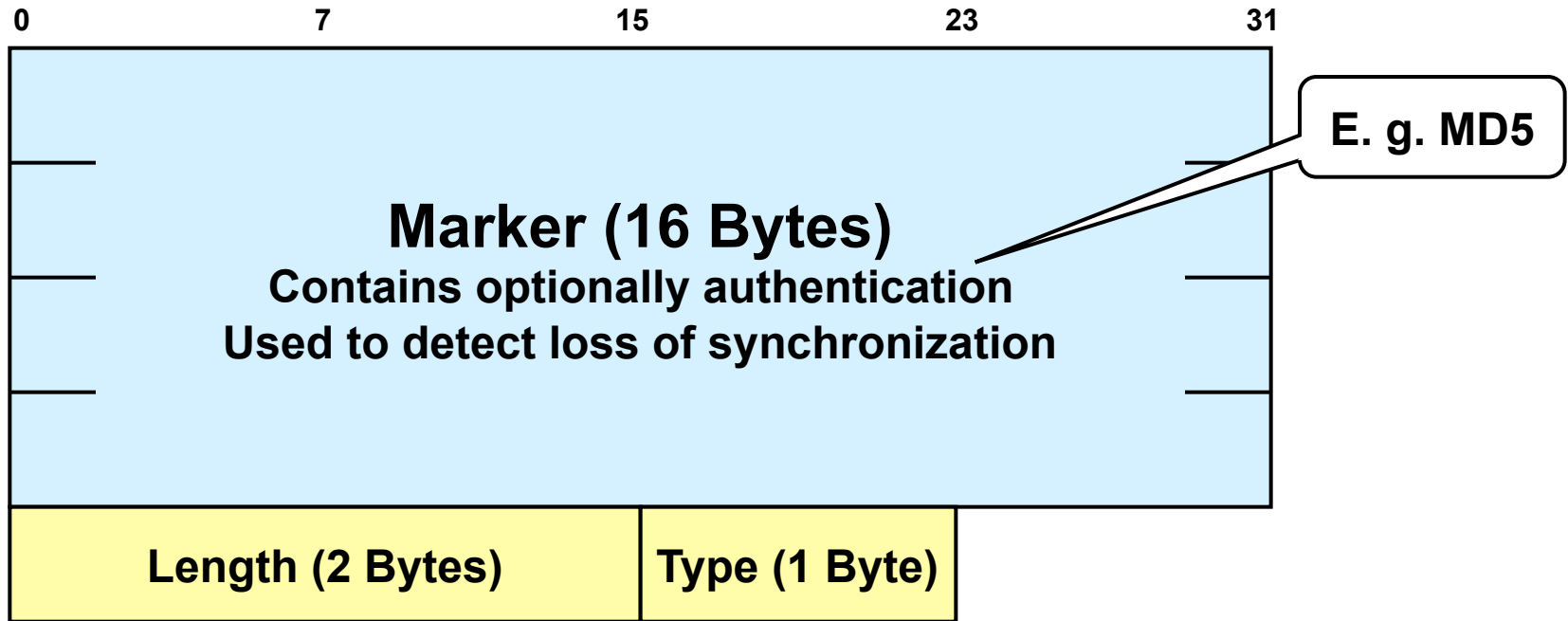
- BGP routing information is stored in RIBs
- RIBs might be combined (vendor specific)
- **Only best paths are forwarded to the neighboring ASs**
- **Alternative paths remain in the BGP table**
 - "Feasible routes" in Adj-RIB-In
 - Are used if the original path is withdrawn

BGP Routing Information Bases



BGP Header Format

FYI

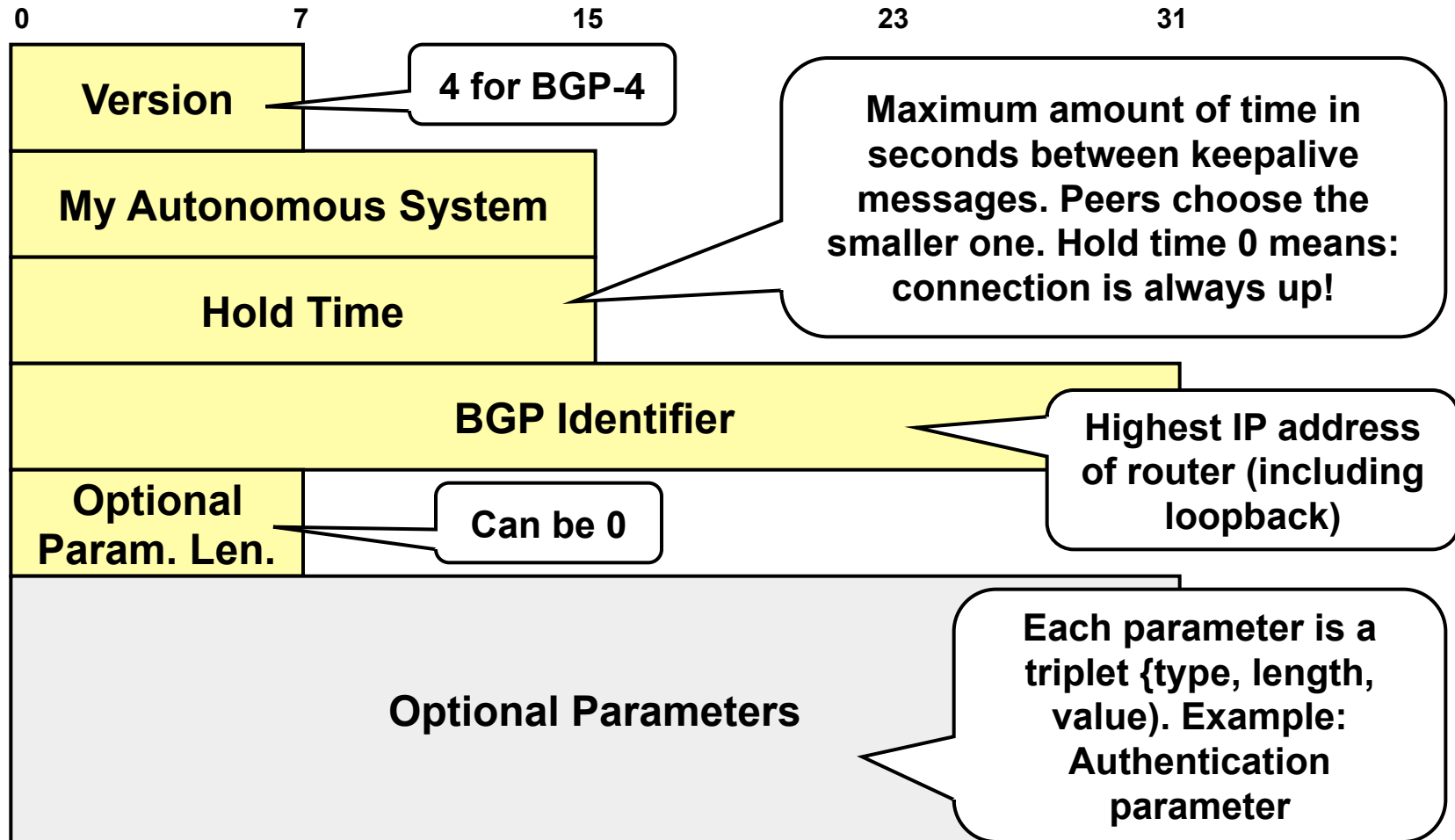


The smallest BGP message is **19 Bytes**
(no data field, e. g. keepalive)

The maximum length is **4,096 Bytes**
(also including header)

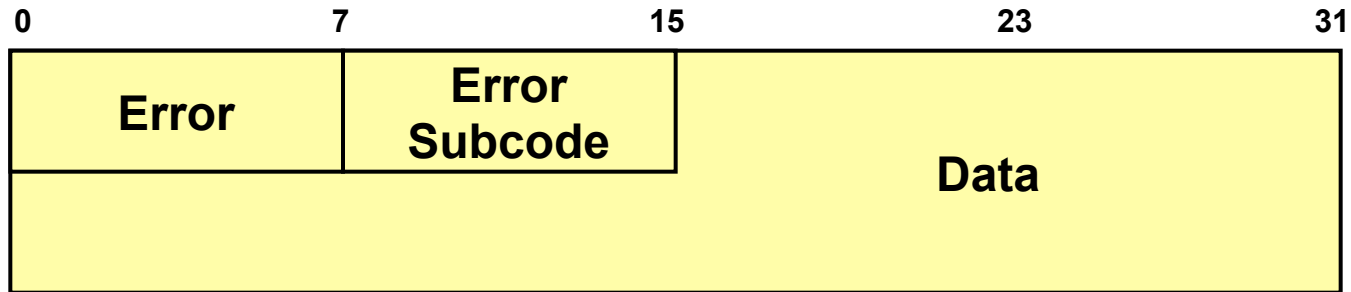
Open Message (Type 1)

FYI



Notification Message (Type 3)

FYI



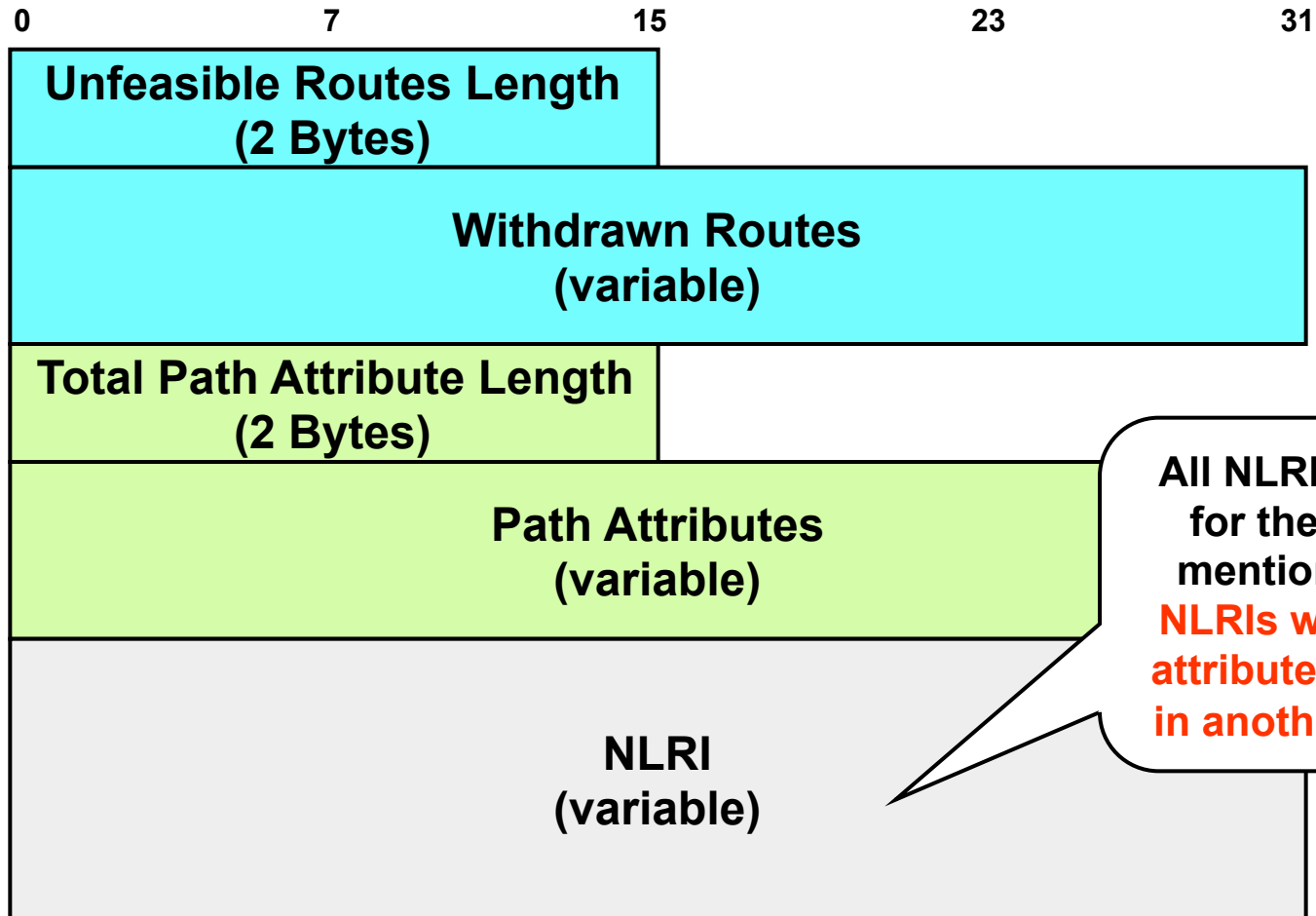
Notification is always sent when an error is detected.

After that, the connection is closed.

