Routing Theory

Vladimír Veselý
veselyv@fit.vutbr.cz

Motivation

- Operating System
- Router
- Switch

Agenda

- **Addressing**
- **Decision Tables**
- **Routing Protocol Basics**
- **Evolution of**
  - distance-vector
  - link-state
  - path-vector
  - multicast
  - switching
- **Summary**

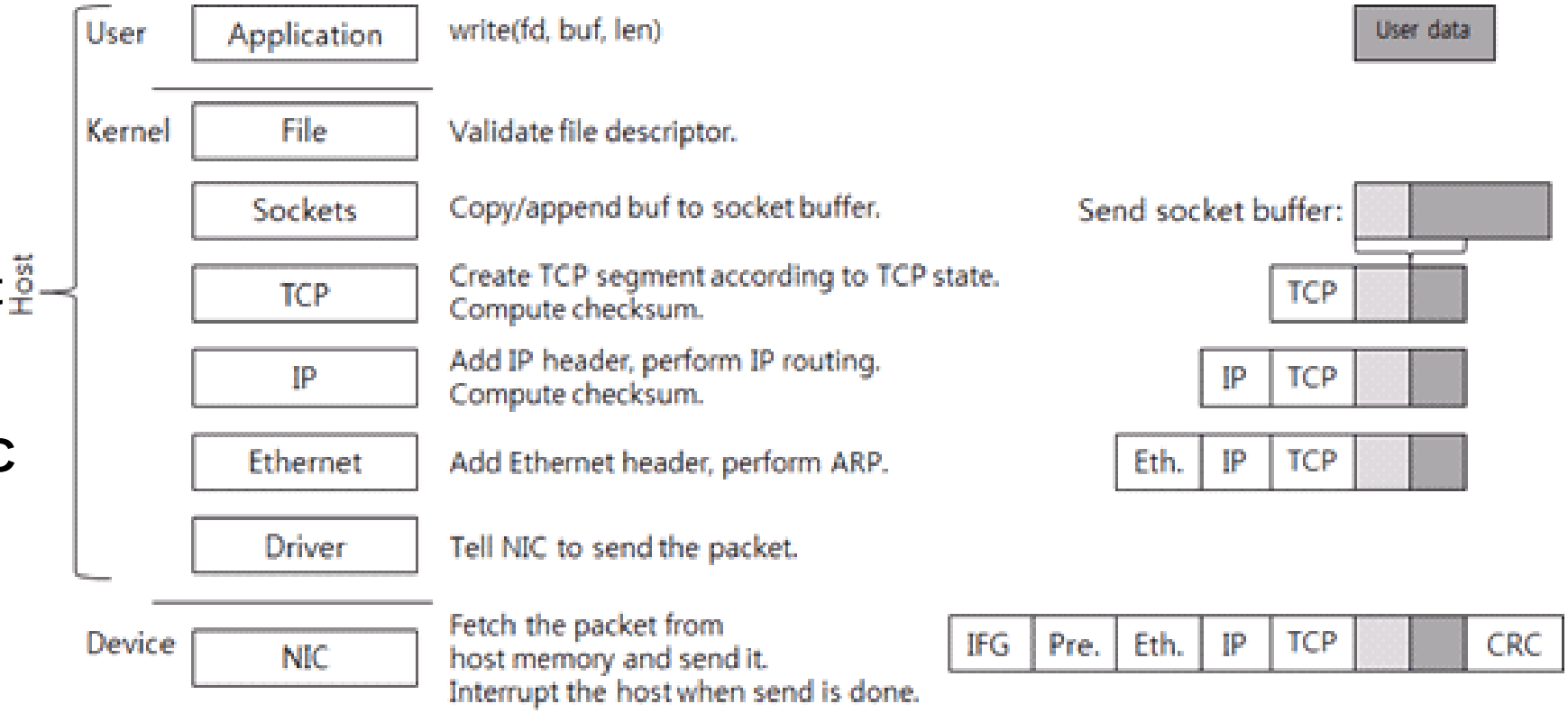
Addressing

“Now you people have names. That's because you don't know who you are. We know who we are, so we don't need names.”

Neil Gaiman, Coraline

Identifiers

- **Port**
- **IP**
- **MAC**



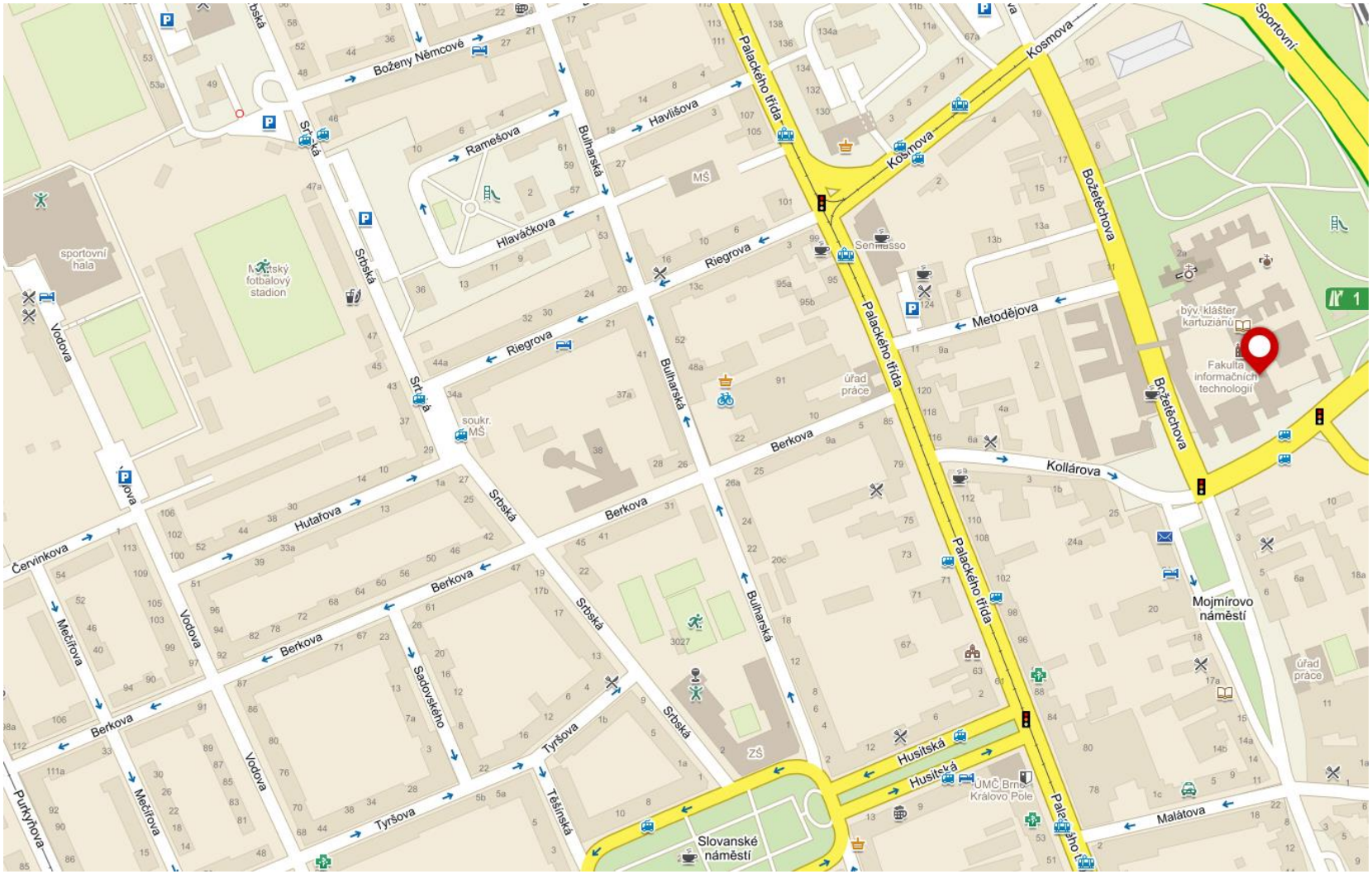
http://blog.cubrid.org:8080/files/attach/images/220547/523/594/operation_process_by_each_layer_of_tcp_ip.png

Type of Communication

- **Unicast**
  - Address identifies a single receiver
  - L3 based on destination IP
  - L2 switching based on destination MAC
- **Multicast**
  - Address identifies a group of receivers
  - L3 routing based on source IP
  - L2 switching based on destination MAC
- **Broadcast**
  - Address is a macro for all recipients
  - Floods L3, L2 traffic
- **Anycast**
  - No distinction from unicast address