Keepalive Message (Type 4)

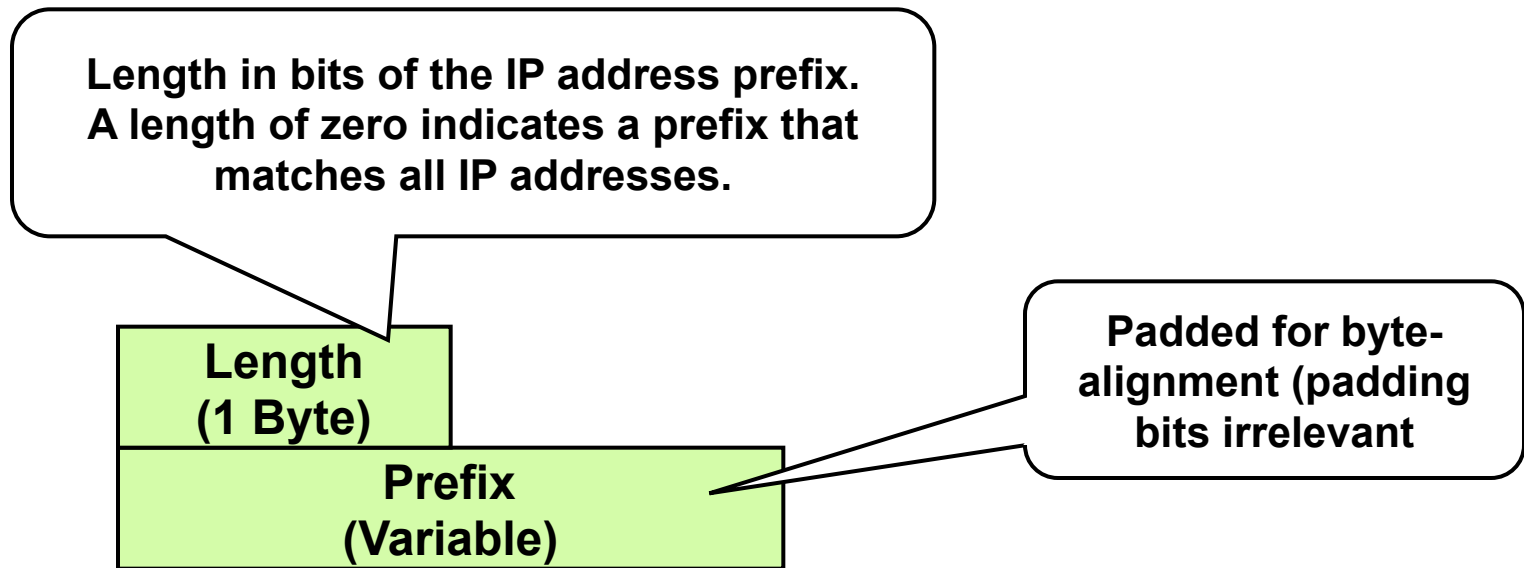
- Consists of header only (19 bytes)
- **Must be sent before hold time expires**
- Recommended keepalive rate
= 1/3 of hold time
- Not necessary if update message is sent

The Update Message (Type 2)



Withdrawn Routes

FYI

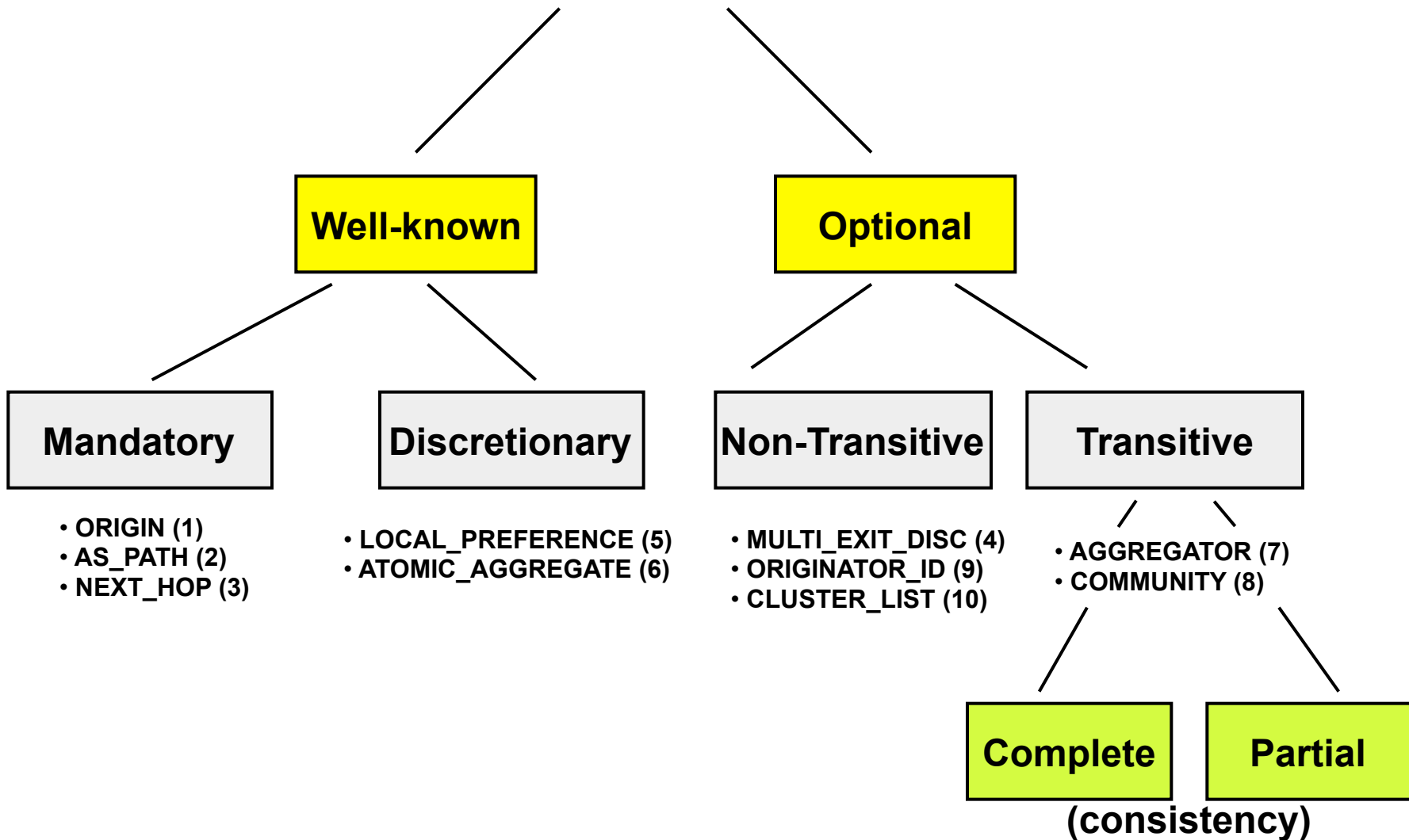


...How destinations are specified within an update

Agenda

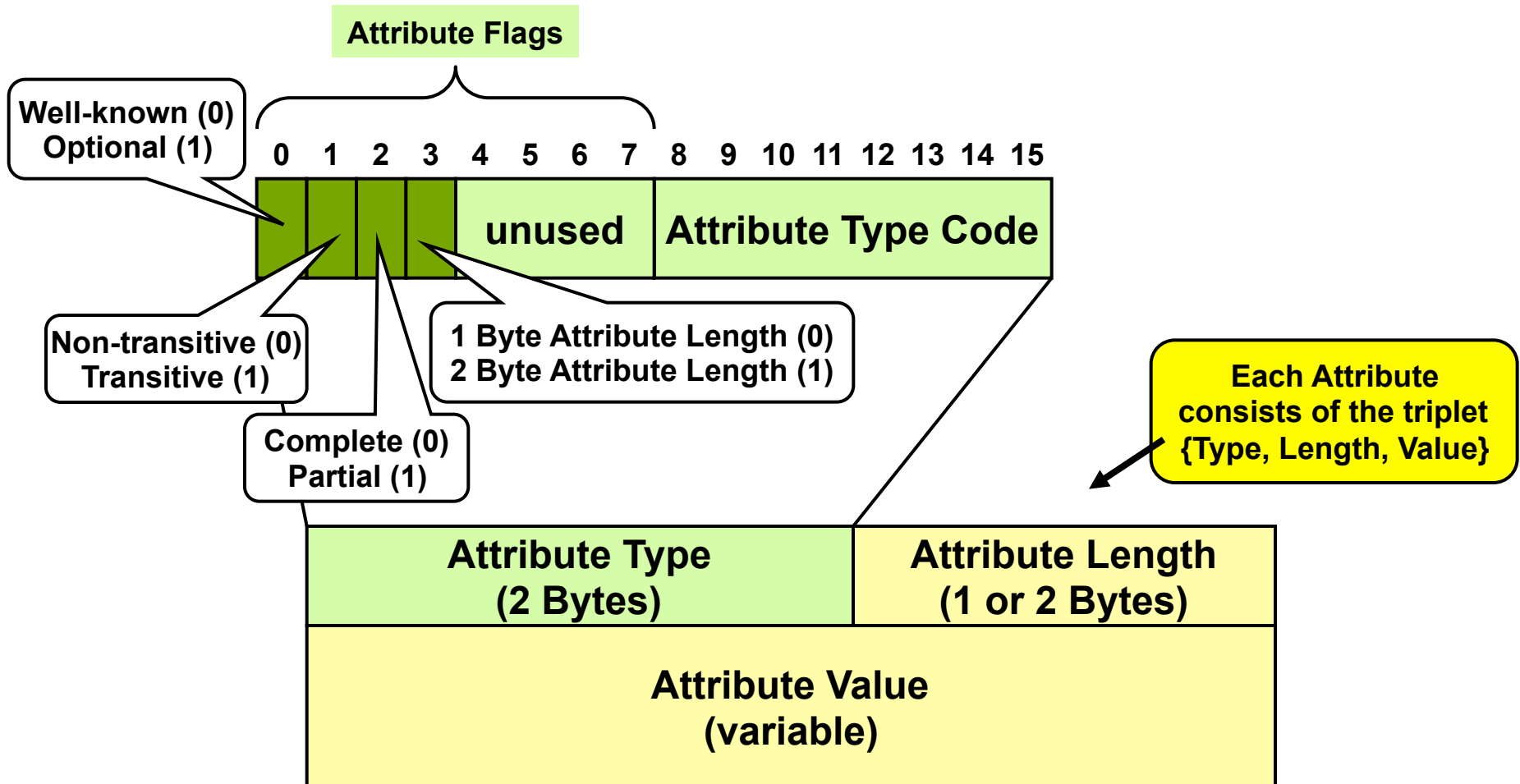
- **Introduction to IP Routing**
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**
 - Introduction
 - BGP Basics
 - BGP Attributes
 - BGP Special Topics
 - CIDR

Attribute Types



Path Attributes

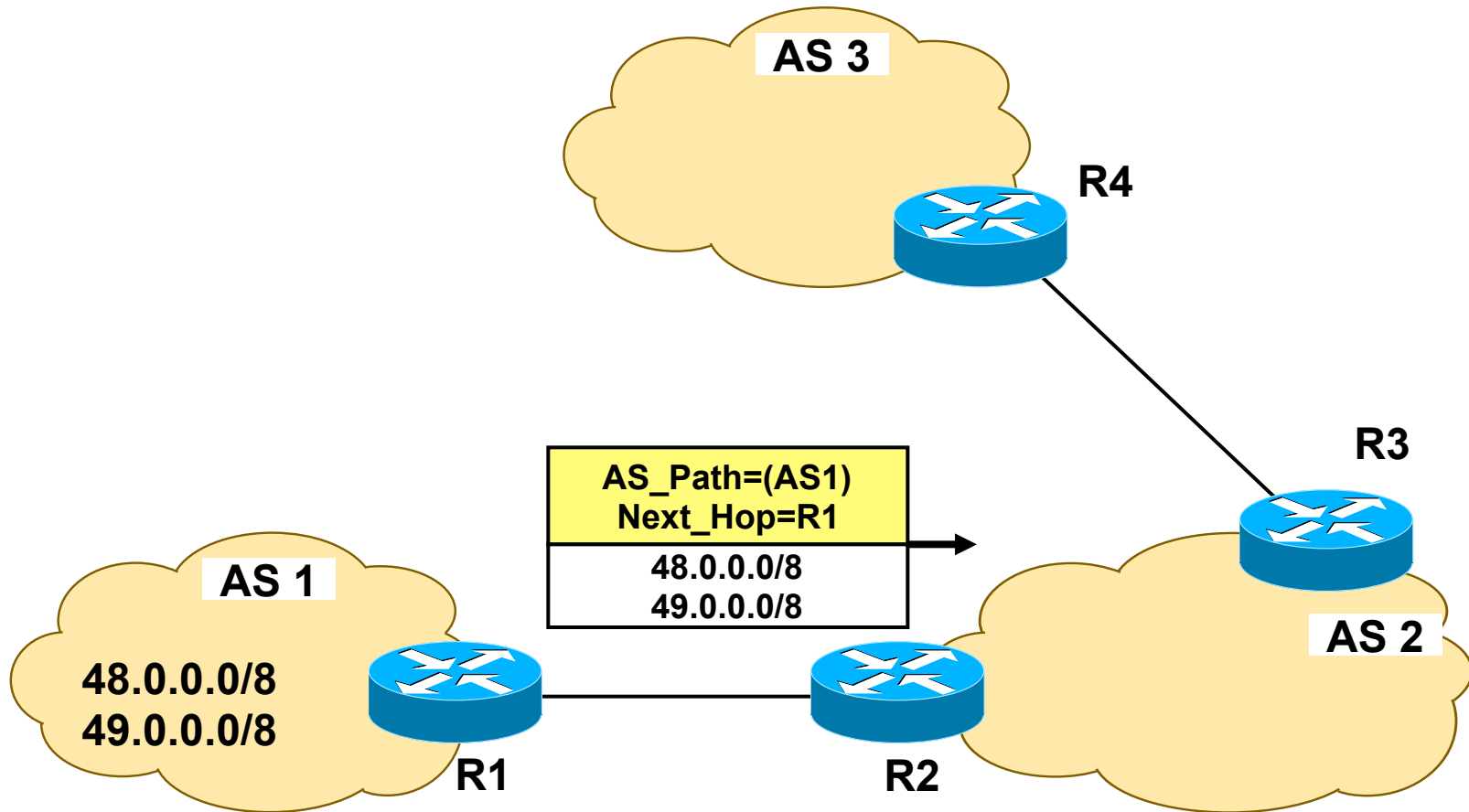
FYI



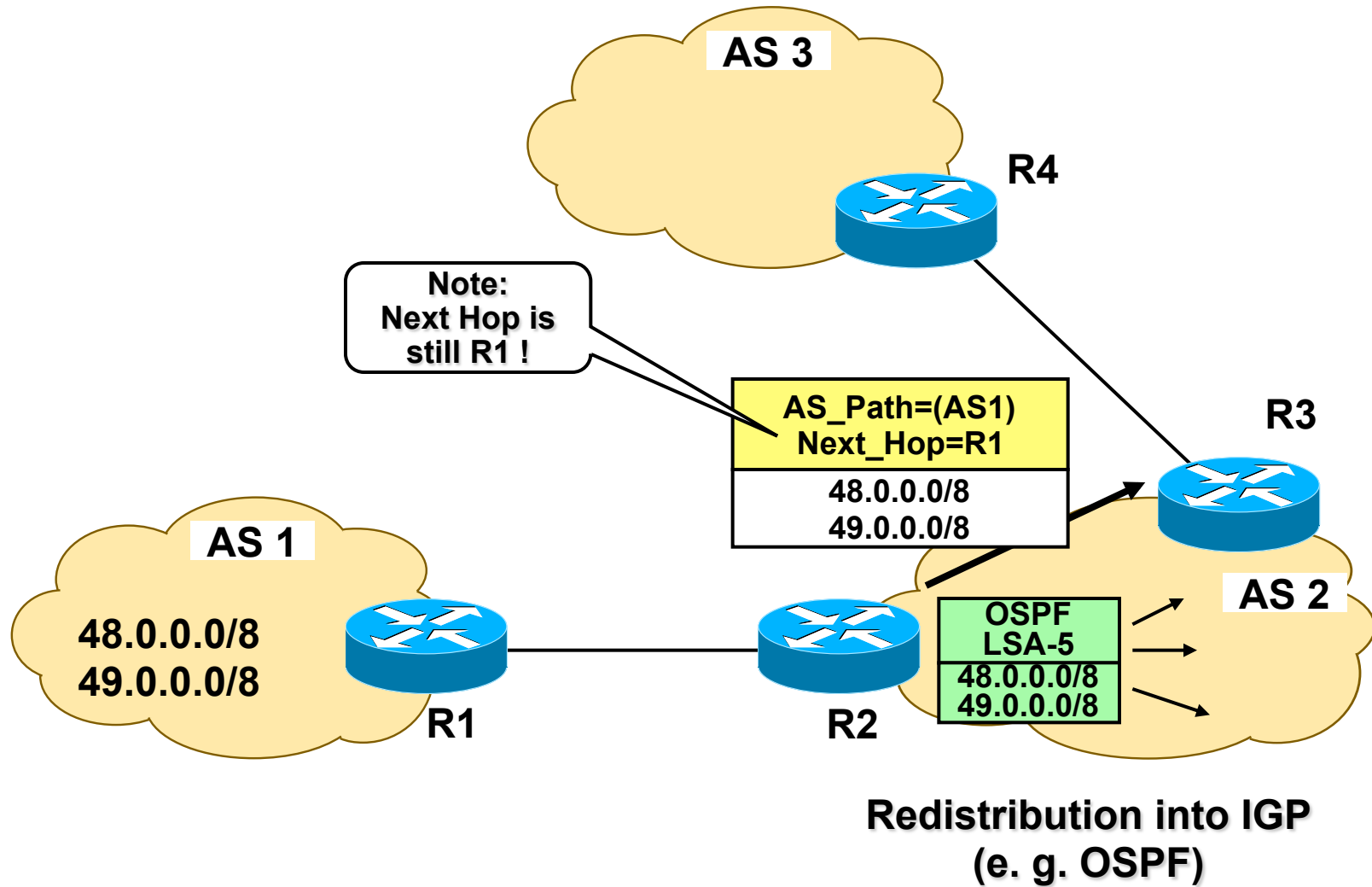
Well-known Mandatory

- **AS_Path** contains all ASs traversed for this route
- **Next_Hop** indicates the last EBGP router leading to this route
 - Not necessarily the physical next hop
- **Origin** indicates how this route was learned

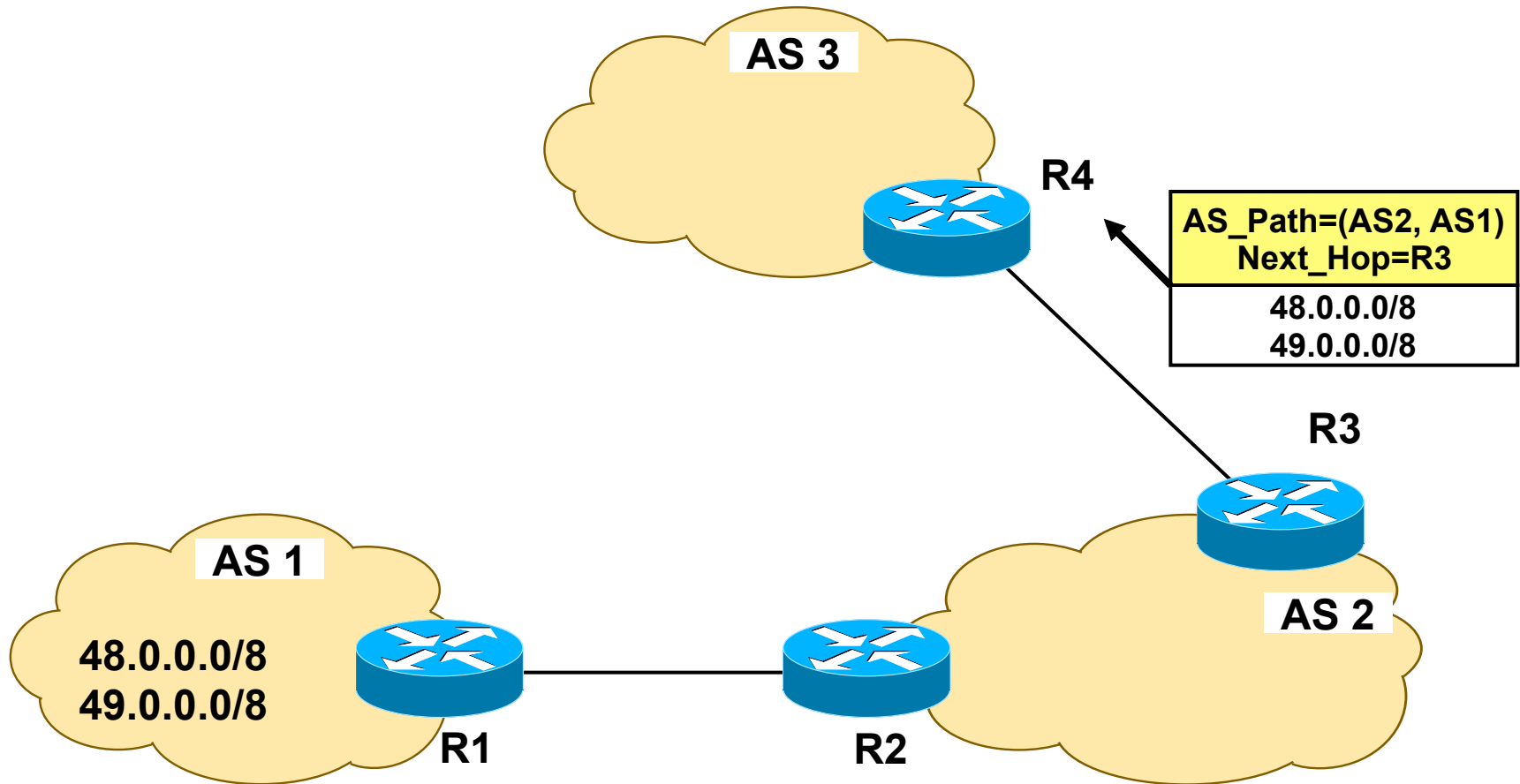
Path Vector Protocol (1)



Path Vector Protocol (2)



Path Vector Protocol (3)



1	Well-known
	Mandatory

FYI

- **Value 0: IGP**
 - Routes learned via **network statement** (NLRI is member of originating AS)
- **Value 1: EGP**
 - Learned via **redistribution from EGP to BGP**
- **Value 2: INCOMPLETE**
 - Learned via redistribution from IGP to BGP
 - Example: redistribute static (Cisco)

2	Well-known
	Mandatory

- **Composed of a sequence of AS path segments**
- **An AS path segment is represented by a triple**
 - Path segment type (1 byte)
 - 1 = **AS_Set** (unordered set of ASs)
 - 2 = **AS_Sequence** (ordered set of ASs)
 - Path segment length (1 byte)
 - Path segment value (variable, 2 bytes per AS)

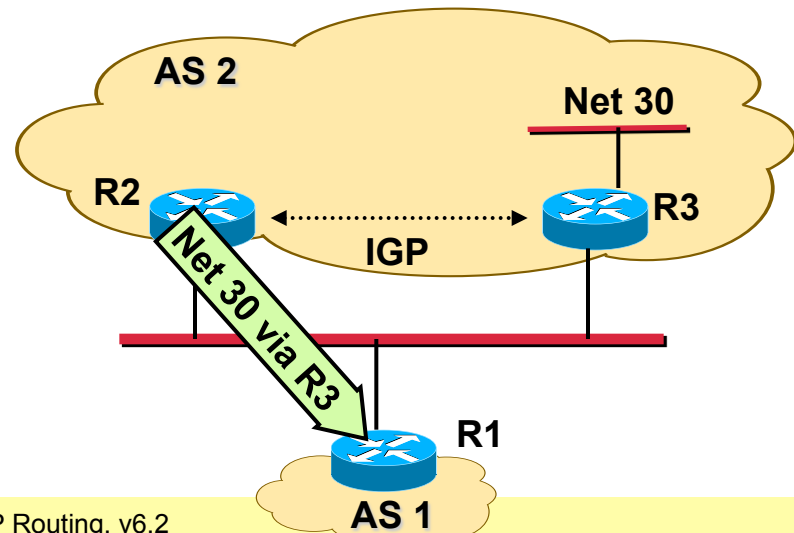
Who is NEXT_HOP?

3	Well-known
	Mandatory

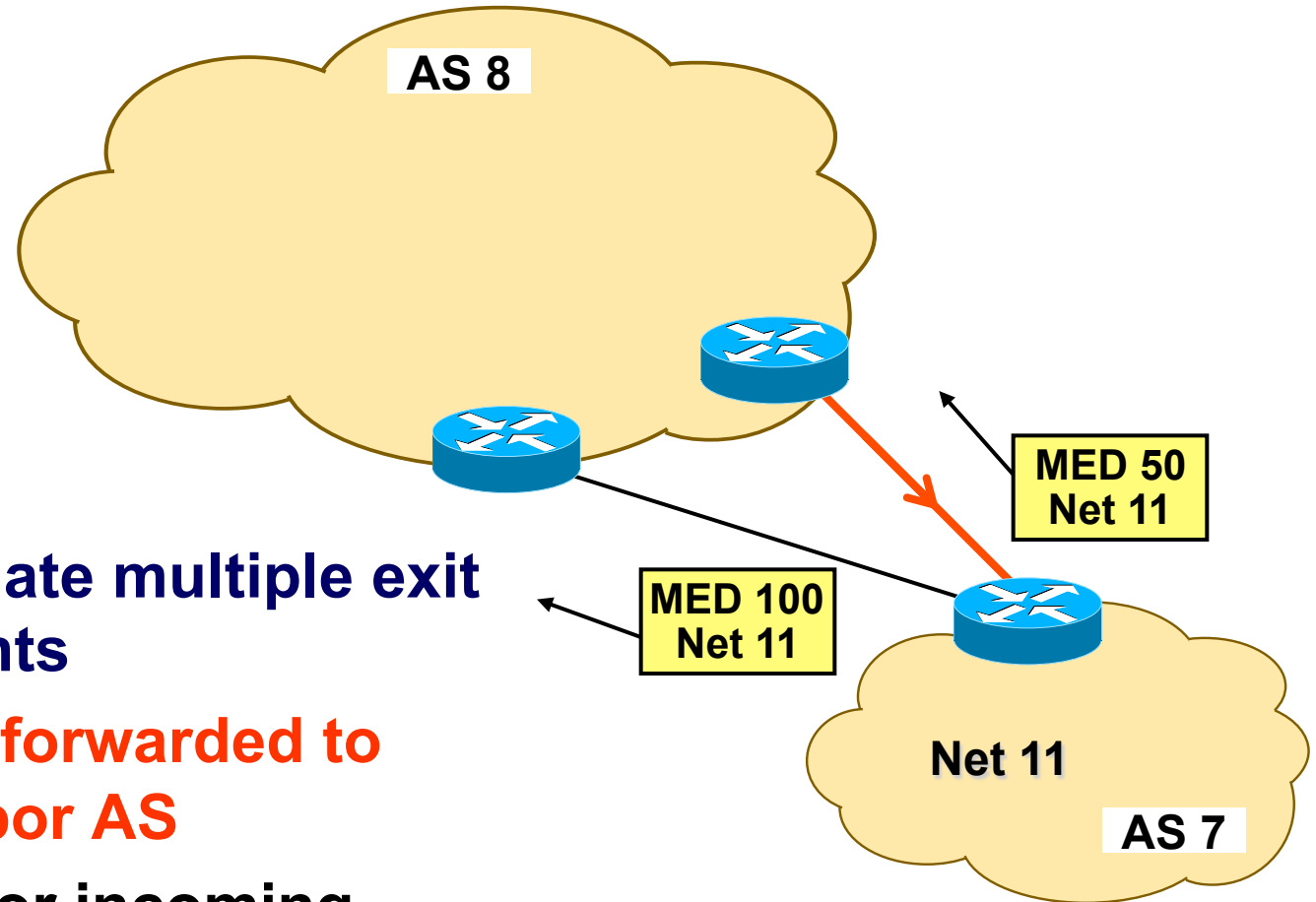
- The **boundary router** that advertized the route in this AS is the next hop
 - Recursive routing table lookup might be necessary to determine the true physical next hop
- **Exception:**
 - On multi-access media (Ethernet, FDDI) always the physical next hop must be indicated

R1 and R2 have BGP session established, R3 speaks IGP only.