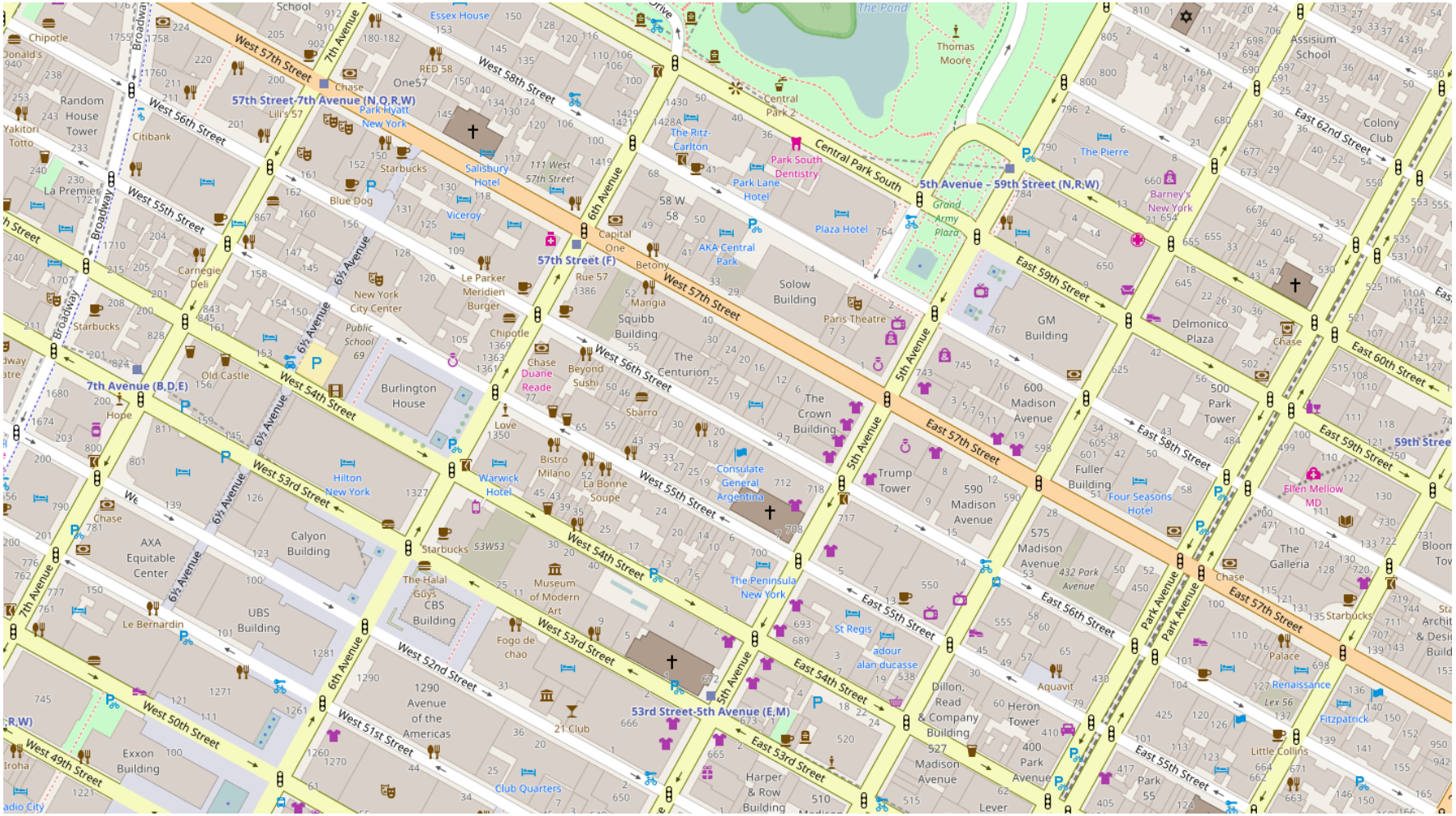
Flat Addresses

- Božetěchova 2, Brno, 612 66



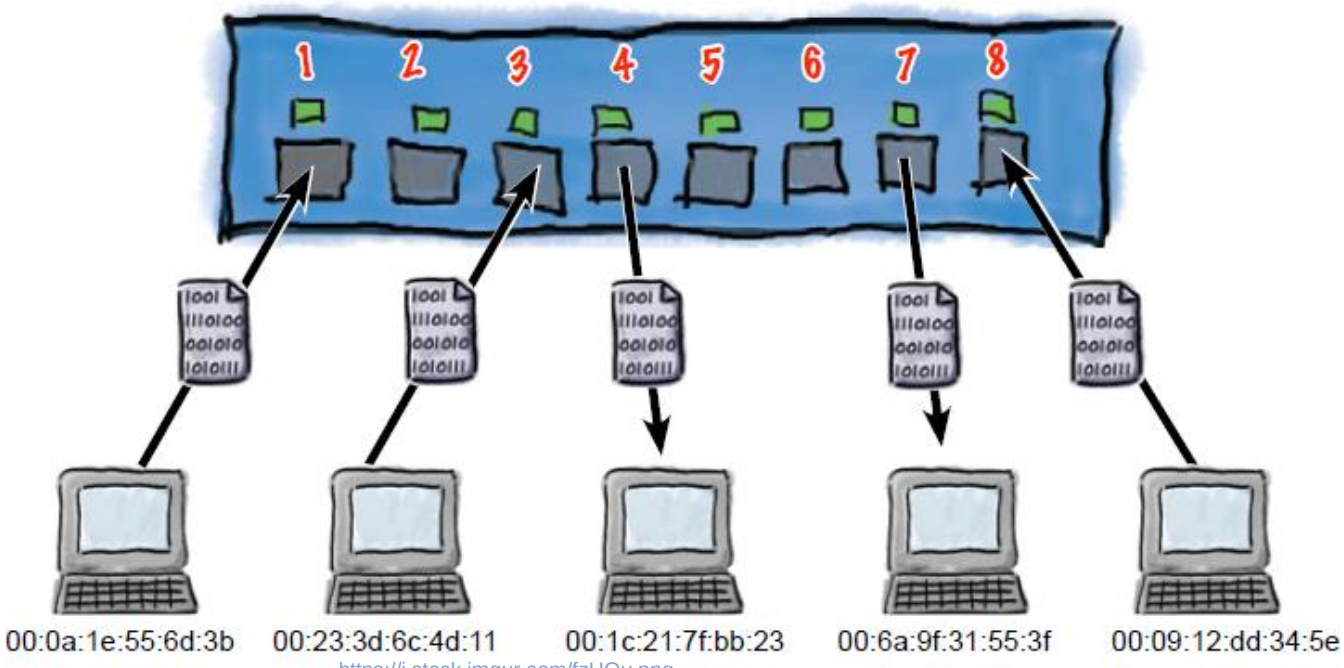
Hierarchical Addresses

- 5th Avenue – 59th Street Crossing, New York



MAC and Port

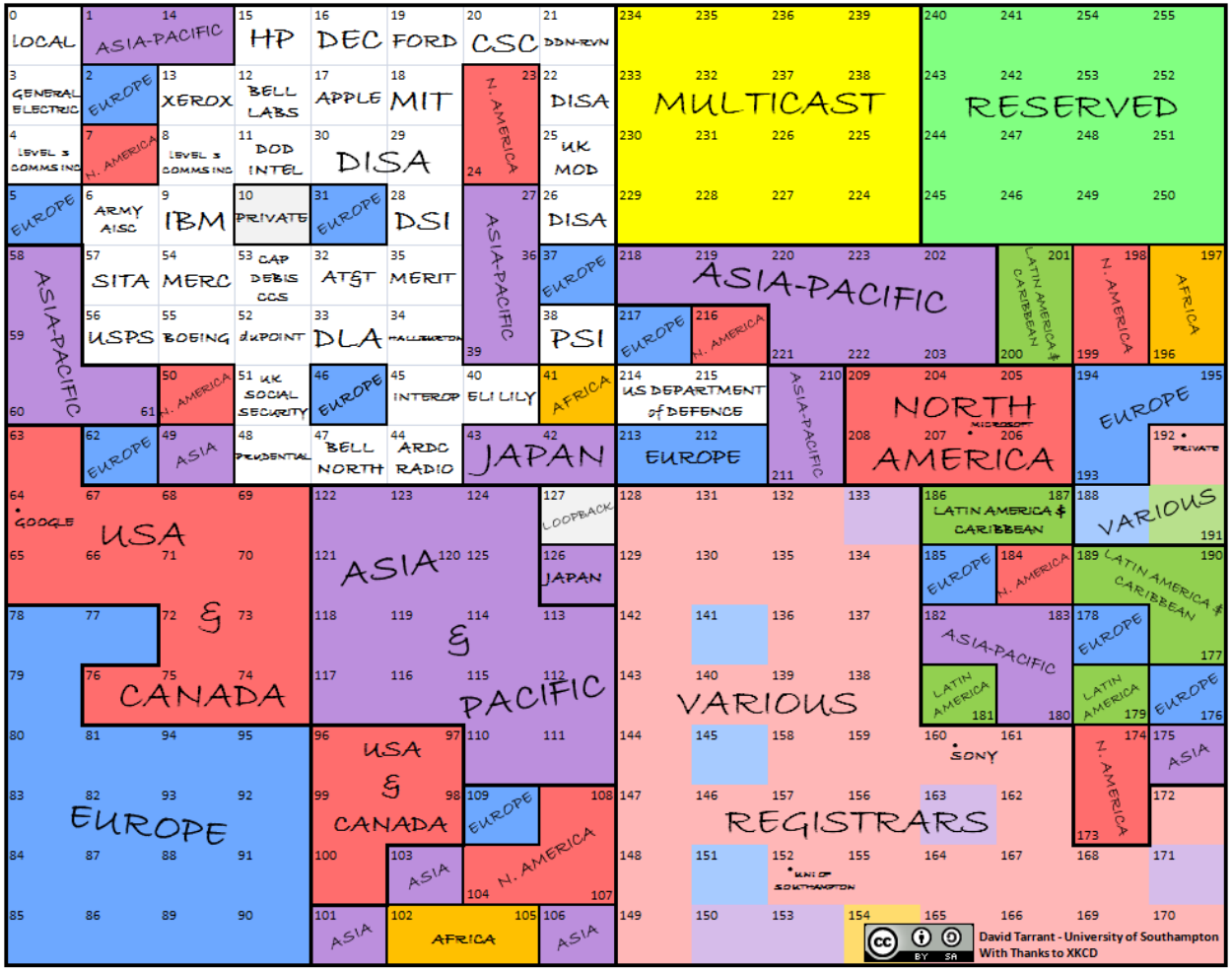
- Port
  - 16 bits
  - Catalog value
    - https://www.iana.org/assignments/service-names-port-numbers/
  - *Nearness for applications?*
- MAC
  - flat
  - 48 bits



https://i.stack.imgur.com/fzHQv.png

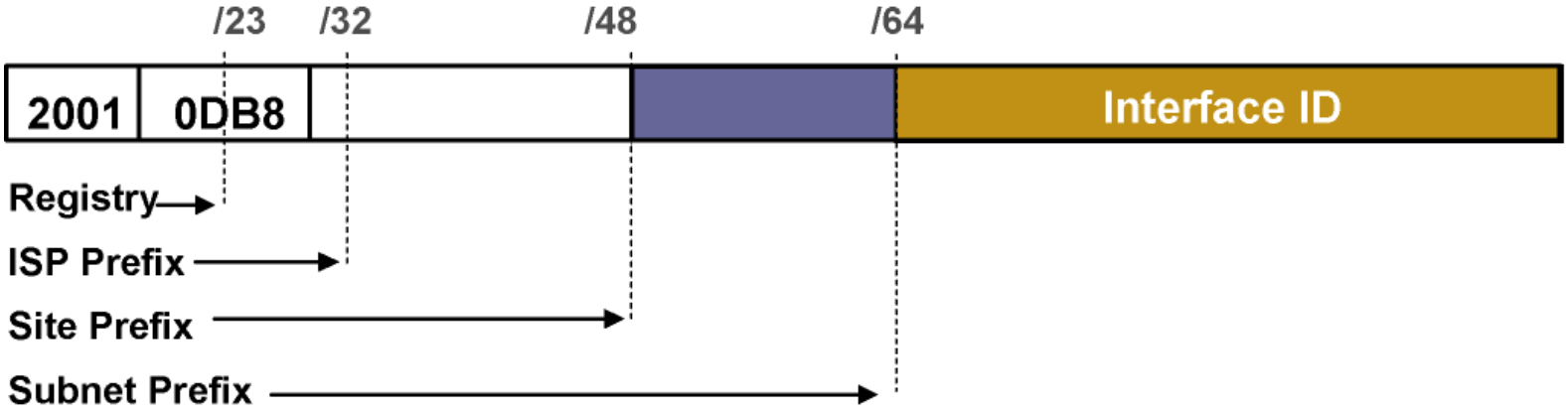
IPv4 Address

- pseudohierarchical (NetID + HostID)
- 32 bits



IPv6 Address

- Pseudohierarchical (Prefix + InterfaceId)
- 128 bit



Unicast Decision Tables

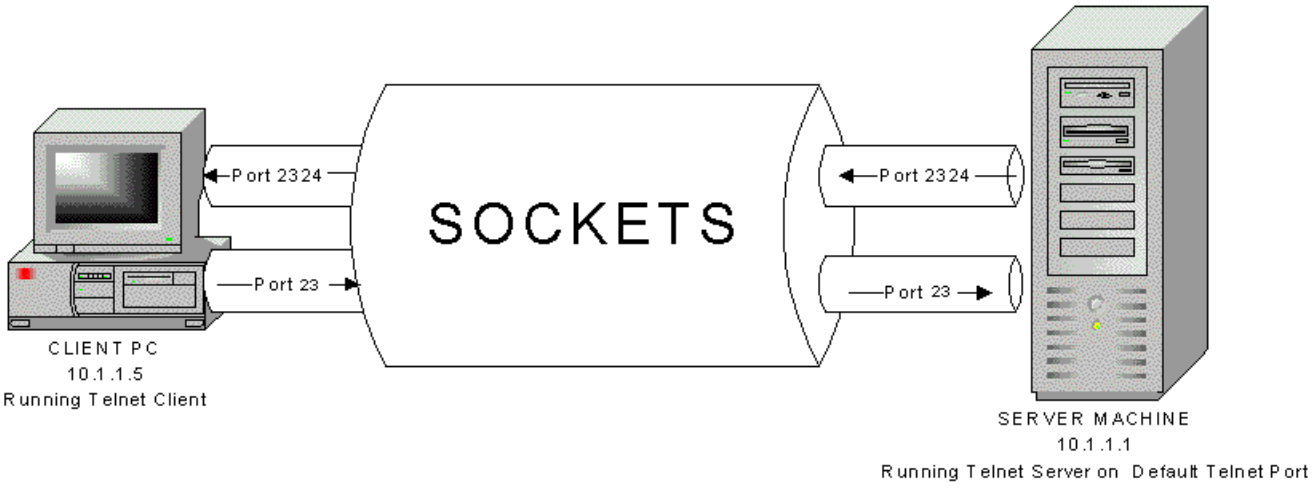
“It is hard to imagine a more stupid or more dangerous way of making decisions than by putting those decisions in the hands of people who pay no price for being wrong.”

Thomas Sowell, Wake Up Parents

L4: Open Sockets

- Windows,
  Linux: `netstat`

```
C:\Users\Mordeth>netstat

Active Connections

  Proto  Local Address          Foreign Address        State
  TCP    10.20.40.10:49681      db5sch101110829:https  ESTABLISHED
  TCP    10.20.40.10:49706      db5sch101110125:https  ESTABLISHED
  TCP    10.20.40.10:49718      13.69.158.96:https     ESTABLISHED
  TCP    10.20.40.10:49725      40.77.226.192:https    ESTABLISHED
  TCP    10.20.40.10:49726      91.190.219.42:12350    ESTABLISHED
  TCP    10.20.40.10:49741      cst2-115-212:imaps     ESTABLISHED
  TCP    10.20.40.10:49749      tereza:5222            ESTABLISHED
  TCP    10.20.40.10:49788      bos-m016c:https        ESTABLISHED
  TCP    10.20.40.10:52175      edge-mqtt-shv-01-frt3:https  ESTABLISHED
  TCP    10.20.40.10:53412      lb:5222                ESTABLISHED
  TCP    10.20.40.10:54004      ip-94-112-78-55:https  ESTABLISHED
  TCP    10.20.40.10:54008      ip-94-112-78-55:https  ESTABLISHED
  TCP    10.20.40.10:54017      104.16.120.49:https    CLOSE_WAIT
  TCP    10.20.40.10:54087      ip-94-112-78-55:https  ESTABLISHED
  TCP    10.20.40.10:54093      ip-94-112-78-55:https  ESTABLISHED
                                                       55:https  ESTABLISHED
                                                       55:https  ESTABLISHED
                                                       55:https  ESTABLISHED
                                                       55:https  ESTABLISHED
                                                       55:https  ESTABLISHED
                                                       55:https  ESTABLISHED
                                                       55:https  ESTABLISHED
                                                       55:https  ESTABLISHED
                                                       55:https  ESTABLISHED
                                                       55:https  ESTABLISHED
```



http://www.ortizonline.com/images/sockets.gif

L4: Socket Multiplexing

- Linux: `lsof`

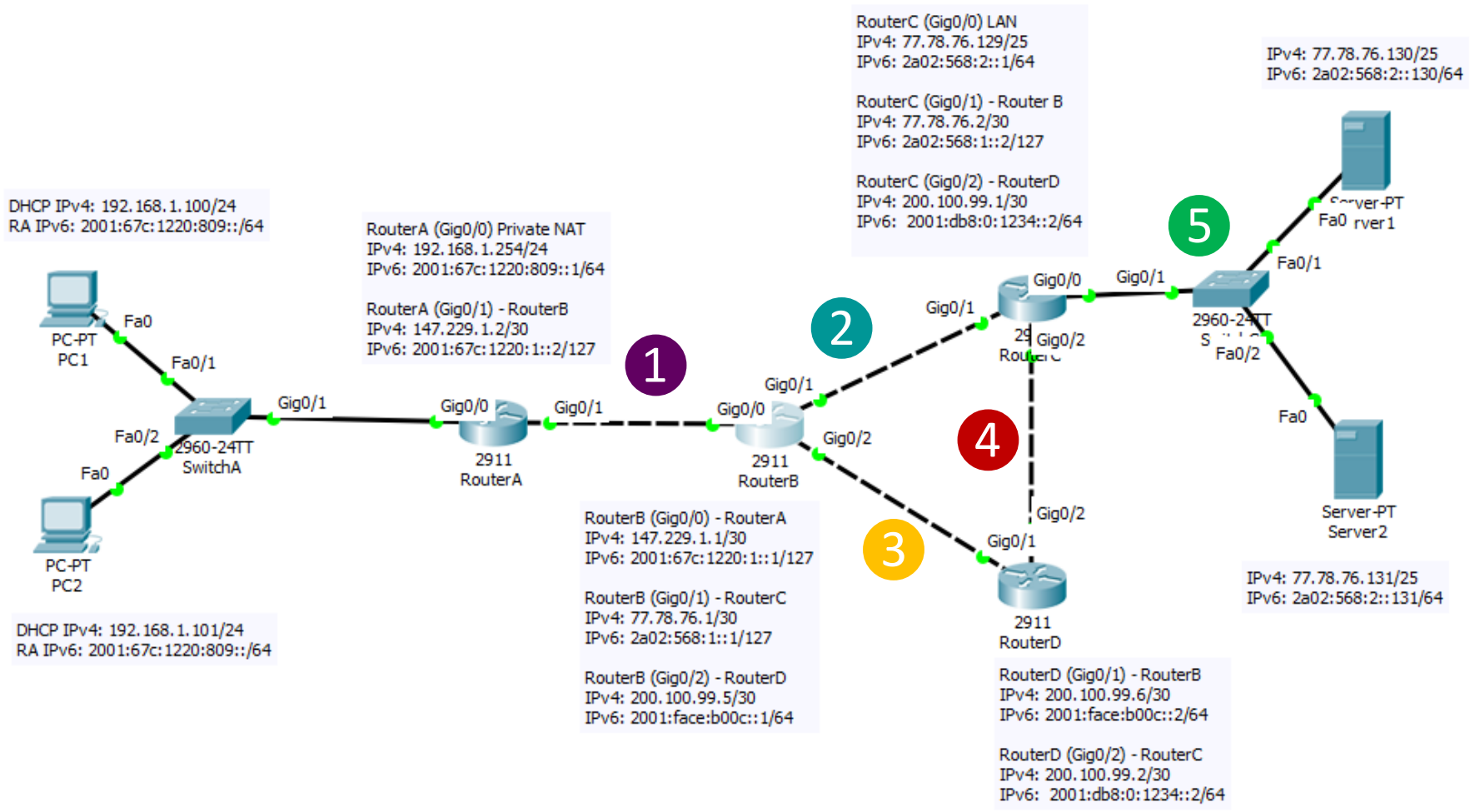
root@ciscolab: ~

```
root@ciscolab:~# lsof -i
COMMAND     PID     USER   FD   TYPE     DEVICE SIZE/OFF NODE NAME
apache2    1453 www-data    4u  IPv6 2179069994      0t0  TCP *:http (LISTEN)
apache2    1569 www-data    4u  IPv6 2179069994      0t0  TCP *:http (LISTEN)
qemu-syst  2629     root   13u  IPv4 2758114985      0t0  TCP *:46474 (LISTEN)
apache2   11845 www-data    4u  IPv6 2179069994      0t0  TCP *:http (LISTEN)
qemu-syst 13214     root   13u  IPv4 2125449868      0t0  TCP *:40457 (LISTEN)
qemu-syst 13391     root   13u  IPv4 2125512308      0t0  TCP *:40456 (LISTEN)
qemu-syst 13756     root   13u  IPv4 2125246965      0t0  TCP *:40454 (LISTEN)
apache2   15126 www-data    4u  IPv6 2179069994      0t0  TCP *:http (LISTEN)
ovs-contr 19541     root    4u  IPv4 2178996945      0t0  TCP *:6633 (LISTEN)
apache2   20175     root    4u  IPv6 2179069994      0t0  TCP *:http (LISTEN)
sshd      20311     root    3u  IPv4 2179068991      0t0  TCP *:666 (LISTEN)
sshd      20311     root    4u  IPv6 2179068993      0t0  TCP *:666 (LISTEN)
qemu-syst 27275     root   13u  IPv4 2761684172      0t0  TCP *:47873 (LISTEN)
apache2   27640 www-data    4u  IPv6 2179069994      0t0  TCP *:http (LISTEN)
apache2   28830 www-data    4u  IPv6 2179069994      0t0  TCP *:http (LISTEN)
apache2   28831 www-data    4u  IPv6 2179069994      0t0  TCP *:http (LISTEN)
apache2   33282 www-data    4u  IPv6 2179069994      0t0  TCP *:http (LISTEN)
apache2   33283 www-data    4u  IPv6 2179069994      0t0  TCP *:http (LISTEN)
qemu-syst 37788     root   13u  IPv4 2757201287      0t0  TCP *:45706 (LISTEN)
qemu-syst 41712     root   13u  IPv4 2757981793      0t0  TCP *:48010 (LISTEN)
sh        42665    unl81    7u  IPv6 1465027947      0t0  TCP *:43139 (LISTEN)
L2-advent 42671    unl81    7u  IPv6 1465027947      0t0  TCP *:43139 (LISTEN)
dhclient  50827     root    7u  IPv4   46382458      0t0  UDP *:bootpc
dhclient  50827     root   20u  IPv4   46378082      0t0  UDP *:9579
dhclient  50827     root   21u  IPv6   46378083      0t0  UDP *:44523
qemu-syst 54293     root   13u  IPv4 2759810296      0t0  TCP *:46218 (LISTEN)
sshd      54405     root    3u  IPv4 2779665963      0t0  TCP sec6net-mv17.fit.vutbr.cz    ->nat24.bozka.vutbr.net
dhclient  56928     root    7u  IPv4   46441602      0t0  UDP *:bootpc
dhclient  56928     root   20u  IPv4   46395053      0t0  UDP *:28491
dhclient  56928     root   21u  IPv6   46395054      0t0  UDP *:37330
qemu-syst 57655     root   13u  IPv4 2758101001      0t0  TCP *:45834 (LISTEN)
apache2   60214 www-data    4u  IPv6 2179069994      0t0  TCP *:http (LISTEN)
root@ciscolab:~# [2~
```

L3: IP Routing Table

- Meeting place for routes from different routing protocols
- The best route is determined by **metric**
  - Different equations how to reach this number
  - Lower means better
  - Optionally load-balancing between the best routes with the same metric
- The best source of routing information is determined by its **trustworthiness**
  - Administrative distance
  - Lower means better

L3: Unicast Topology



L3: Unicast IPv4 RT

```
RouterB#show ip route
Codes: L - local, C - connected, S - static, R - RIP, M - mobile, B - BGP
       D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2, E - EGP
       i - IS-IS, L1 - IS-IS level-1, L2 - IS-IS level-2, ia - IS-IS inter area
       * - candidate default, U - per-user static route, o - ODR
       P - periodic downloaded static route

Gateway of last resort is not set

     77.0.0.0/8 is variably subnetted, 3 subnets, 3 masks
C       77.78.76.0/30 is directly connected, GigabitEthernet0/1
L       77.78.76.1/32 is directly connected, GigabitEthernet0/1
D       77.78.76.128/25 [90/3072] via 77.78.76.2, 01:29:26, GigabitEthernet0/1
     147.229.0.0/16 is variably subnetted, 2 subnets, 2 masks
C       147.229.1.0/30 is directly connected, GigabitEthernet0/0
L       147.229.1.1/32 is directly connected, GigabitEthernet0/0
     200.100.99.0/24 is variably subnetted, 3 subnets, 2 masks
D       200.100.99.0/30 [90/3072] via 77.78.76.2, 00:53:40, GigabitEthernet0/1
                        [90/3072] via 200.100.99.6, 00:53:20, GigabitEthernet0/2
C       200.100.99.4/30 is directly connected, GigabitEthernet0/2
L       200.100.99.5/32 is directly connected, GigabitEthernet0/2
RouterB#
```

2 5 1 4 3

L3: Unicast IPv6 RT

```
RouterB#show ipv6 route
IPv6 Routing Table - 11 entries
Codes: C - Connected, L - Local, S - Static, R - RIP, B - BGP
       U - Per-user Static route, M - MIPv6
       I1 - ISIS L1, I2 - ISIS L2, IA - ISIS interarea, IS - ISIS summary
       O - OSPF intra, OI - OSPF inter, OE1 - OSPF ext 1, OE2 - OSPF ext 2
       ON1 - OSPF NSSA ext 1, ON2 - OSPF NSSA ext 2
       D - EIGRP, EX - EIGRP external
C   2001:67C:1220:1::/127 [0/0]
     via GigabitEthernet0/0, directly connected
L   2001:67C:1220:1::1/128 [0/0]
     via GigabitEthernet0/0, receive
O   2001:DB8:0:1234::/64 [110/2]
     via FE80::209:7CFF:FEAD:7402, GigabitEthernet0/1
     via FE80::201:64FF:FE75:B802, GigabitEthernet0/2
C   2001:FACE:B00C::/64 [0/0]
     via GigabitEthernet0/2, directly connected
L   2001:FACE:B00C::1/128 [0/0]
     via GigabitEthernet0/2, receive
C   2A02:568:1::/127 [0/0]
     via GigabitEthernet0/1, directly connected
L   2A02:568:1::1/128 [0/0]
     via GigabitEthernet0/1, receive
O   2A02:568:2::/64 [110/2]
     via FE80::209:7CFF:FEAD:7402, GigabitEthernet0/1
L   FF00::/8 [0/0]
     via Null0, receive
RouterB#
```

1
4
3
2
5

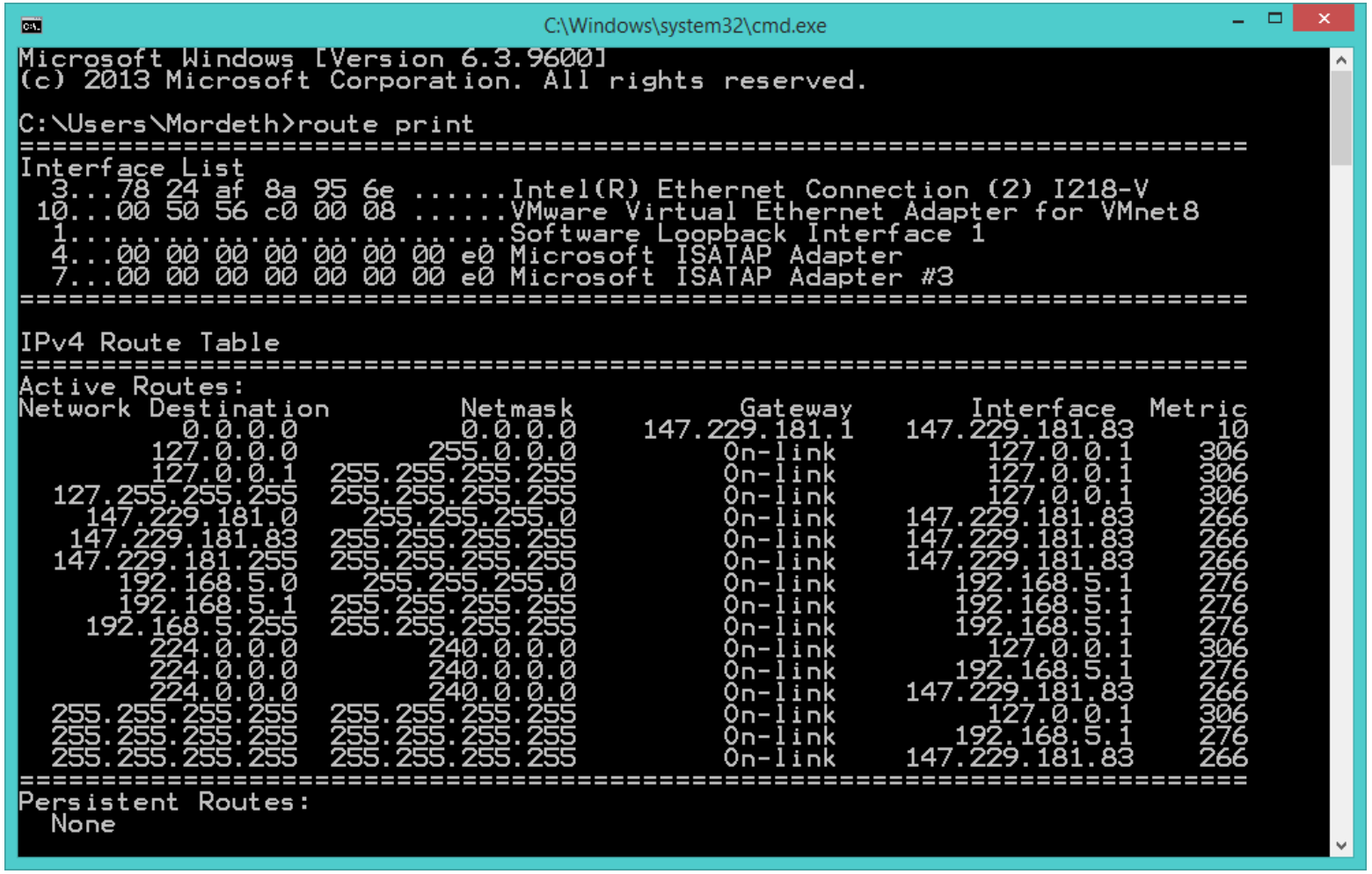
Cisco's Administrative Distance

Route Source	Default Distance Values
Connected interface	0
Static route	1
Enhanced Interior Gateway Routing Protocol (EIGRP) summary route	5
External Border Gateway Protocol (BGP)	20
Internal EIGRP	90
IGRP	100
OSPF	110
Intermediate System-to-Intermediate System (IS-IS)	115
Routing Information Protocol (RIP)	120
Exterior Gateway Protocol (EGP)	140
On Demand Routing (ODR)	160
External EIGRP	170
Internal BGP	200
Unknown*	255