R2 advertises R3 as next hop to Net 30 because R3 is on the same physical media.



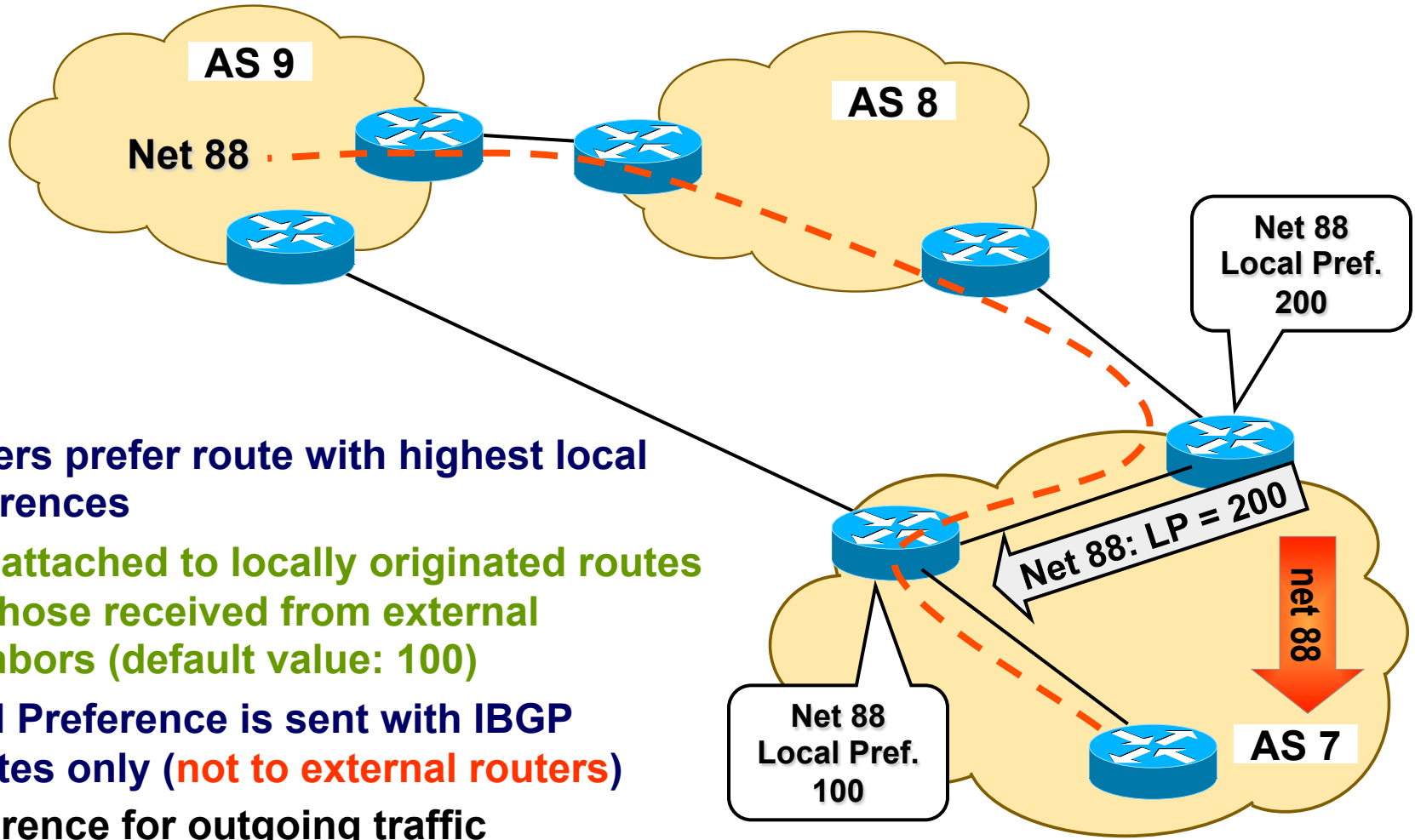
4	Optional
	Non-transitive



- To discriminate multiple exit or entry points
- **Must not be forwarded to other neighbor AS**
- Preference for incoming traffic

LOCAL_PREF

5	Well-known
	Discretionary



- Routers prefer route with highest local preferences
- Only attached to locally originated routes and those received from external neighbors (default value: 100)
- Local Preference is sent with IBGP updates only (not to external routers)
- Preference for outgoing traffic

6	Well-known
	Discretionary

- Optionally the **Atomic_Aggregate** attribute indicates that some BGP router made an AS aggregation
 - When selecting the less specific route on overlapping routes (rejecting the more specific route)
- **Length 0**

AGGREGATOR

FYI

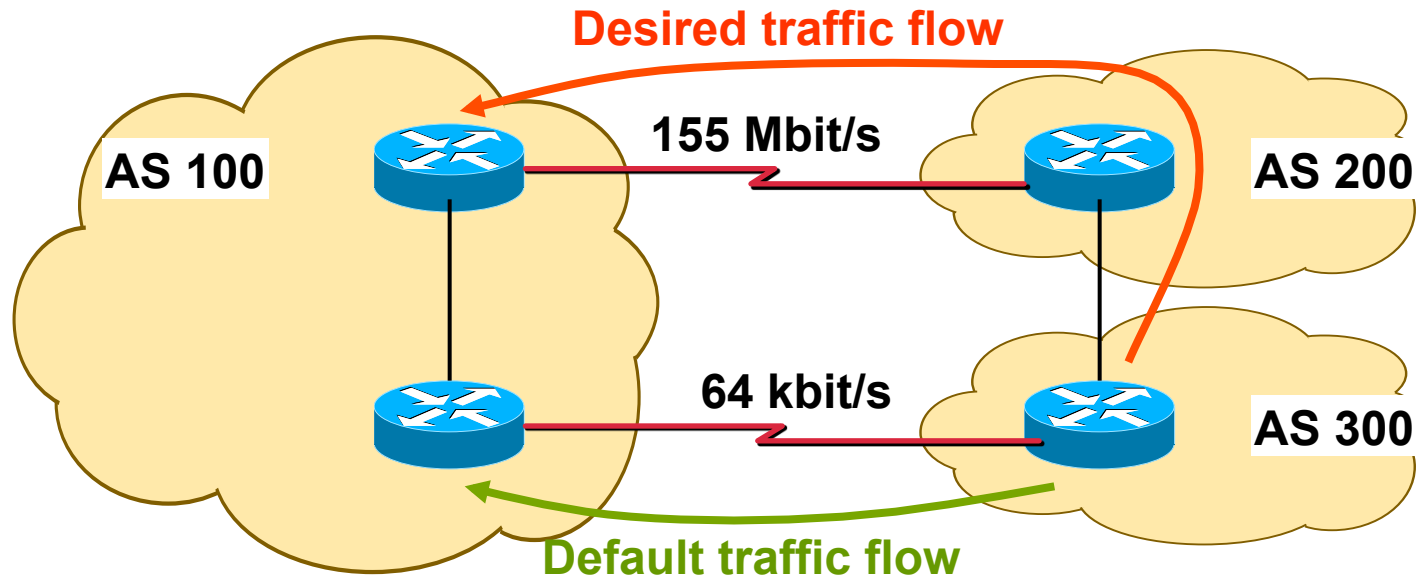
7	Optional
	Transitive

- Contains the AS number and IP address of the BGP speaker that formed the aggregate route
- Useful for troubleshooting

8	Optional
	Transitive

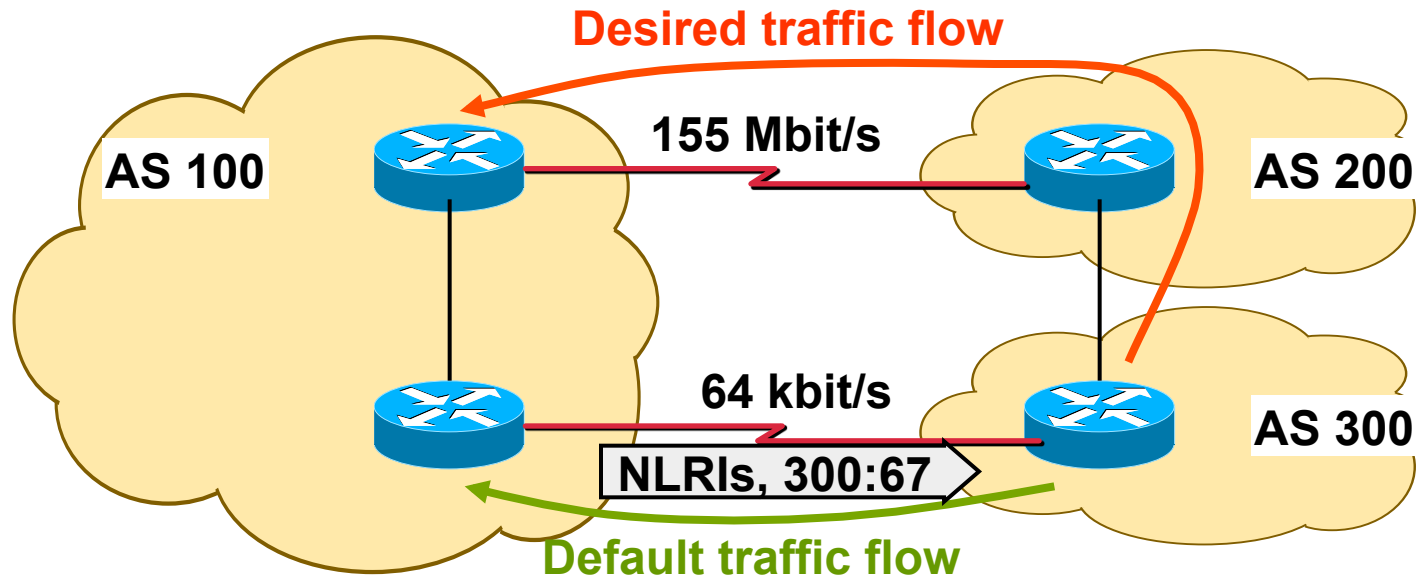
- **Group of destinations that share a common policy**
 - Each destination could be member of multiple communities
 - Carried across ASs
- **Community strings are simple policy labels**
 - Any BGP router can **tag** routes in incoming and outgoing routing updates or when doing redistribution
 - Any BGP router can **filter** routes in incoming or outgoing updates or select preferred routes based on communities

Community Example (1)



- **Assume AS 100 wants AS 300 to use the 155 Mbit/s link to reach own networks**
 - MED: not possible (non-transitive)
 - Local Preference: will admin of AS 300 set it?
- **Best and easiest: Use community !**

Community Example (2)



- Receiving a community string means "apply the predefined policy"
- In our example 300:67 means: "set local preference to 50"

- **More than one BGP community per route allowed**
 - By default, communities are stripped in outgoing BGP updates
- **Private range:**
0x00010000 - 0xFFFEFFFF
- **Common practice**
 - High order 16 bit: **AS number**
 - Low order 16 bit: **Local significance**

- **Reserved ranges:** 0x00000000 - 0x0000FFFF and 0xFFFF0000 - 0xFFFFFFFF
- **0xFFFFFFFF01 means: NO_EXPORT**
 - Routes received carrying this value should not be advertised to EBGP peers, except ASs of a confederation
- **0xFFFFFFFF02 means: NO_ADVERTISE**
 - Routes received carrying this value should not be advertised at all (both IBGP and EBGP peers)
- **0xFFFFFFFF03 means: NO_EXPORT_SUBCONFED**
 - Routes received carrying this value should not be advertised to EBGP peers, including members of a confederation (Cisco: LOCAL_AS)

Administrative Weight (Cisco)

FYI

- **No attribute – just a **local** parameter**
- **Applies only to routes within an individual router**
- **Number between 0 and 65535**
 - The higher the weight the more preferable the route
- **Initially invented to translate public routing policies (EGP)**

Decision Hierarchy

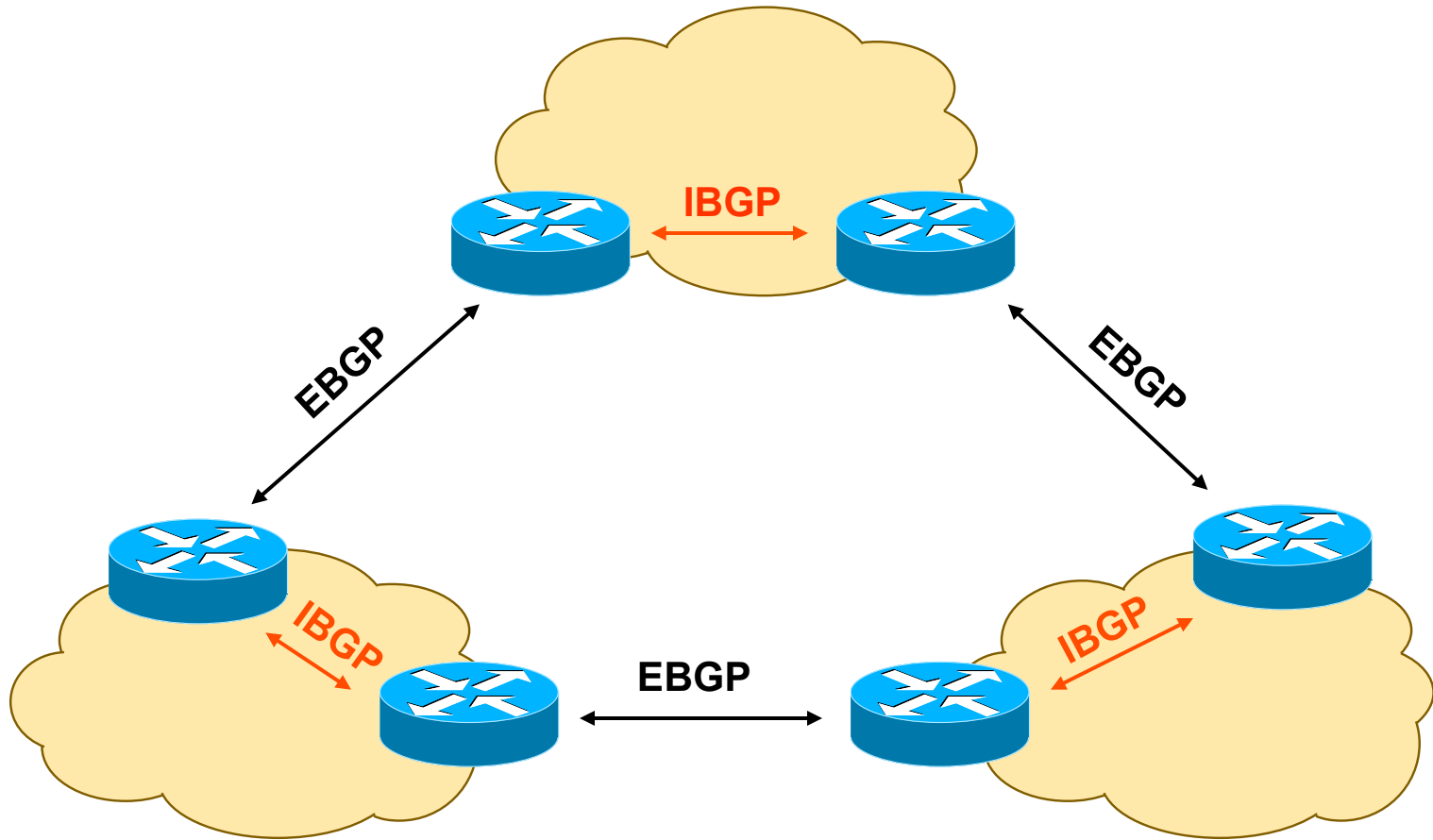
FYI

- 1. Prefer highest weight (Cisco)**
- 2. Prefer highest local preference**
- 3. Prefer locally originated routes**
- 4. Prefer shortest AS-Path**
- 5. Prefer lowest origin code**
- 6. Prefer lowest MED**
- 7. Prefer EBGP path over IBGP path**
- 8. Lowest IGP metric to next hop**
- 9. Prefer oldest route for EBGP paths**
- 10. Prefer path with lowest neighbor BGP router ID**

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP)**
 - Introduction
 - BGP Basics
 - BGP Attributes
 - BGP Special Topics
 - CIDR

EBGP and IBGP



Internal and External BGP

- **EBGP messages are exchanged between peers of different ASs**
 - EBGP peers should be directly connected
- **Inside an AS this information is forwarded via IBGP to the next BGP router**
 - IBGP messages have same structure like EBGP messages
- **Administrative Distance**
 - IBGP: 200
 - EBGP: 20 (preferred over all IGP)

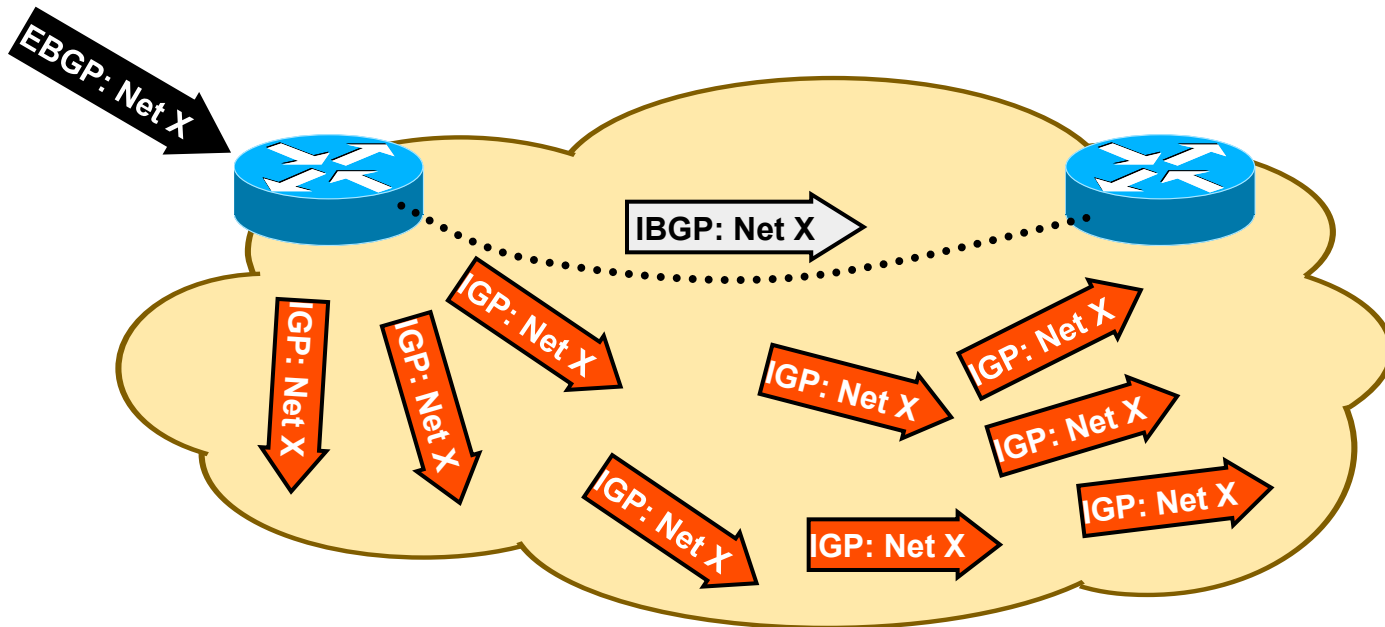
Loop Detection



- **Update is only forwarded if own AS number is not already contained in AS_Path**
- **Thus, routing loops are avoided easily**
- **But this principle doesn't work with IBGP updates (!)**
- **Therefore IBGP speaking routers must be fully meshed !!!**

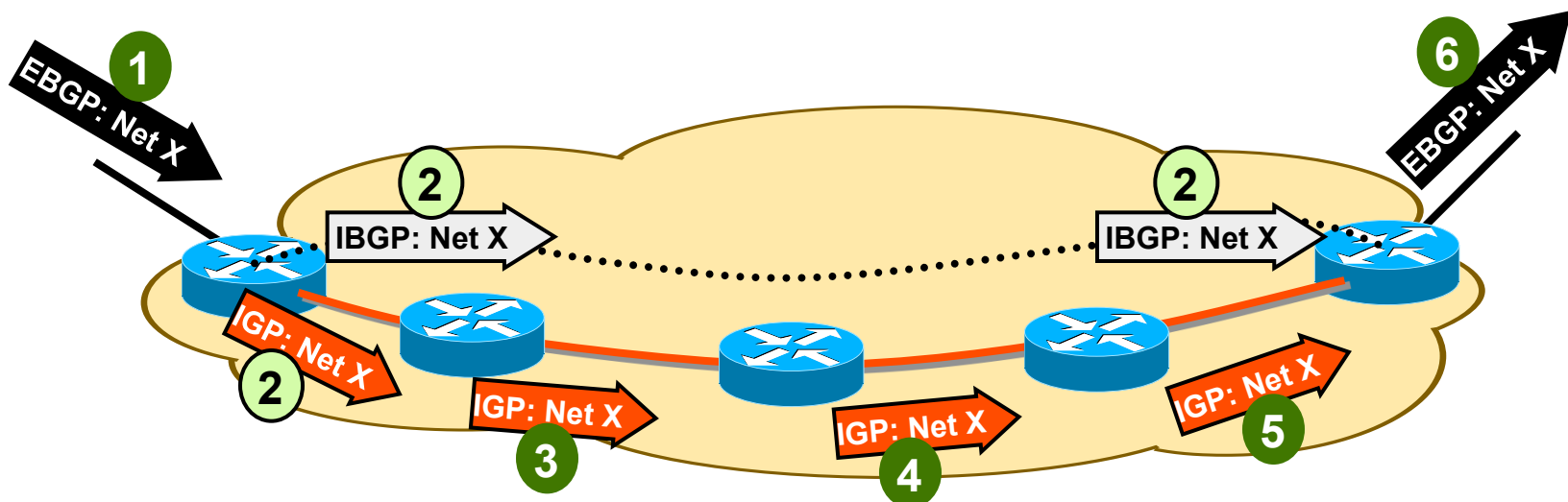
BGP → IGP Redistribution

- **Only routes learned via EBGP are redistributed into IGP**
 - To assure optimal load distribution
 - Cisco-IOS default filter behavior



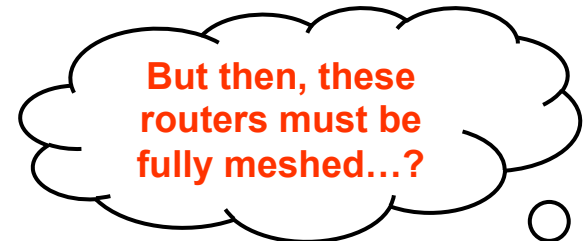
Synchronization With IGP

- **Routes learned via IBGP may only be propagated via EBGP if same information has been also learned via IGP**
 - That is, same routes also found in routing table (= are really reachable)
- **Without this "IGP-Synchronization" black holes might occur**



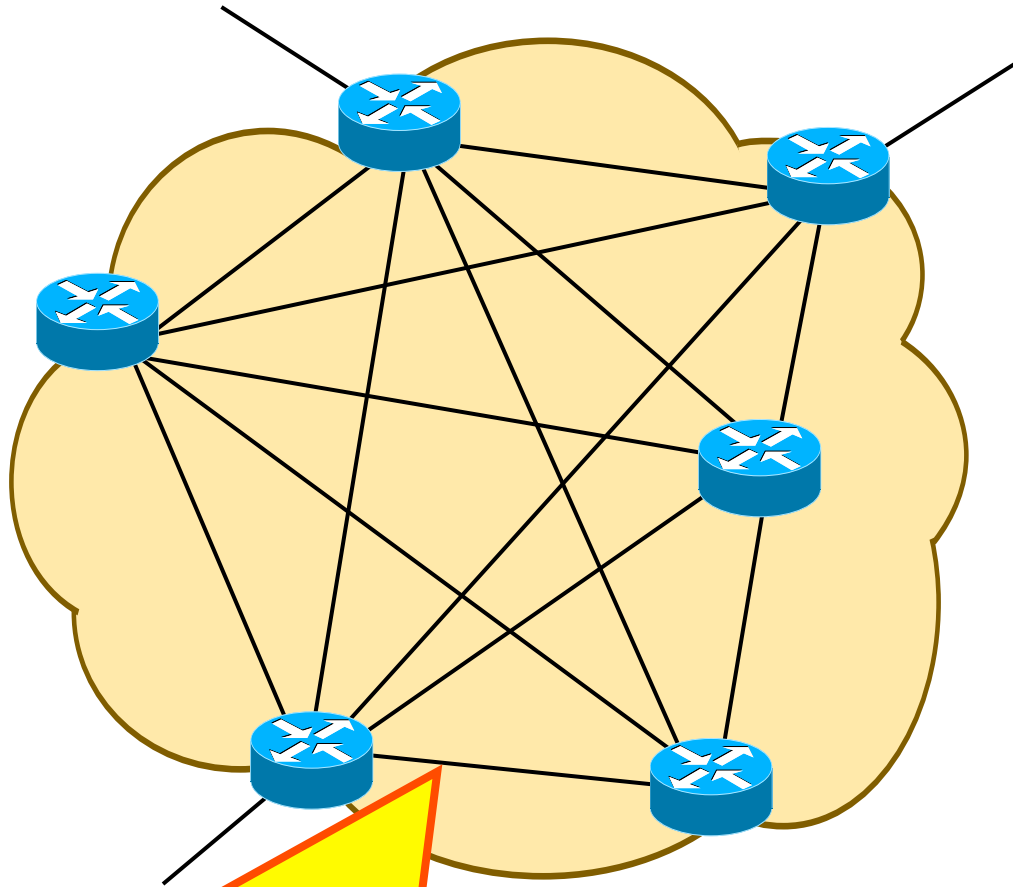
Avoid Synchronization

- **Synchronization with IGP means injecting thousands of routes into IGP**
 - IGP might get overloaded
 - Synchronization dramatically affects BGP's convergence time
- **Alternatives**
 - Set default routes leading to BGP routers (might lead to suboptimal routing)
 - **Use only BGP-routers inside the AS !**



Fully Meshed IBGP Routers

FYI

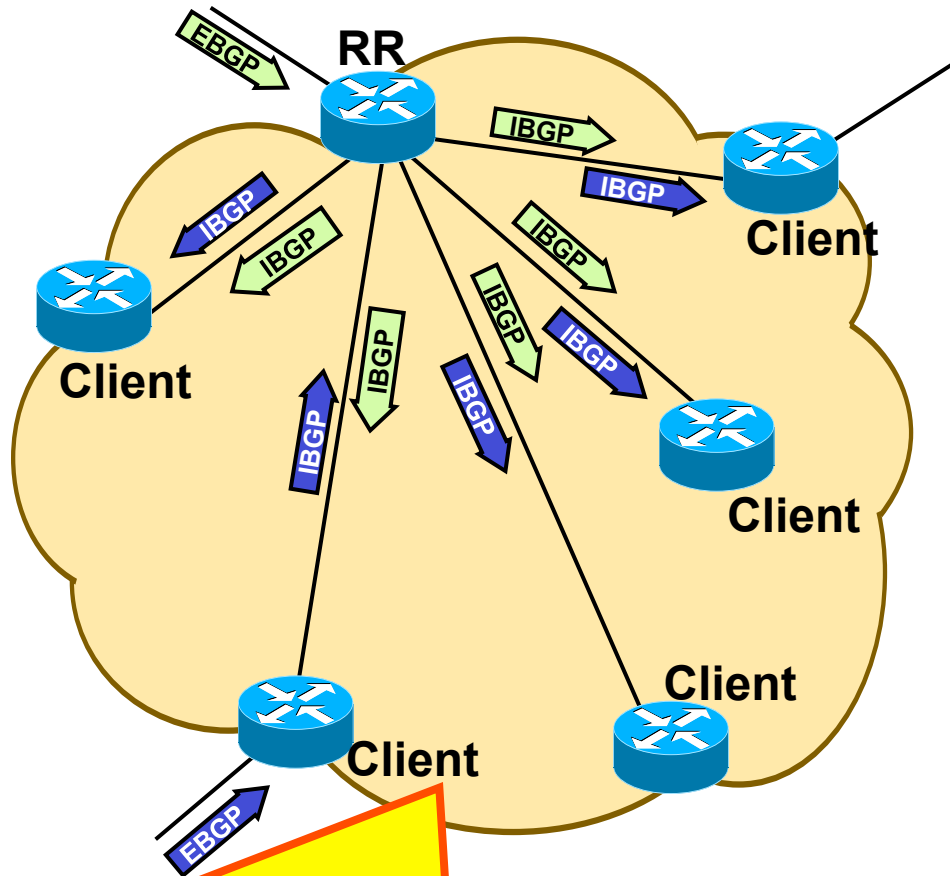


- **Does not scale**
 - $n(n-1)/2$ links
- **Resource and configuration challenge**
- **Solutions:**
 - Route Reflectors
 - Confederations