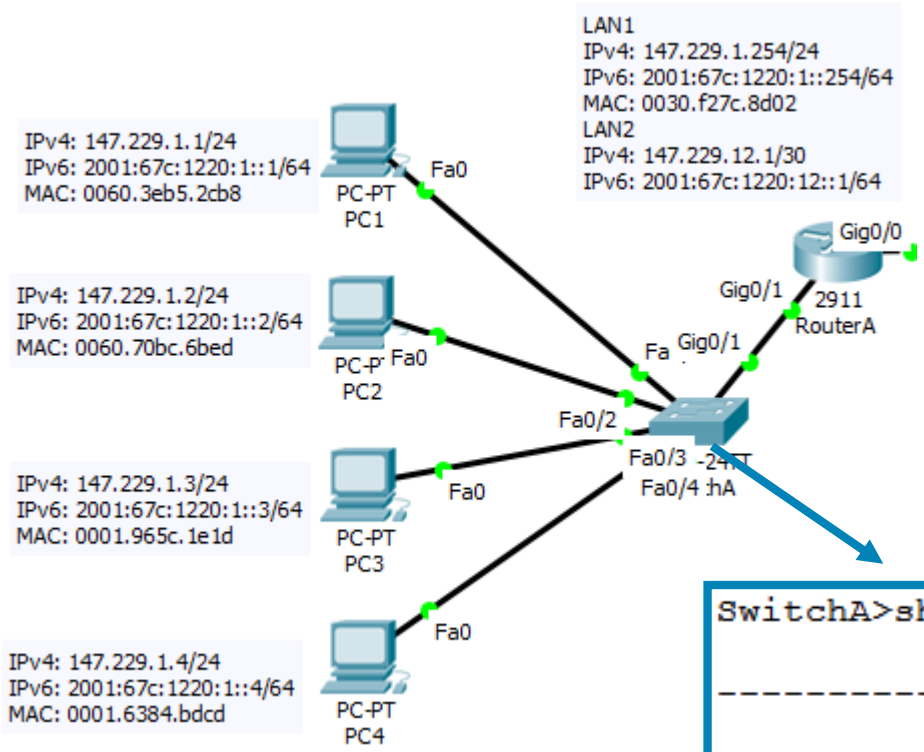
Route Source	Cisco Value	Juniper Value
Connected interface	0	0
Static route	1	5
External Border Gateway Protocol (BGP)	20	170
OSPF	110	150
Intermediate System-to-Intermediate System (IS-IS)	115	160/165
Routing Information Protocol (RIP)	120	100
Internal BGP	200	170

L3: Routing Table Examples

- Windows: `route print`



L2: CAM Table




```
SwitchA>show mac address-table
          Mac Address Table
-------------------------------------------

Vlan    Mac Address       Type        Ports
----    -----------       --------    -----

   1    0001.6384.bdcd    DYNAMIC     Fa0/4
   1    0001.965c.1e1d    DYNAMIC     Fa0/3
   1    0030.f27c.8d02    DYNAMIC     Gig0/1
   1    0060.3eb5.2cb8    DYNAMIC     Fa0/1
   1    0060.70bc.6bed    DYNAMIC     Fa0/2
```

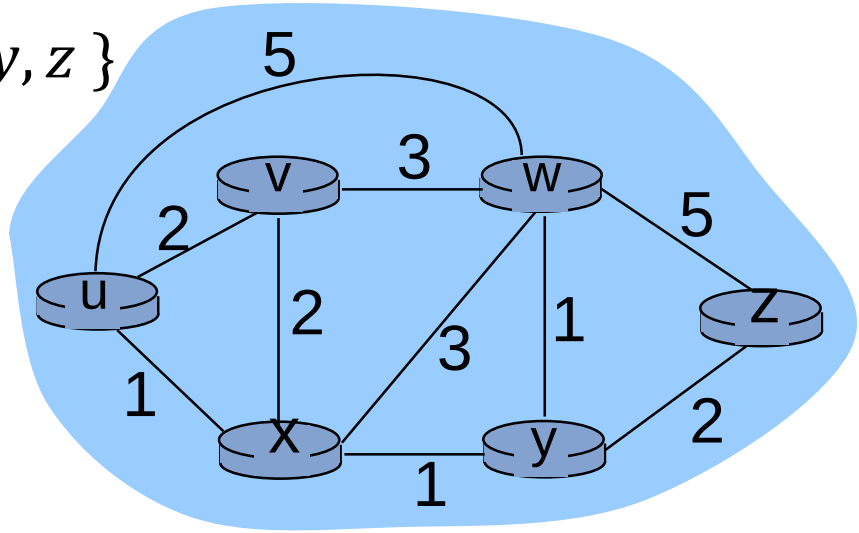
Routing Protocol Basics

“Begin at the beginning,” the King said, gravely, “and go on till you come to the end; then stop”

Lewis Carroll, Alice in the Wonderland

Graph Abstraction

- **graph** := $G = (N, E)$
- N := **set of routers** = $\{ u, v, w, x, y, z \}$
- E := **set of links** =
  $\{(u, v), (u, x), (v, x), (v, w), (x, w),$
  $(x, y), (w, y), (w, z), (y, z)\}$
- **cost of path** $(x_1, x_2, x_3, \ldots, x_p)$ =
  $c(x_1, x_2) + c(x_2, x_3) +$
  $\ldots + c(x_{p-1}, xp)$
- $c(x, x')$:= **cost of link** (x, x'),
  e.g., $c(w, z) = 5$



key question: what is the least-cost path between u and z?
routing algorithm: algorithm that finds that least cost path

Types

- *Global or decentralized routing information?*
- **Distance-vector**
  - Single metric
  - *“Routing by rumor”*
  - router knows physically-connected neighbors, link costs to neighbors
  - iterative process of computation, exchange of info with neighbors
- **Link-state**
  - Single metric
  - *Everyone knows how network globaly looks like*
  - all routers have complete topology, link cost info
- **Path-vector**
  - Multi-metric
  - Network viewed as a set of autonomous systems

Bellman-Ford Equation

let

$d_x(y)$:= cost of least-cost path from x to y

then

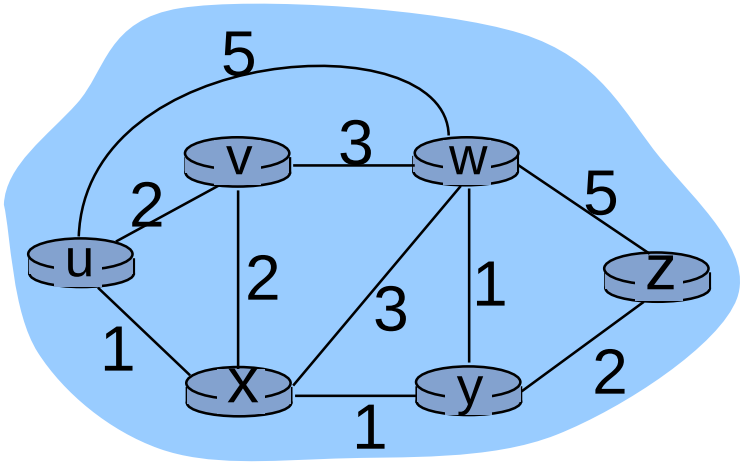
$$d_x(y) = \min\{c(x,v) + d_v(y)\}$$

cost from neighbor v to y

cost to neighbor v

min taken over all neighbors v of x

Bellman-Ford Example



clearly:

$d_v(z) = 5,$
$d_x(z) = 3,$
$d_w(z) = 3$

B-F equation says:

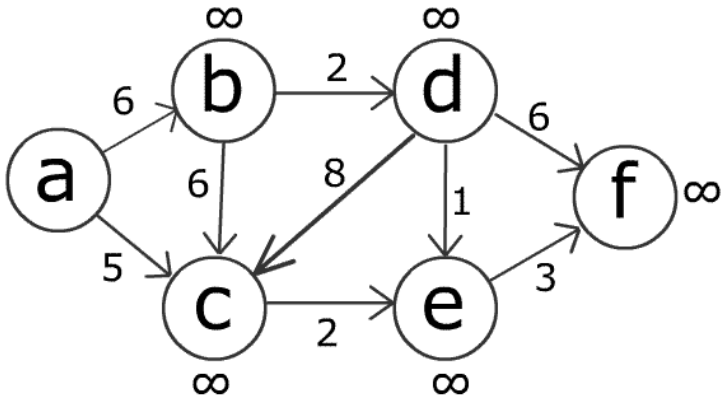
$$d_u(z)$$
$$= \min\{c(u,v) + dv(z),$$
$$c(u,x) + dx_{(z)},$$
$$c(u,w) + d_w(z)\}$$
$$= \min\{2+5,$$
$$1+3,$$
$$5+3\} = 4$$

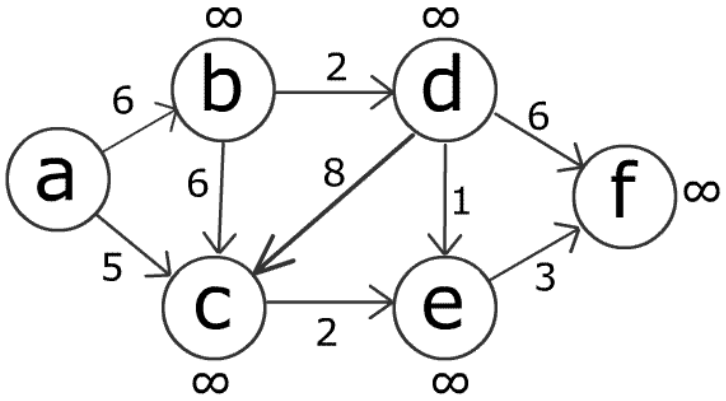
node achieving minimum is:

- next hop in shortest path
- used in forwarding table

Bellman-Ford Algorithm

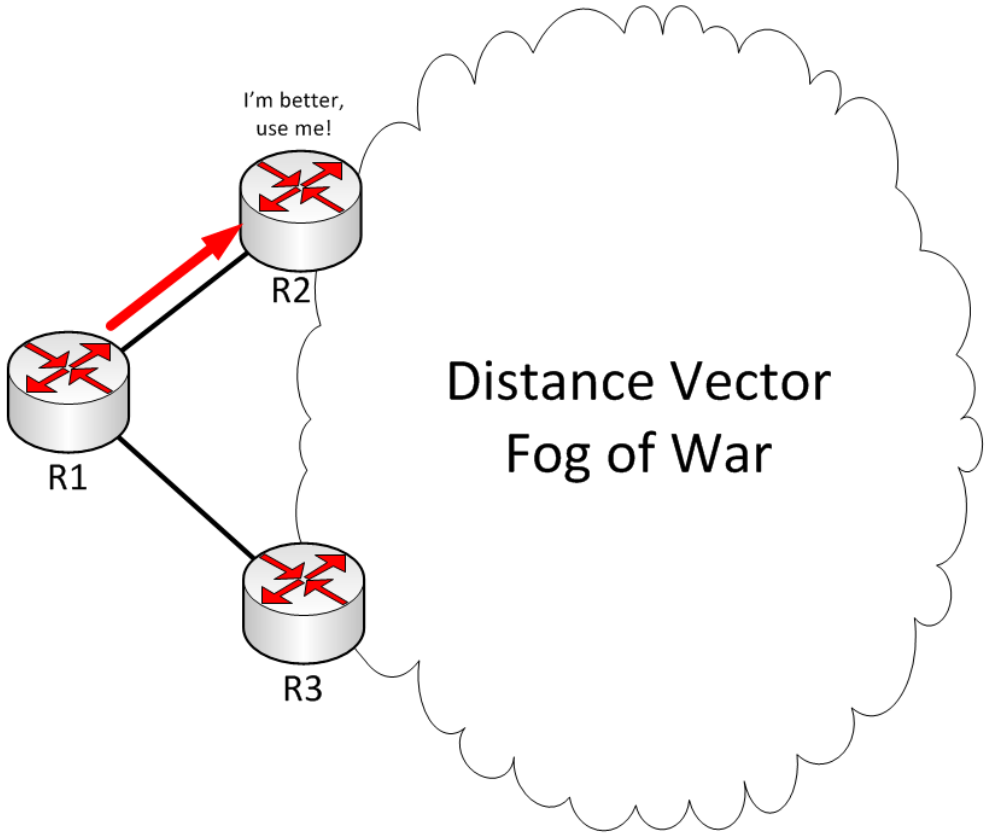
```
BELLMAN-FORD(G,w,s)
1.   INITIALIZE-SINGLE-SOURCE(G,s)
2.   for i = 1 to |G.V|-1
3.      for each edge (u,v) ∈ G.E
4.         RELAX(u,v,w)
5.   for each edge (u,v) ∈ G.E
6.      if v.d > u.d + w(u,v)
7.         return FALSE
8.   return TRUE

INITIALIZE-SINGLE-SOURCE(G,s)
1.   for each vertex v ∈ G.V
2.      v.d = ∞
3.      v.pi = NIL
4.   s.d = 0

RELAX(u,v,w)
1.   if v.d > u.d + w(u,v)
2.      v.d = u.d + w(u,v)
3.      v.pi = u
```





Distance-Vector Protocol

- Each router knows little about network topology
  - Only best next-hops are chosen by each router for each destination network
  - Best end-to-end paths result from composition of all next-hop choices
- Does not require uniform policies at all routers
  - Routing by rumor
- Examples: RIP, EIGRP, Babel



Distance-Vector Notation

- $D_x(y)$:= estimate of least cost from x to y
  - x maintains distance vector $\boldsymbol{D}_x = [D_x(y): y \in N]$
- node x:
  - knows cost to each neighbor v: $c(x, v)$
  - maintains its neighbors' distance vectors.
  - for each neighbor v, node x maintains

$$\boldsymbol{D}_v = [D_v(y): y \in N]$$

Distance-Vector Example

$D_x(y) = \min\{c(x,y) + D_y(y), c(x,z) + D_z(y)\}$
$= \min\{2+0 , 7+1\} = 2$

$D_x(z) = \min\{c(x,y) +$
$D_y(z), c(x,z) + D_z(z)\}$
$= \min\{2+1 , 7+0\} = 3$

node x table

from \ cost to	x	y	z
x	0	2	7
y	∞	∞	∞
z	∞	∞	∞

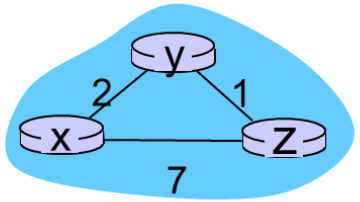
from \ cost to	x	y	z
x	0	2	3
y	2	0	1
z	7	1	0

node y table

from \ cost to	x	y	z
x	∞	∞	∞
y	2	0	1
z	∞	∞	∞

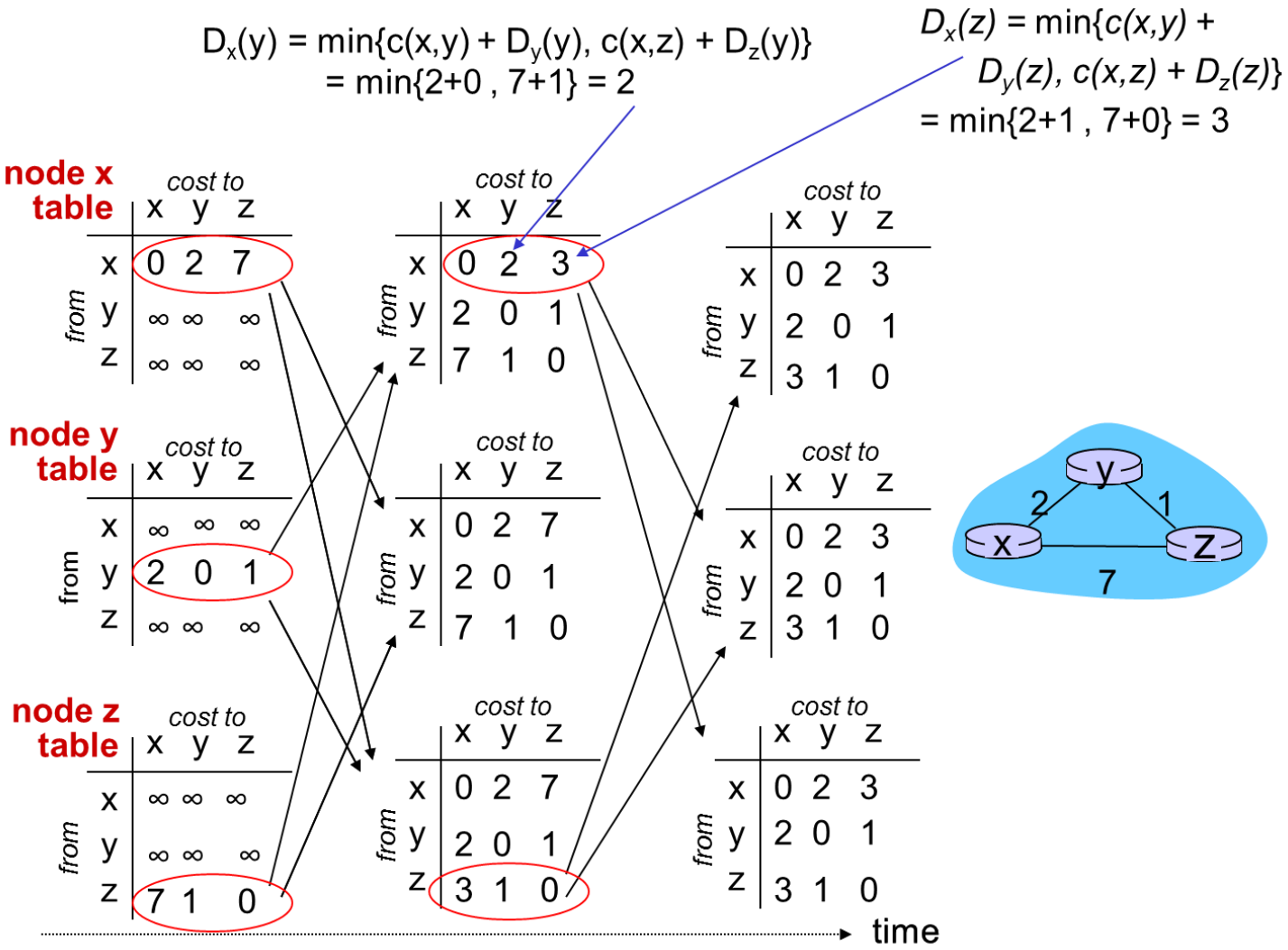
node z table

from \ cost to	x	y	z
x	∞	∞	∞
y	∞	∞	∞
z	7	1	0



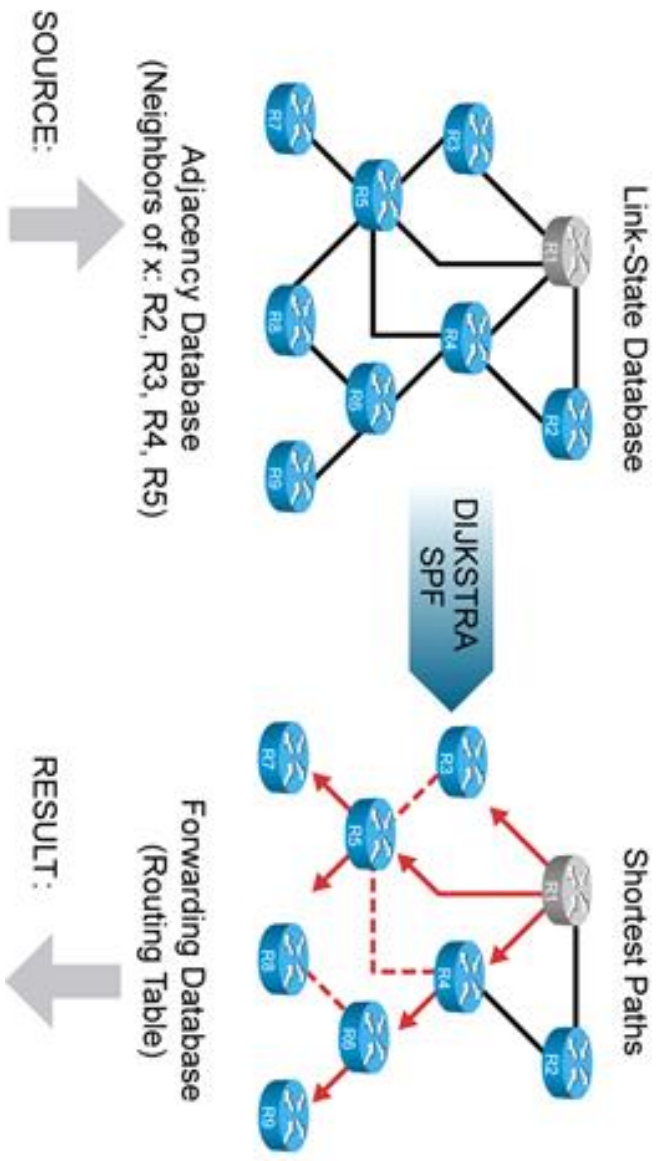
time

Distance-Vector Example



Link-State Protocol

- Network graph known to all nodes
  - Accomplished via “link state broadcast”
  - Topology information is flooded within the routing domain
- Computes least cost paths from one node (source) to all other nodes
  - Best end-to-end paths are computed locally at each router
  - Best end-to-end paths determine next-hops
- Iterative
  - after k iterations,
    know least cost path to k destinations
- Works only if policy is shared and uniform
- Examples: OSPF, IS-IS



Link-State Notation

$c(x, y)$

- := link cost from node x to y
- = ∞ if not direct neighbors

$D(v)$

- := current value of cost of path from source to dest. v

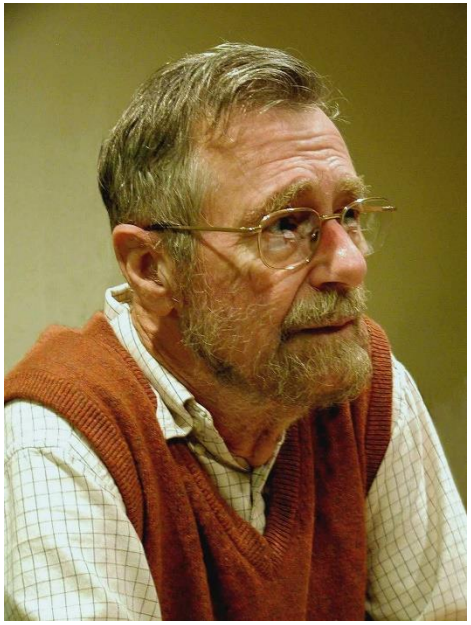
$p(v)$

- := predecessor node along path from source to v

N'

- := set of nodes whose least cost path definitively known

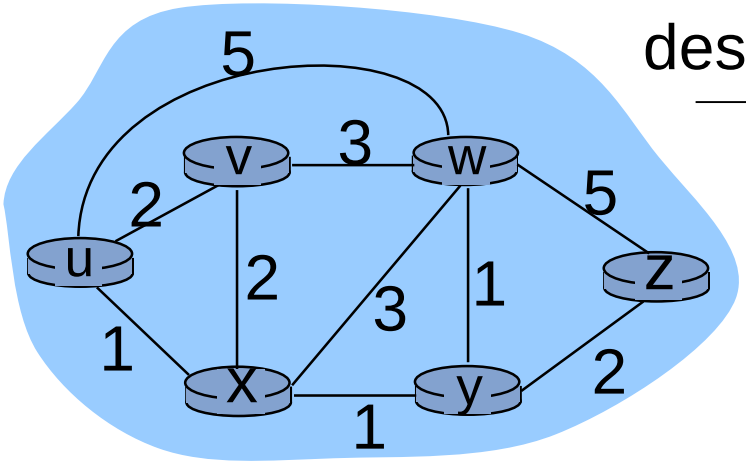
Dijkstra's Algorithm



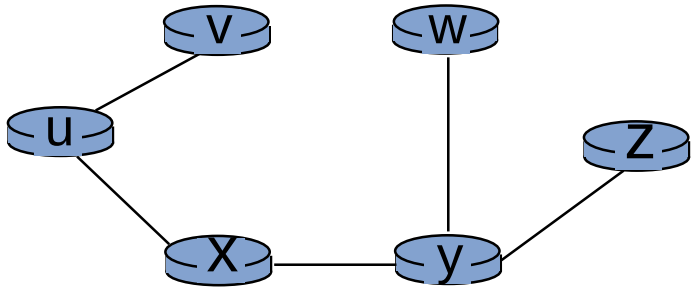
1 ***Initialization:***
2 $N' = \{u\}$
3 for all nodes v
4 if v adjacent to u
5 then $D(v) = c(u,v)$
6 else $D(v) = \infty$
7
8 ***Loop***
9 find w not in N' such that $D(w)$ is a minimum
10 add w to N'
11 update $D(v)$ for all v adjacent to w and not in N' :
12 $D(v) = \min(\boldsymbol{D}(v), \boldsymbol{D}(w) + c(w,v))$
13 */* new cost to v is either old cost to v or known*
14 *shortest path cost to w plus cost from w to v */*
15 ***until all nodes in*** N'