Note: These are **logical IBGP connections!
The physical topology might look different!**

Route Reflector

FYI

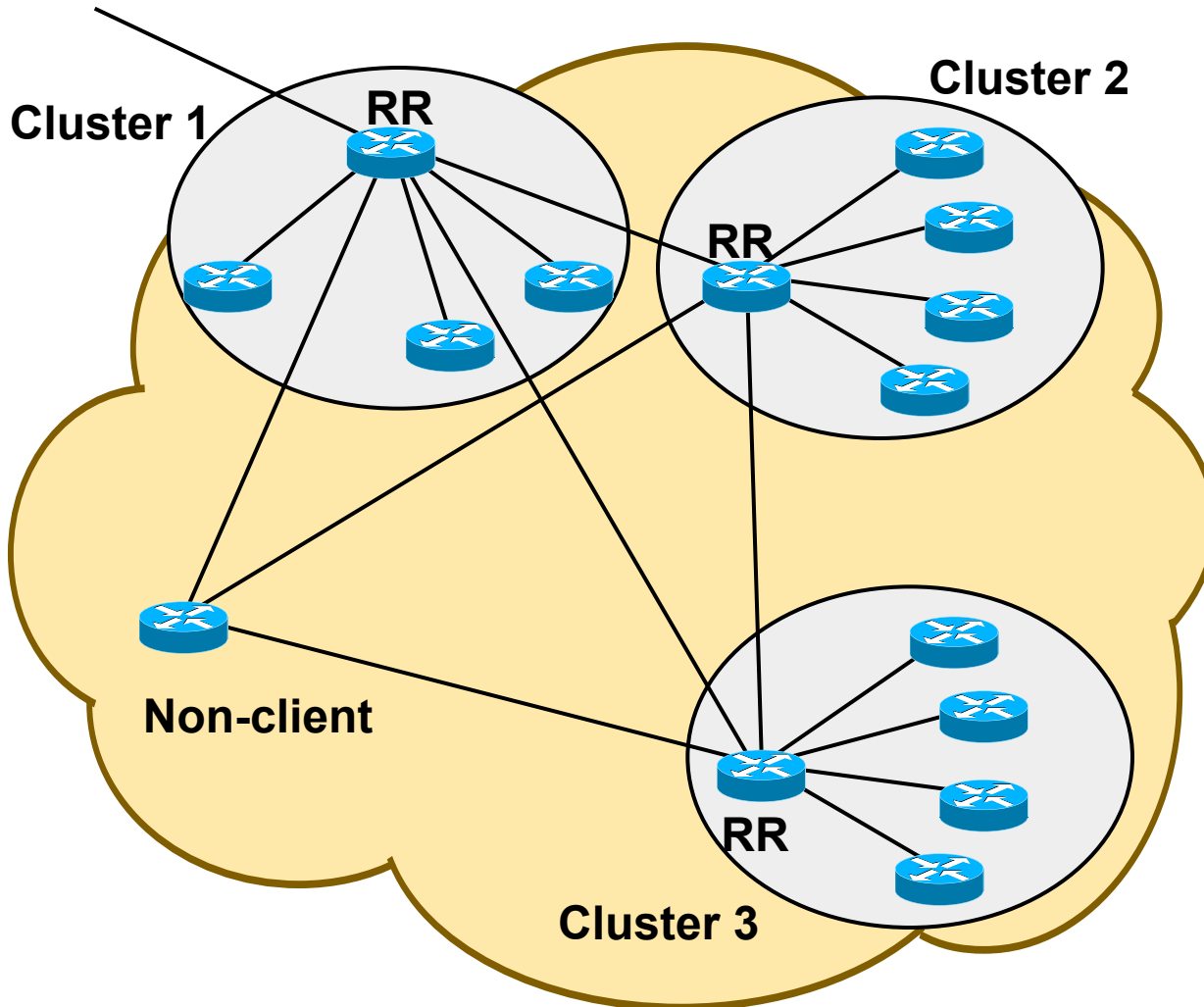


- RR mirrors BGP messages for "clients"
- RR and clients belong to a "cluster"
- Only RR must be configured
 - Clients are not aware of the RR

Note: Although these are logical IBGP connections, the physical topology should be the **main indicator** for an efficient cluster design (which router becomes RR)

RR Clusters

FYI



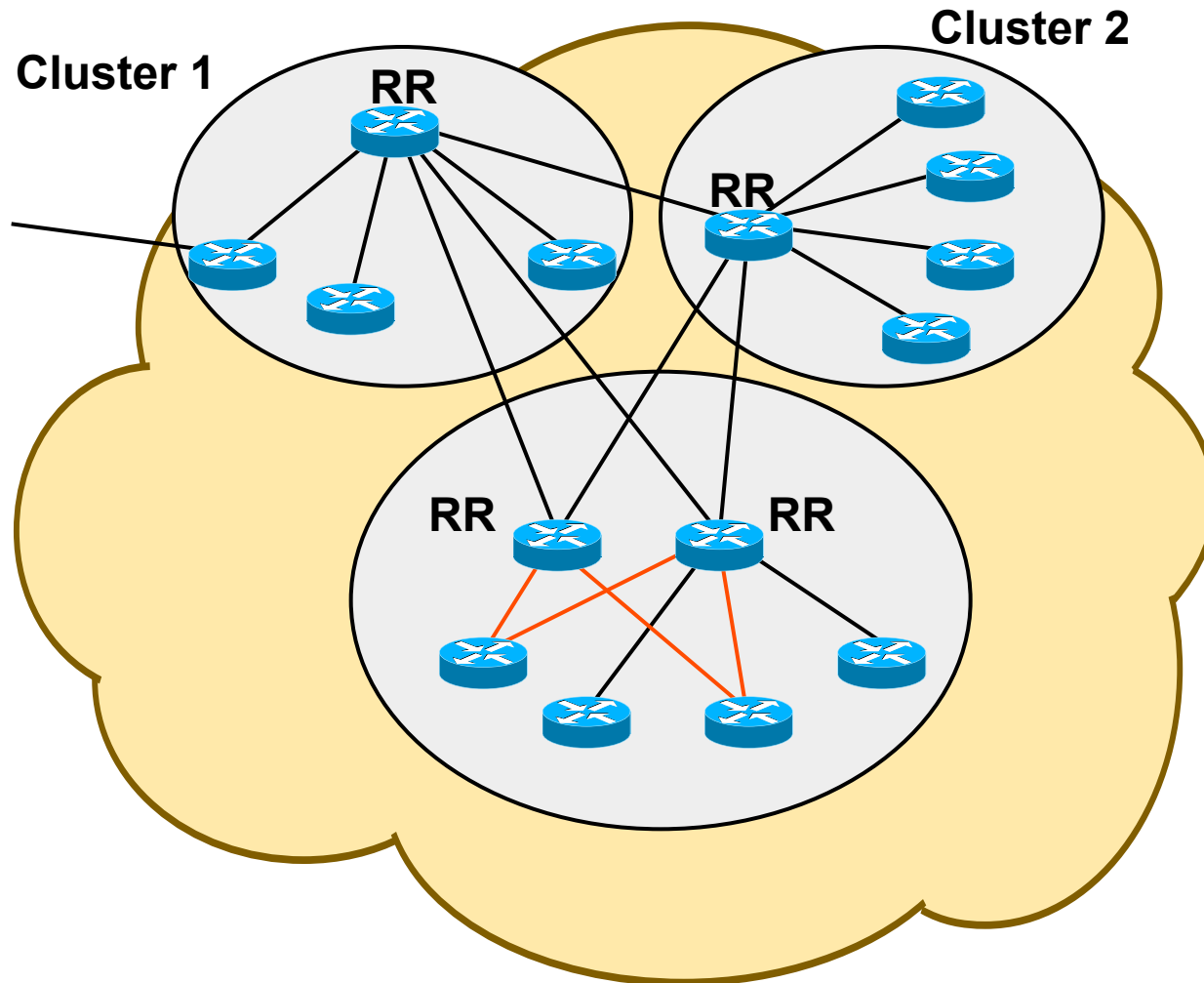
- Only RRs are fully meshed
- Special Attributes care for loop-avoidance
- **"Non-clients"** must be fully meshed with RRs
 - And with other non-clients

- RRs do **not change** IBGP behavior or attributes
- RRs only propagate **best routes**
- Special attributes to avoid routing updates **reentering the cluster (routing loops)**
 - **ORIGINATOR_ID**

Contains router-id of the route's originator in the local AS; attached by RR (Optional, Non-Trans.)
 - **CLUSTER_LIST**

Sequence of cluster-ids; RR appends own cluster-id when route is sent to non-clients outside the cluster (Optional, Non-Transitive)

Redundant RRs

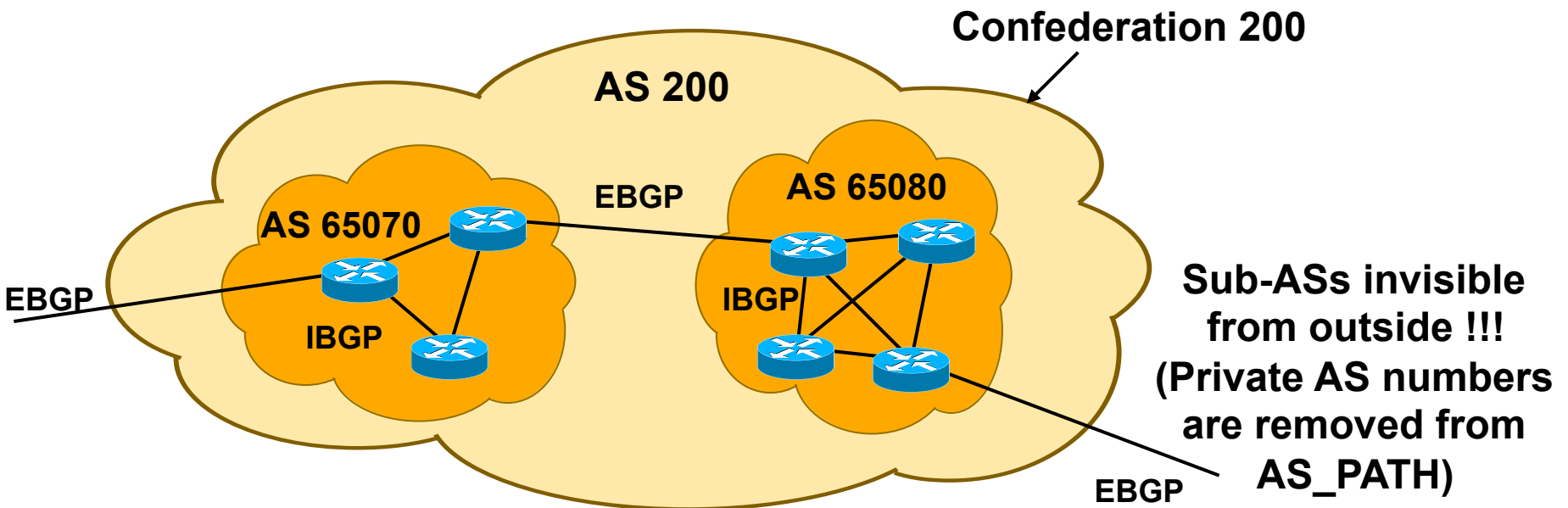


- **RR is single point of failure**
 - Other than fully meshed approach
- **Redundant RRs can be configured**
 - Clients attached to several RRs

Confederations

FYI

- Alternative to route reflectors
- Idea: AS can be broken into multiple sub-ASs
- Loop-avoidance based on AS_Path
- All BGP routers inside a sub-AS must be fully meshed
- EBGP is used between sub-ASs



RRs versus Confederations

FYI

- **RRs are more popular**
 - Simple migration (only RRs needs to be configured accordingly)
 - Best scalability
- **Confederations drawbacks**
 - Introducing confederations require complete AS-renumbering inside an AS
 - Major change in logical topology
 - Suboptimal routing (Sub-ASs do not influence external AS_PATH length)
- **Confederations benefits**
 - Can be used with RRs
 - Policies could be applied to route traffic between sub-ASs

Agenda

- Introduction to IP Routing
- RIP
- OSPF
- Introduction to Internet Routing (BGP, CIDR)
 - Introduction
 - BGP Basics
 - BGP Attributes
 - BGP Special Topics
 - CIDR **FYI**