Dijkstra's algorithm: Example

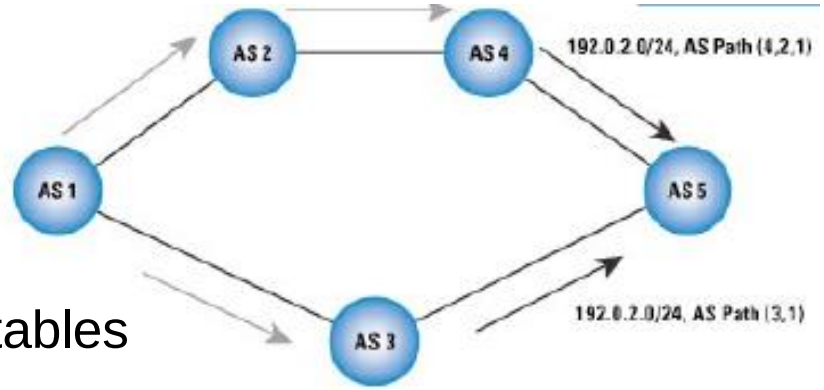
Step	N'	$D(v), p(v)$	$D(w), p(w)$	$D(x), p(x)$	$D(y), p(y)$	$D(z), p(z)$
0	u	$2, u$	$5, u$	$1, u$	∞	∞
1	ux	$2, u$	$4, x$		$2, x$	∞
2	uxy	$2, u$	$3, y$			$4, y$
3	$uxyv$		$3, y$			$4, y$
4	$uxyvw$					$4, y$
5	$uxyvwz$					



destination	link
v	(u, v)
x	(u, x)
y	(u, x)
w	(u, x)
z	(u, x)



Hierarchical Routing



- Scaling issues
  - can't store all destinations in routing tables
  - routing table exchange would swamp links
- Administrative autonomy
  - Internet = network of networks
  - each network admin may want to control routing in its own network
- **Autonomous systems**
  - aggregate routers into regions
  - http://www.cidr-report.org/as2.0/autnums.html
  - 16-bit or 32-bit long
  - Public (range from 1 to 64 511) and private (64 512 – 65 534) parts

address prefix	AS-path

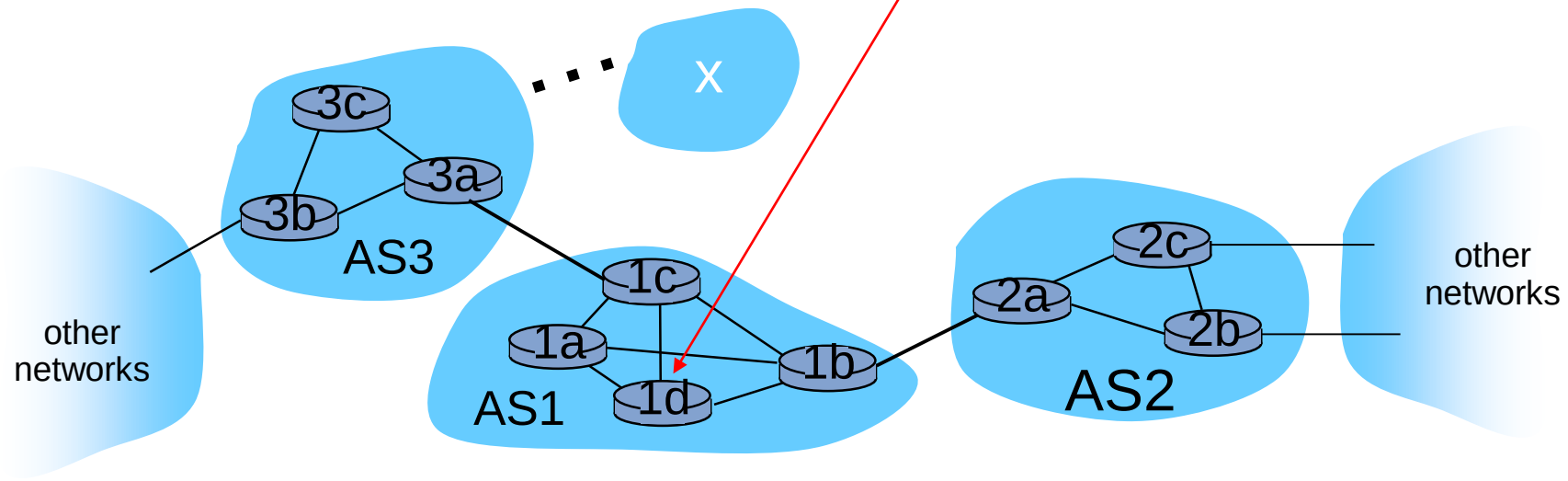
Inter-AS Routing: Exterior GW Protocol

- Suppose router in AS1 receives datagram destined outside of AS1:
- *Router should forward packet to gateway router, but which one?*

- AS1 must:

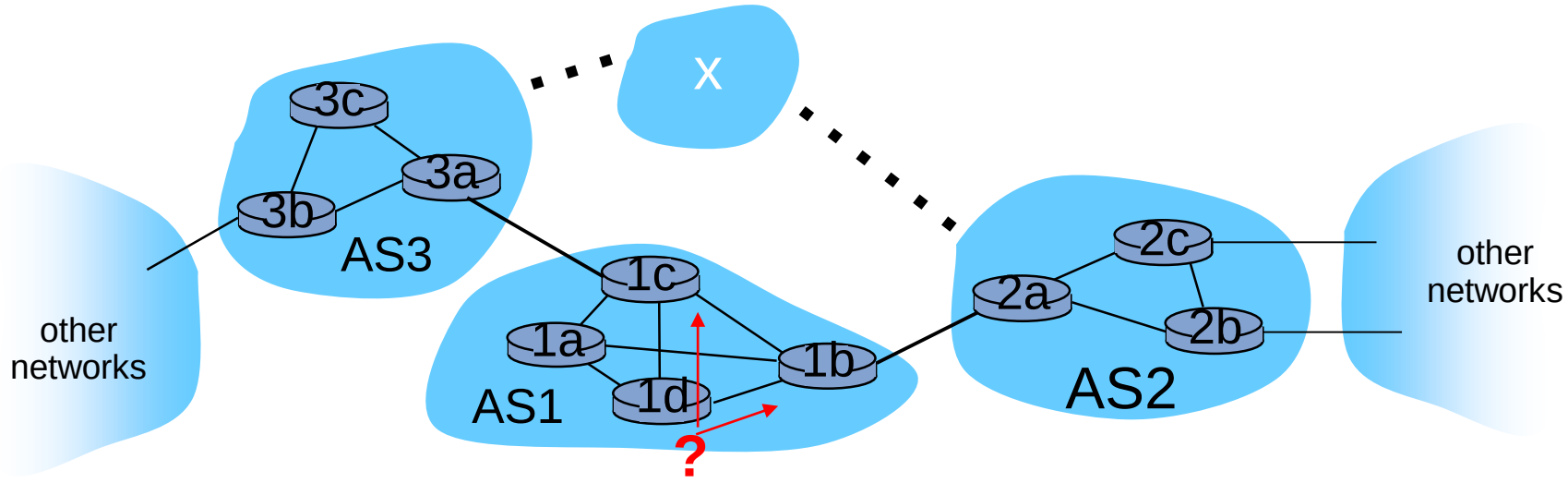
1) learn which destinations are reachable through AS2, which through AS3
2) propagate this reachability info to all routers in AS1

job of inter-AS routing!



Intra-AS Routing: Interior GW Protocol

- Now suppose AS1 learns from inter-AS protocol that subnet x is reachable from AS3 and from AS2
- To configure forwarding table, router 1d must determine which gateway it should forward packets towards for dest x
  - this is also job of inter-AS routing protocol!



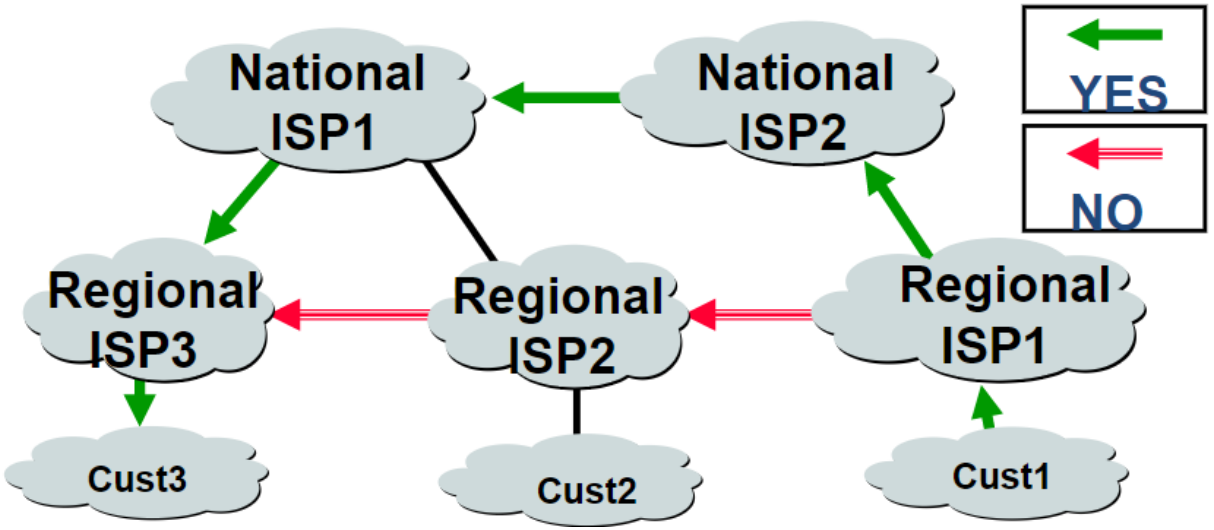
Inter-AS Goals

- **Should scale** for the size of the global Internet
  - Focus on reachability not optimality
  - Use address aggregation techniques to minimize core routing table sizes and associated control traffic

- Allow **policy-based routing** between autonomous systems
  - Policy refers to arbitrary preference among a menu of available routes (based upon routes' attributes)
  - Fully distributed routing (as opposed to a signaled approach) is the only possibility
  - Extensible to meet the demands for newer policies

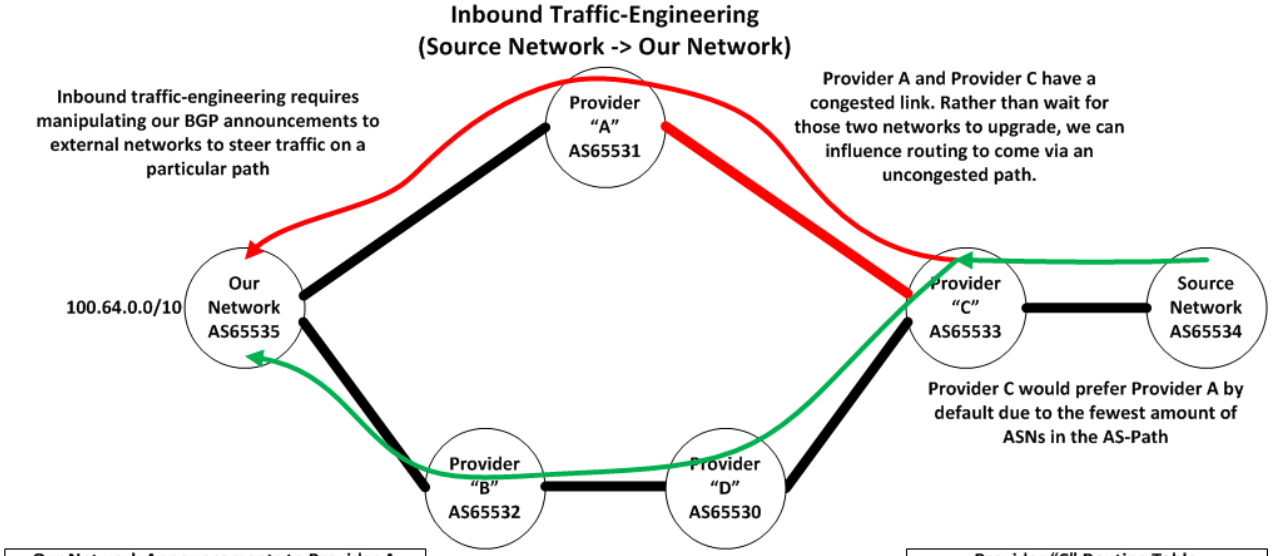
Path-Vector Protocol

- Sends the whole path (not just distance)
  - Loop prevention
- Route contains several different independent metrics
  - Preferred exit from ASN
  - Preferred entry to ASN
  - ASN Path
  - Origin
  - Community



Flexible Policies

- Each node may apply local routing policy
  - to choose a route
  - to change and reannounce route with different attributes
- **Inbound** vs. **outbound** traffic-enginneering



Our Network Announcements to Provider A	
Prefix	AS-Path
100.64.0.0/10	65535 65535 65535