Early IP Addressing

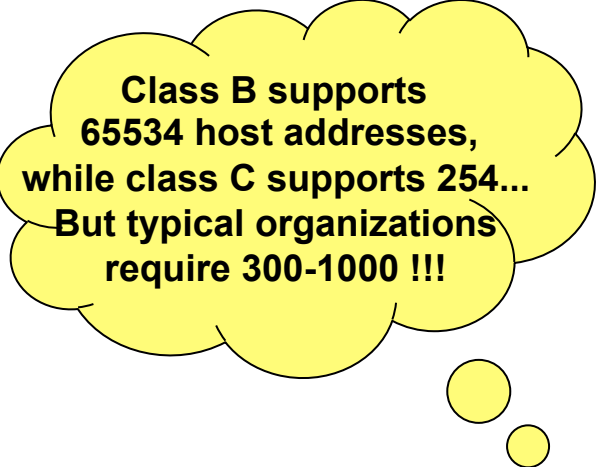
- **Before 1981 only class A addresses were used**
 - Original Internet addresses comprised 32 bits (8 bit net-id = 256 networks)
- **In 1981 RFC 790 (IP) was finished and classes were introduced**
 - 7 bit class A networks
 - 14 bits class B networks
 - 21 bits class C networks

Address Classes

- **From 1981-1993 the Internet was Classful (!)**
- **Early 80s: Jon Postel volunteered to maintain assigned network addresses**
 - Paper notebook
- **Internet Registry (IR) became part of IANA**
- **Postel passed his task to SRI International**
 - Menlo Park, California
 - Called Network Information Center (NIC)

Classful – Drawbacks

- **"Three sizes *don't* fit all" !!!**
 - Demand to assign as little as possible
 - Demand for aggregation as many as possible
- **Assigning a whole network number**
 - Reduces routing table size
 - But wastes address space



**Class B supports
65534 host addresses,
while class C supports 254...
But typical organizations
require 300-1000 !!!**

Subnetting

- **Subnetting introduced in 1984**
 - Net + Subnet (=another level)
 - RFC 791
 - Initially only statically configured
- **Classes A, B, C still used for global routing !**
 - Destination Net might be subnetted
 - Smaller routing tables

Routing Table Growth (88-92)

MM/YY	ROUTES ADVERTISED	MM/YY	ROUTES ADVERTISED
Feb-92	4775	Apr-90	1525
Jan-92	4526	Mar-90	1038
Dec-91	4305	Feb-90	997
Nov-91	3751	Jan-90	927
Oct-91	3556	Dec-89	897
Sep-91	3389	Nov-89	837
Aug-91	3258	Oct-89	809
Jul-91	3086	Sep-89	745
Jun-91	2982	Aug-89	650
May-91	2763	Jul-89	603
Apr-91	2622	Jun-89	564
Mar-91	2501	May-89	516
Feb-91	2417	Apr-89	467
Jan-91	2338	Mar-89	410
Dec-90	2190	Feb-89	384
Nov-90	2125	Jan-89	346
Oct-90	2063	Dec-88	334
Sep-90	1988	Nov-88	313
Aug-90	1894	Oct-88	291
Jul-90	1727	Sep-88	244
Jun-90	1639	Aug-88	217
May-90	1580	Jul-88	173

Growth in routing table size, total numbers
Source for the routing table size data is MERIT

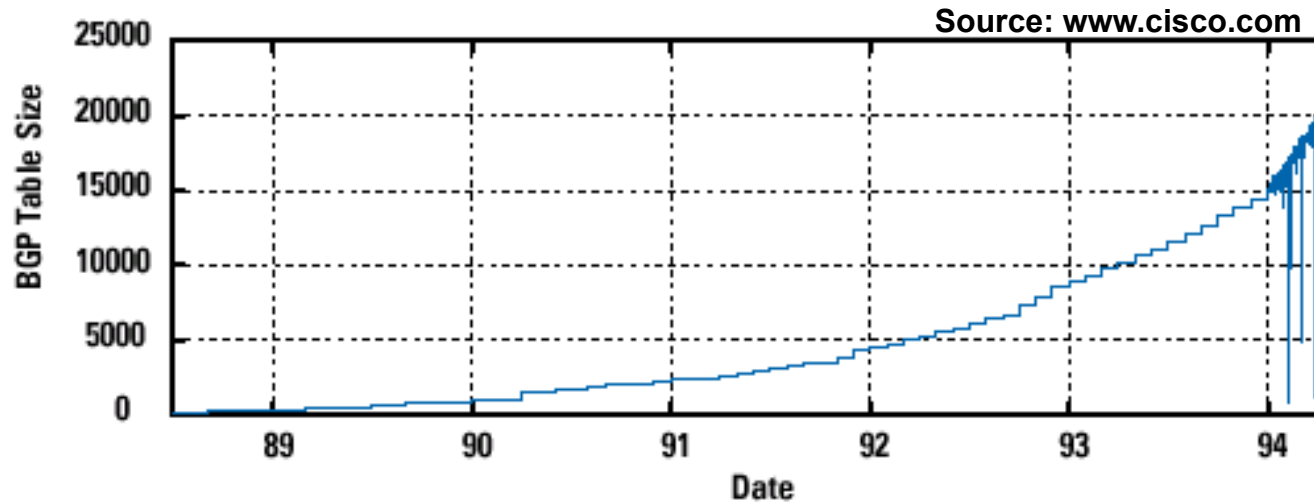
Network Number Statistics, April 1992

	Total	Allocated	Allocated %
Class A	126	48	54%
Class B	16383	7006	43%
Class C	2097151	40724	2%

Only 2% of more than 2 million Class C addresses assigned !!!

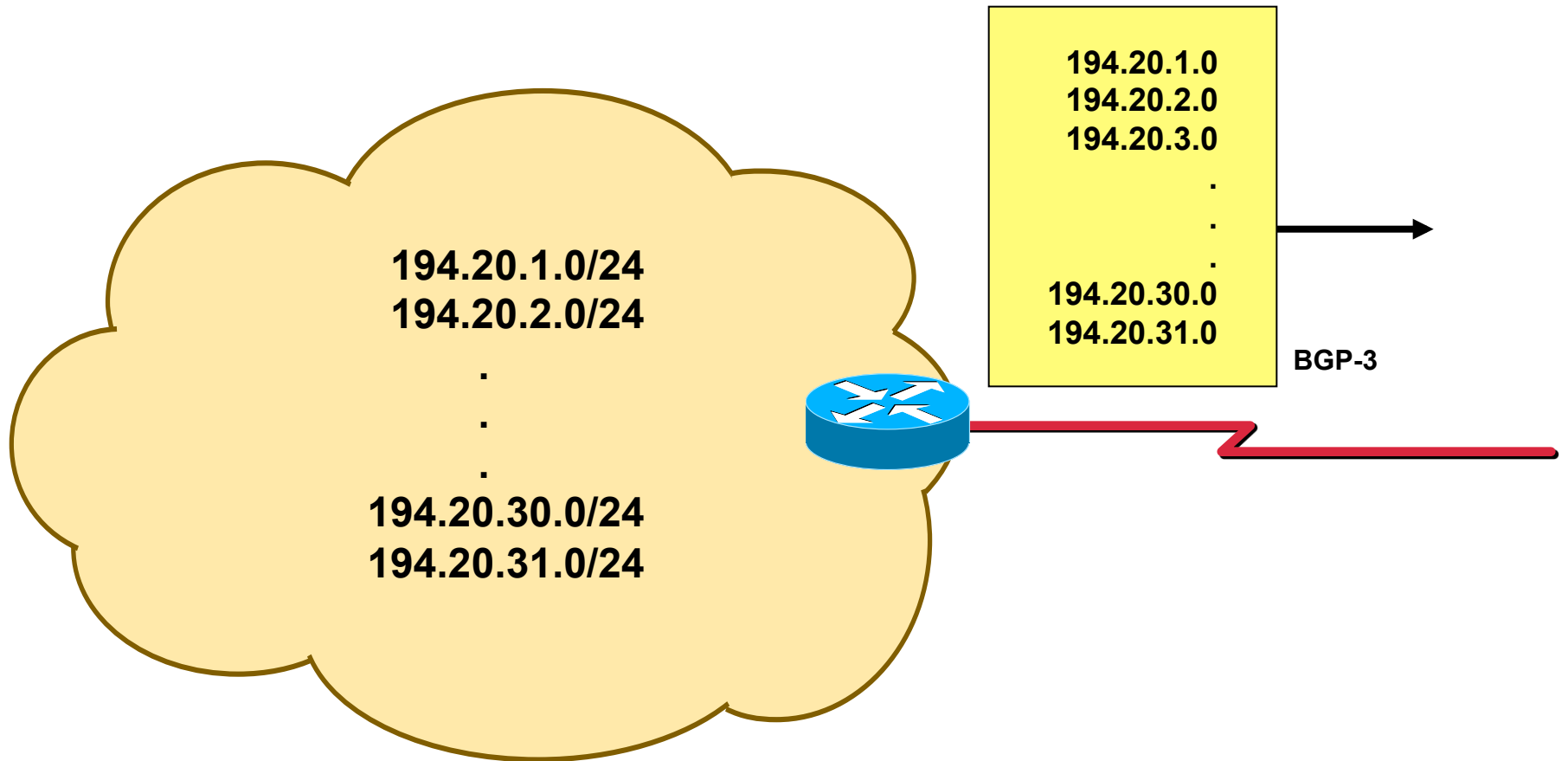
Source: RFC 1335

Supernetting (RFC 1338)

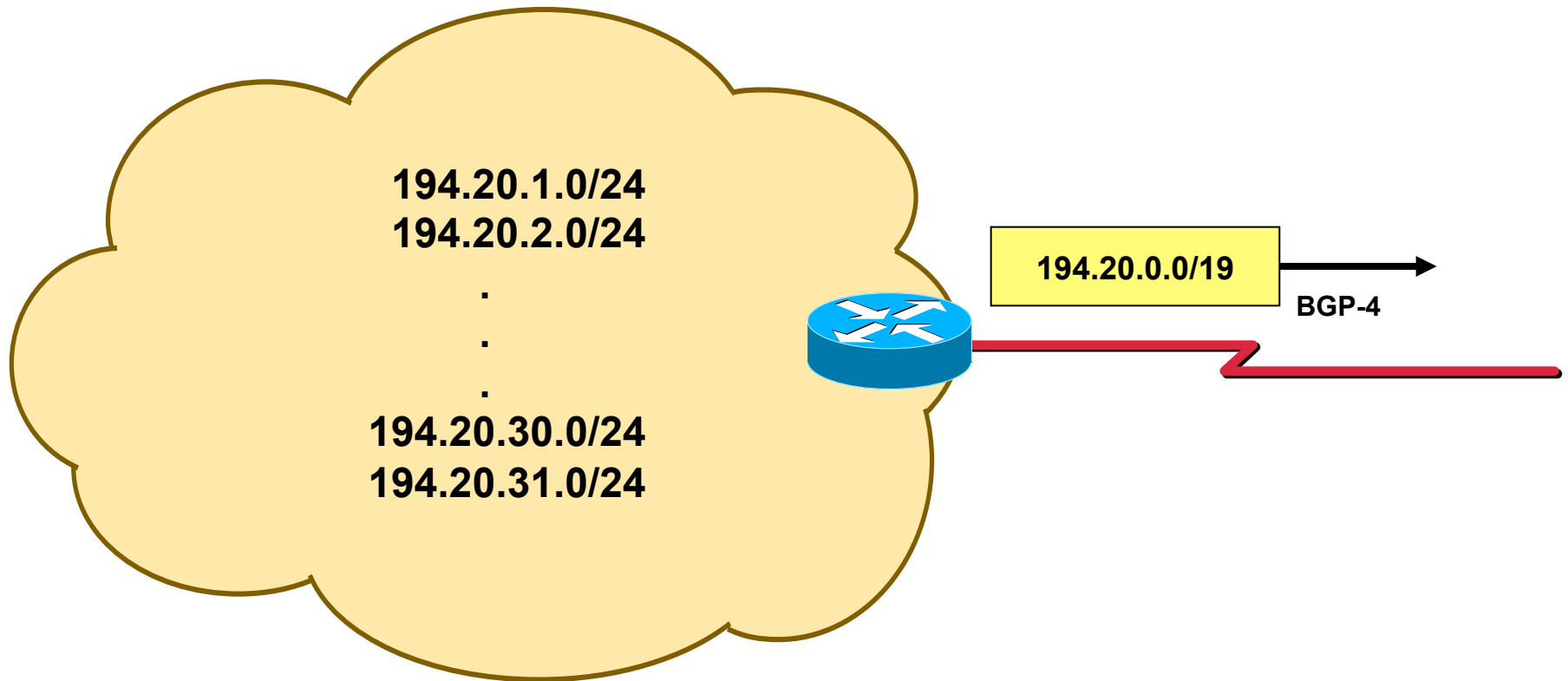


- **In 1992: RFC 1338 stated scaling problem:**
 - Class B exhaustion
 - No class for typical organizations available
 - Unbearable growth of routing table
- **Use subnetting technique also in the Internet !**
 - Do hierarchical IP address assignment !
 - Aggregation = "Supernetting"
(Smaller netmask than natural netmask)

Classful Routing Update



Now Classless and Supernetting



CIDR

- **September 1993, RFC 1519:
Classless Inter-Domain Routing (CIDR)**
- **Requires classless routing protocols**
 - BGP-3 upgraded to BGP-4
 - New BGP-4 capabilities were drawn on a napkin, with all implementers of significant routing protocols present (legend)
 - RFC 1654

Address Management

- **ISPs assign**
*contiguous blocks of
contiguous blocks of
contiguous blocks ...*
of addresses to their customers
- **Aggregation at borders possible !**
- **Tier I providers filter routes with prefix lengths larger than /19**
 - But more and more exceptions today...