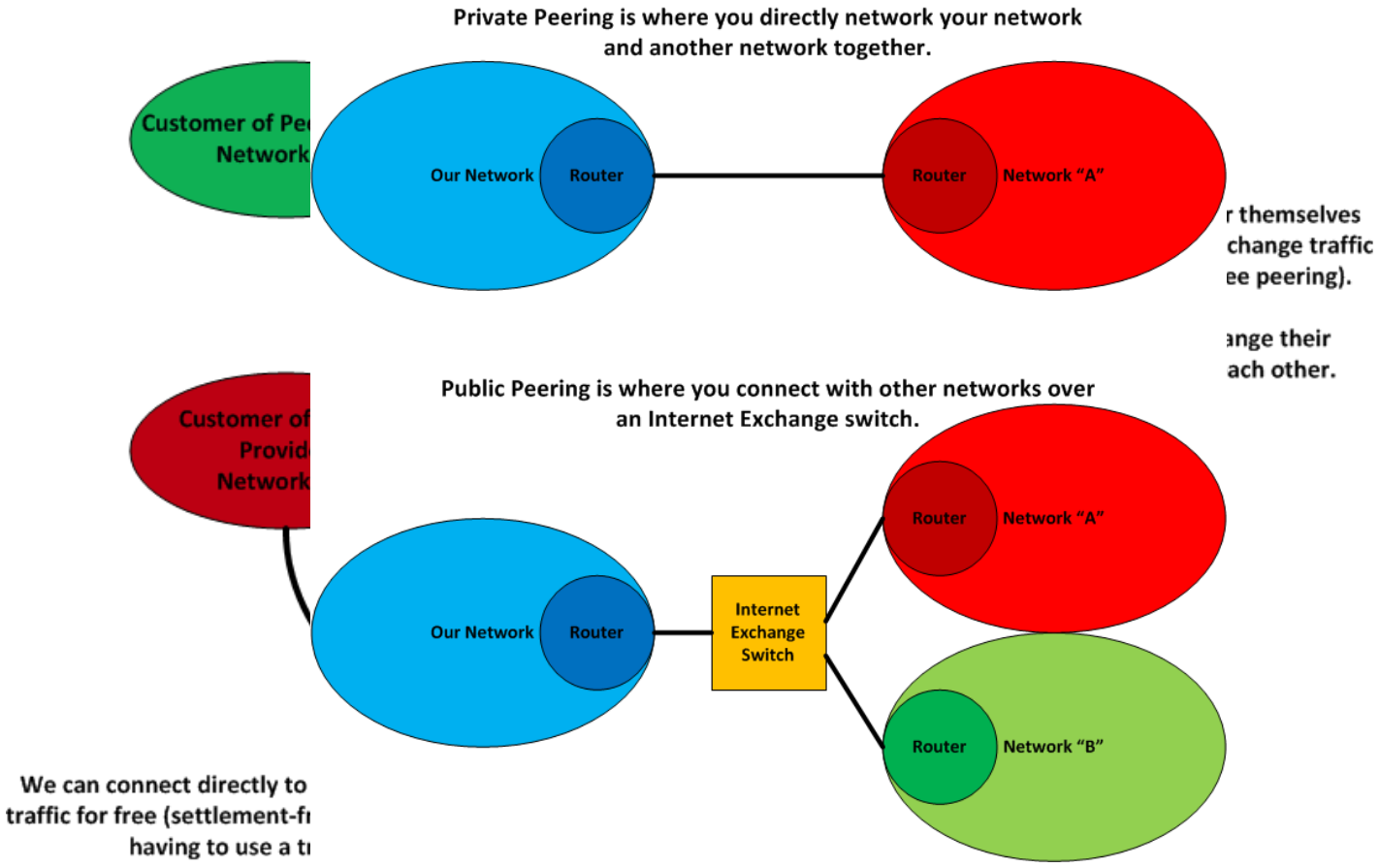
Our Network Announcements to Provider B	
Prefix	AS-Path
100.64.0.0/10	65535

Provider "C" Routing Table	
Prefix	AS-Path
10.0.0.0/8	65531 65535 65535 65535
10.0.0.0/8	65530 65532 65535

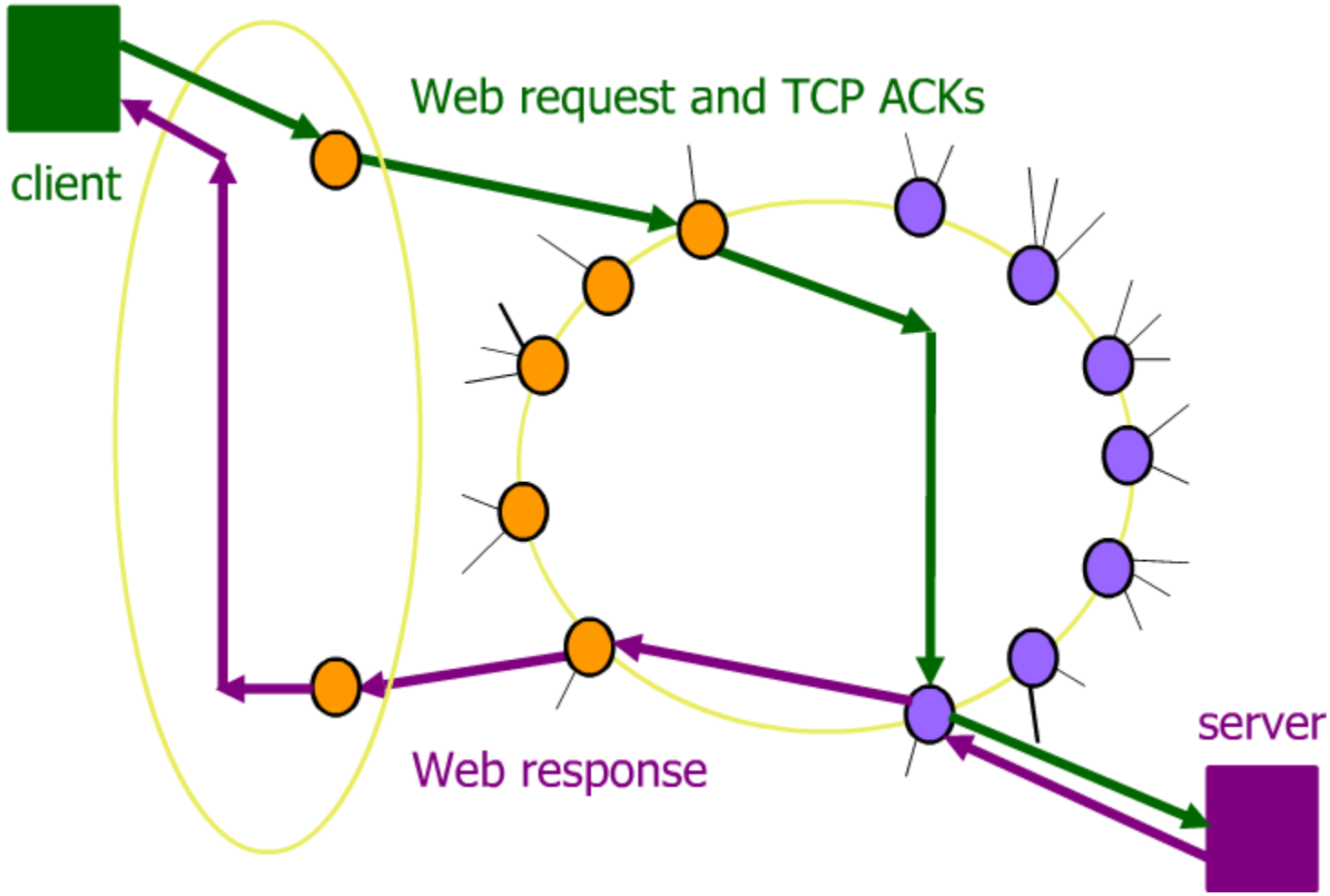
By prepending our ASN multiple times in the path, we can influence Provider C into selecting Provider D to reach our network

Peering

- = exchange of traffic between AS
- BRIX, NIX.CZ (http://www.nix.cz/en/technical#traffic)



Asymmetric Routing

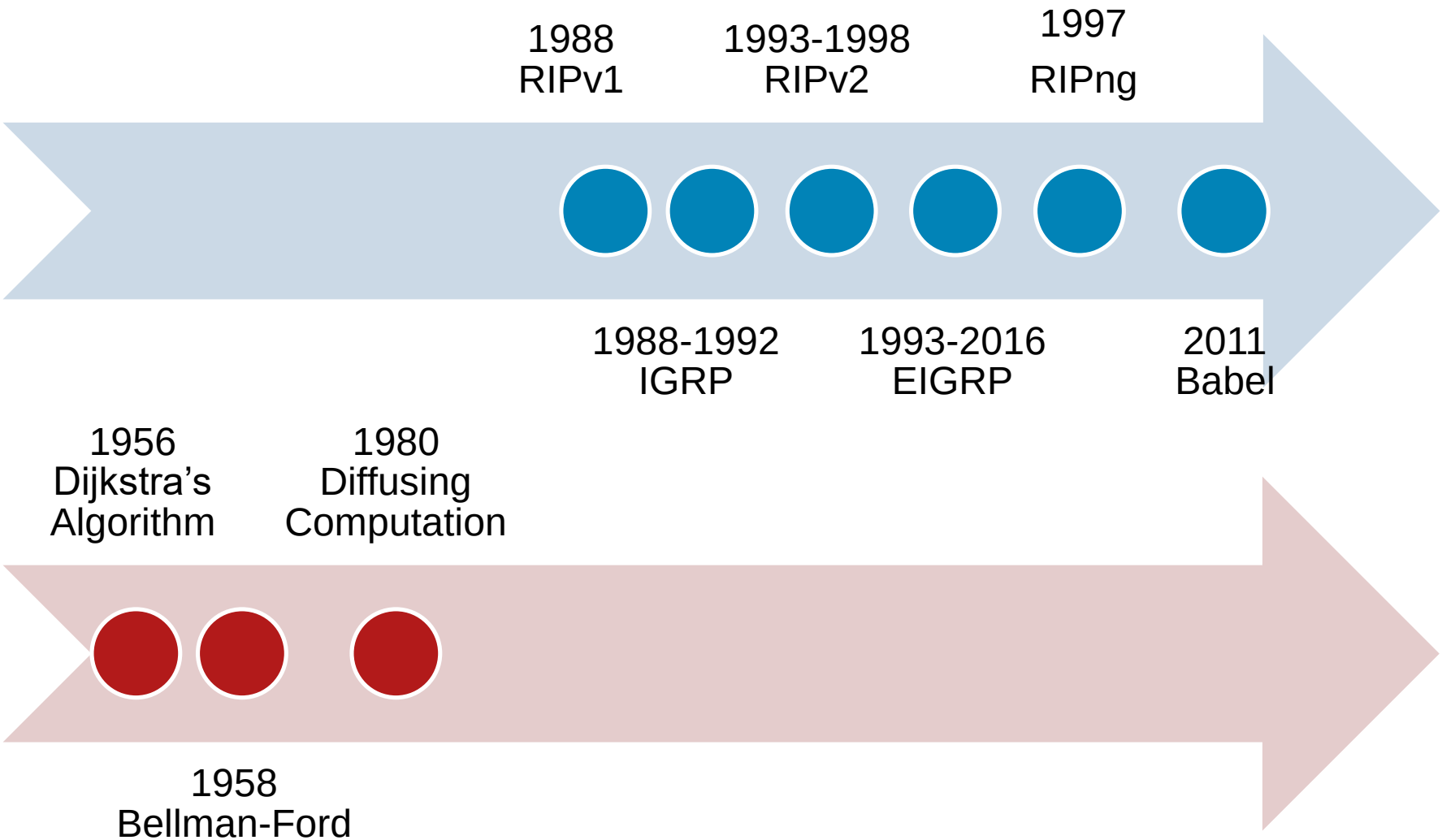


Distance-Vector Protocol Evolution

“Time is the longest distance between two places.”