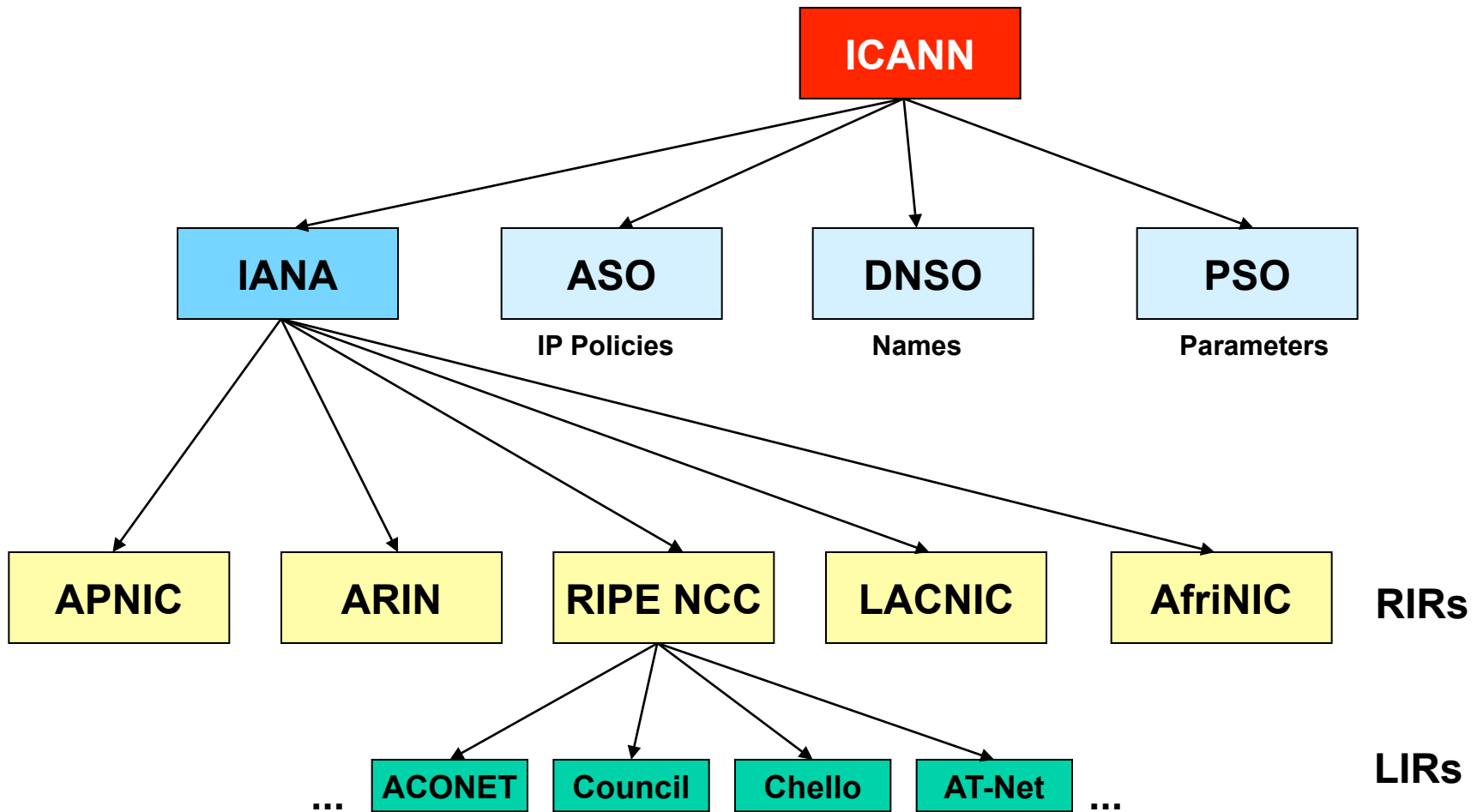
International Address Assignment

- **August 1990, RFC 1174 (by IAB) proposed regionally distributed registry model**
 - Regionally means continental ;-)
- **Regional Internet Registries (RIRs)**
 - RIPE NCC
 - APNIC
 - ARIN

RIRs

- **RIPE NCC (1992)**
 - Réseaux IP Européens (RIPE) founded the Network Coordination Centre (NCC)
- **APNIC (1993)**
 - Asia Pacific Information Centre
- **ARIN (1997)**
 - American Registry for Internet Numbers
- **AfriNIC**
 - Africa
- **LACNIC**
 - Latin America and Caribbean

ICANN, RIRs, and LIRs



CIDR Concepts Summary

- **Coordinated address allocation**
- **Classless routing**
- **Supernetting**

RFC 1366 Address Blocks

- **192.0.0.0 - 193.255.255.255 ...** **Multiregional**
- **194.0.0.0 - 195.255.255.255 ...** **Europe**
- **198.0.0.0 - 199.255.255.255 ...** **North America**
- **200.0.0.0 - 201.255.255.255 ...** **Central/South America**
- **202.0.0.0 - 203.255.255.255 ...** **Pacific Rim**

Class A Assignment

- **IANA responsibility**

- RFC 1366 states: *"There are only approximately 77 Class A network numbers which are unassigned, and these 77 network numbers represent about 30% of the total network number space."*

- **64.0.0.0 – 127.0.0.0 were reserved for the end of (IPv4) days ?**

- Recent assignments
(check IANA website)

Class B Assignment

- **IANA and RIRs requirements**
 - Subnetting plan which documents more than 32 subnets within its organizational network
 - More than 4096 hosts
- **RFC 1366 recommends to use multiple Class Cs wherever possible**

Class C Assignment

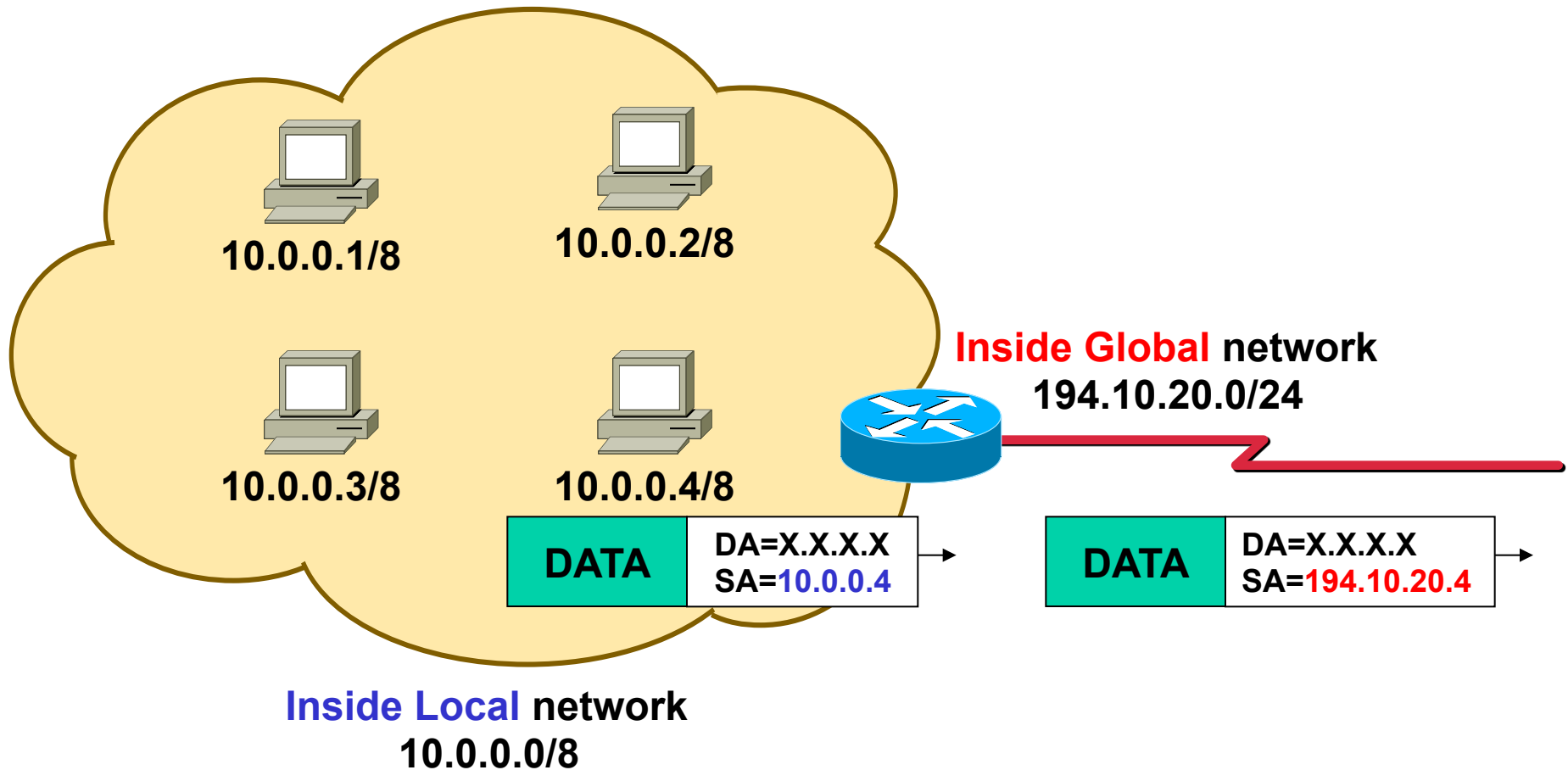
- If an organization requires more than a single Class C, it will be assigned a bit-wise contiguous block from the Class C space
- Up to 16 contiguous Class C networks per subscriber (= one prefix, 12 bit length)

Organization	Assignment
1) requires fewer than 256 addresses	1 class C network
2) requires fewer than 512 addresses	2 contiguous class C networks
3) requires fewer than 1024 addresses	4 contiguous class C networks
4) requires fewer than 2048 addresses	8 contiguous class C networks
5) requires fewer than 4096 addresses	16 contiguous class C networks

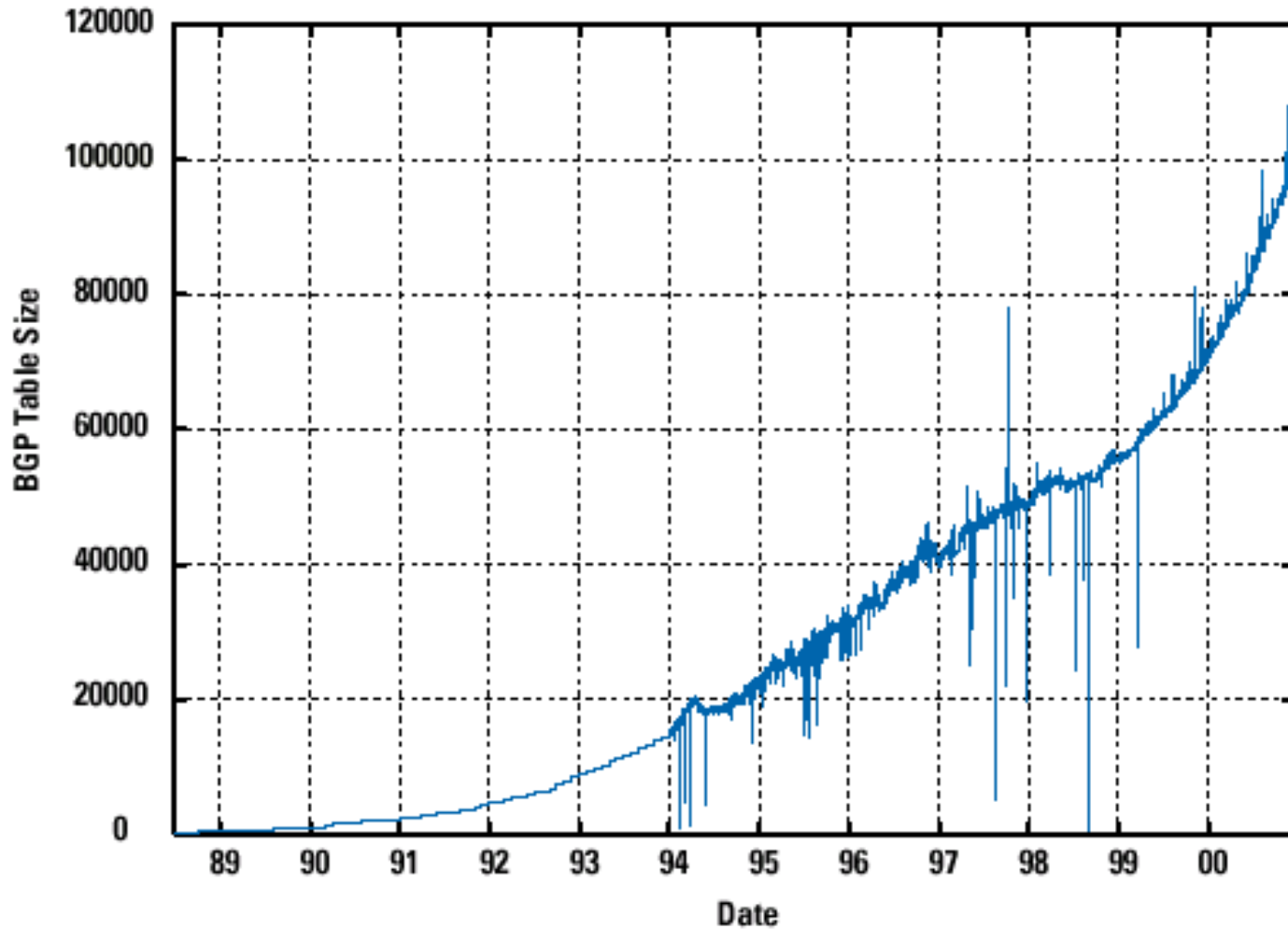
RFC 1918 – Private Addresses

- **In order to prevent address space depletion, RFC 1918 defined three private address blocks**
 - 10.0.0.0 - 10.255.255.255 (prefix: 10/8)
 - 172.16.0.0 - 172.31.255.255 (prefix: 172.16/12)
 - 192.168.0.0 - 192.168.255.255 (prefix: 192.168/16)
- **Connectivity to global space via Network Address Translation (NAT)**

NAT Example



But...



Source: www.cisco.com