Tennessee Williams, The Glass Menagerie

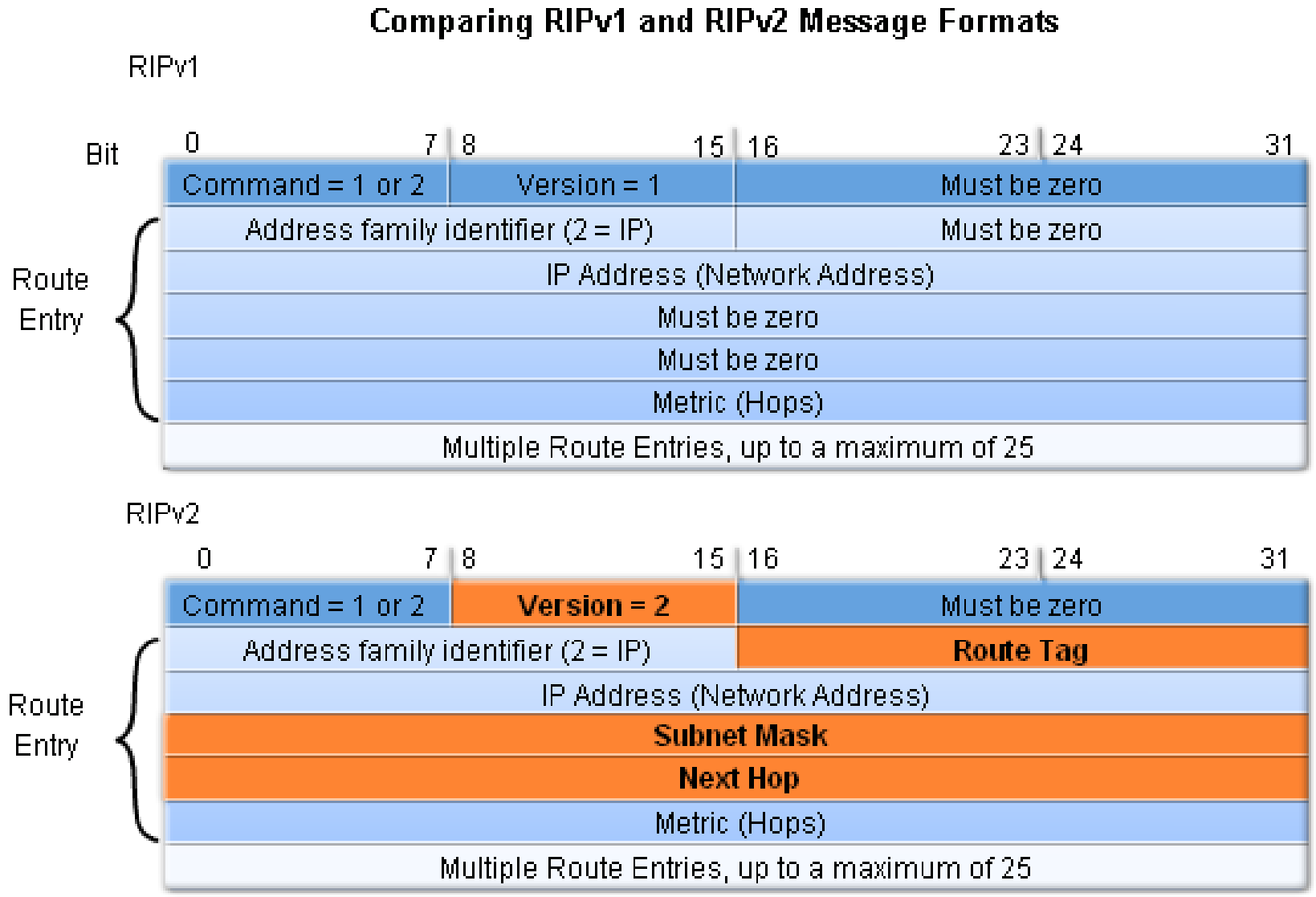
Timeline


1988
RIPv1
1993-1998
RIPv2
1997
RIPng
1988-1992
IGRP
1993-2016
EIGRP
2011
Babel
1956
Dijkstra's
Algorithm
1980
Diffusing
Computation
1958
Bellman-Ford
Algorithm

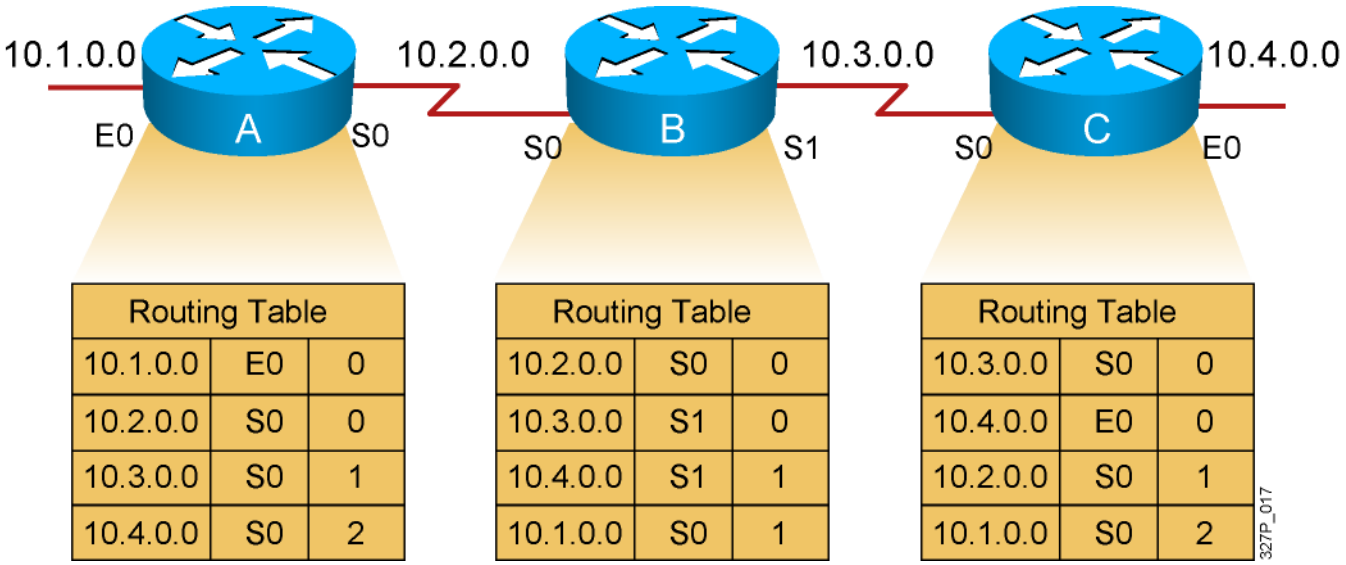
RIP

- Versions
  - RIPv1 – classful, RFC 1058
  - RIPv2 – classless, RFC 1388
  - RIPng – IPv6 support, RFC 2080
- Messages
  - uses UDP datagrams on port 520
  - two types Request and (un)solicited Response
  - size of datagram limited to 512 bytes (allow advertisement of 25 routes)
- Hop count as metric
  - Distance from the advertising router to the destination network
- No neighbor detection
  - just periodic exchange of DV every 30 seconds
  - unsolicit/solicit updates goverened by other timers

RIP Messages

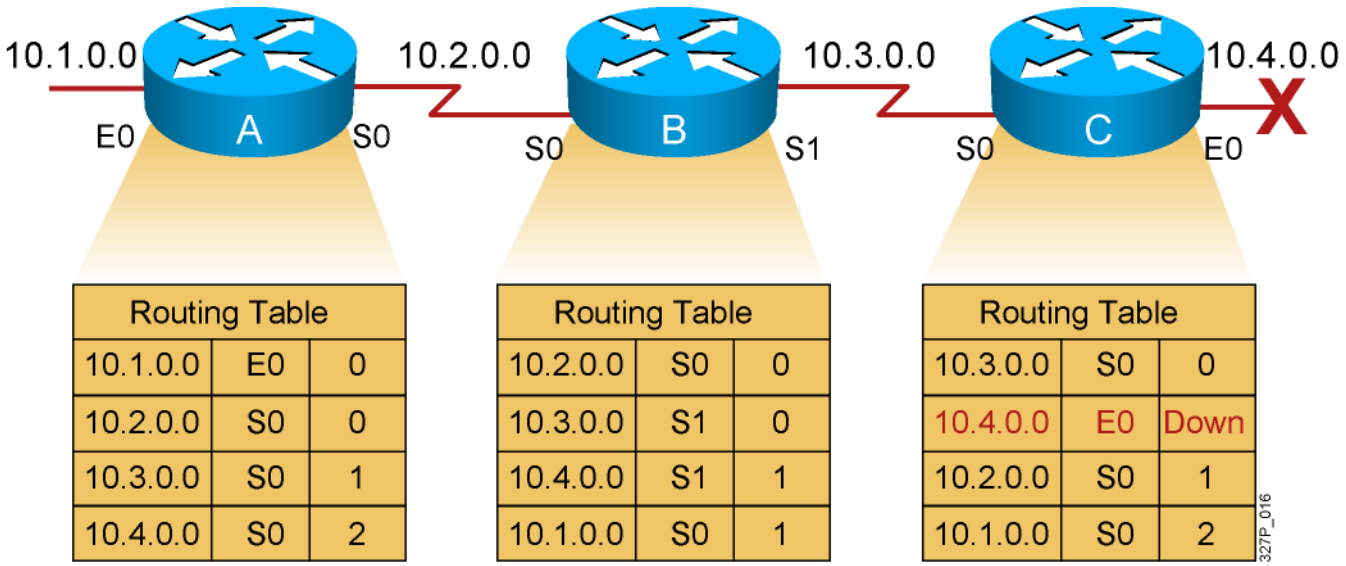


RIP: Counting to Infinity



Each node maintains the distance from itself to each possible destination network.

RIP: Counting to Infinity



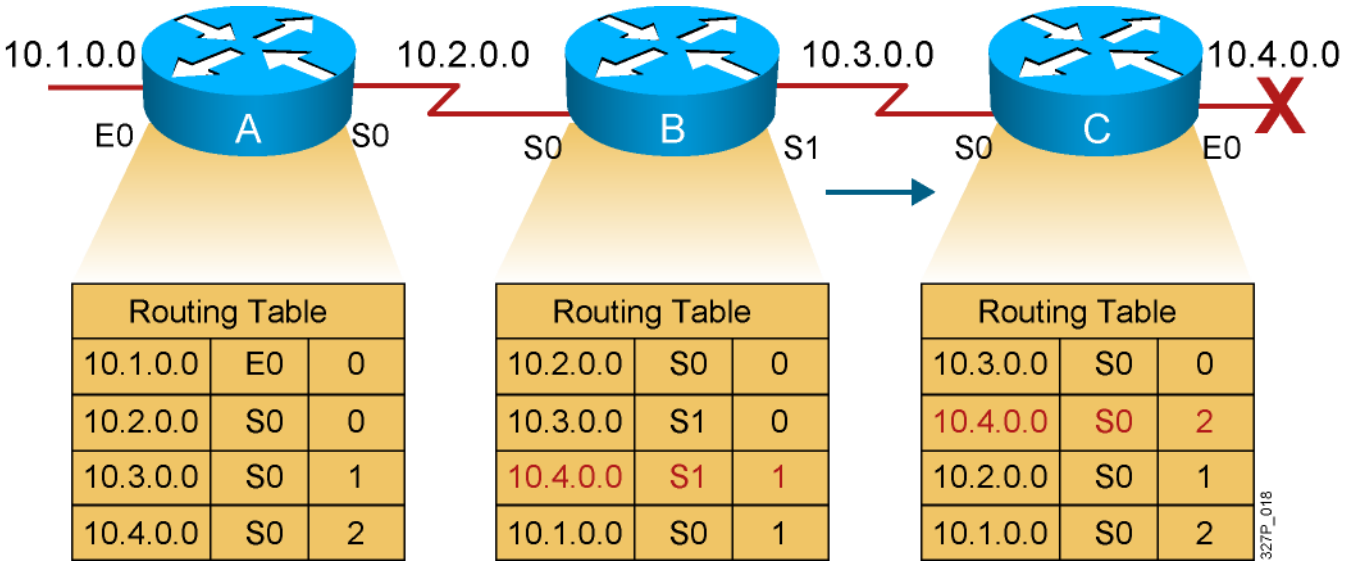
Routing Table		
10.1.0.0	E0	0
10.2.0.0	S0	0
10.3.0.0	S0	1
10.4.0.0	S0	2

Routing Table		
10.2.0.0	S0	0
10.3.0.0	S1	0
10.4.0.0	S1	1
10.1.0.0	S0	1

Routing Table		
10.3.0.0	S0	0
10.4.0.0	E0	Down
10.2.0.0	S0	1
10.1.0.0	S0	2

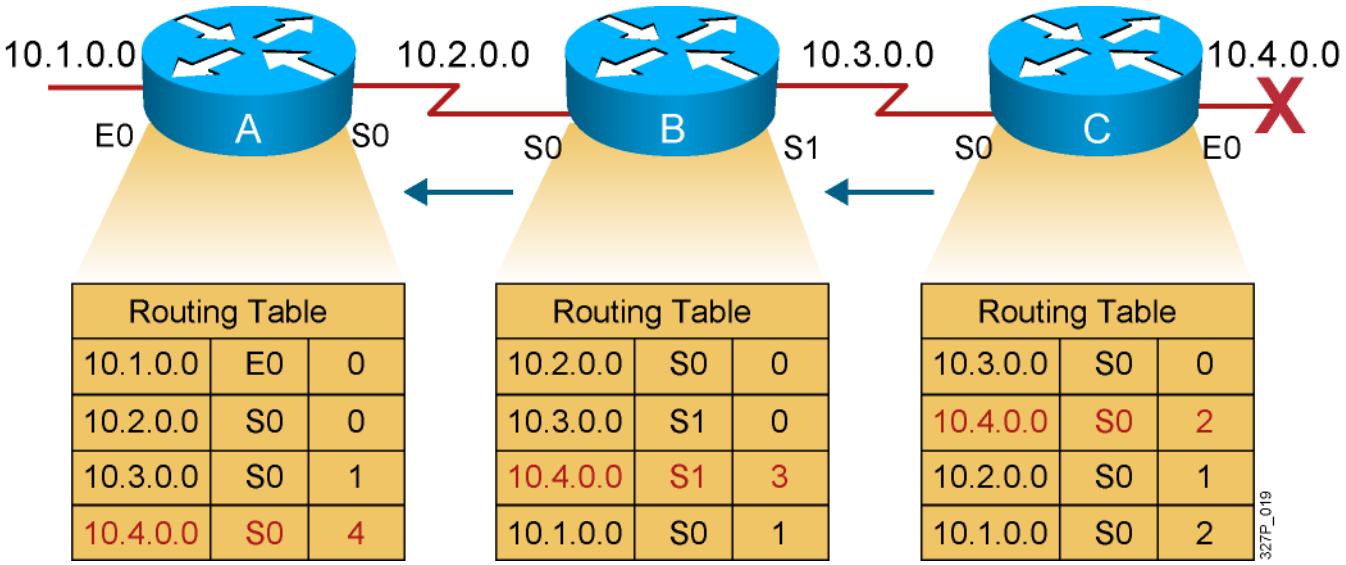
Slow convergence produces inconsistent routing.

RIP: Counting to Infinity (Cont.)



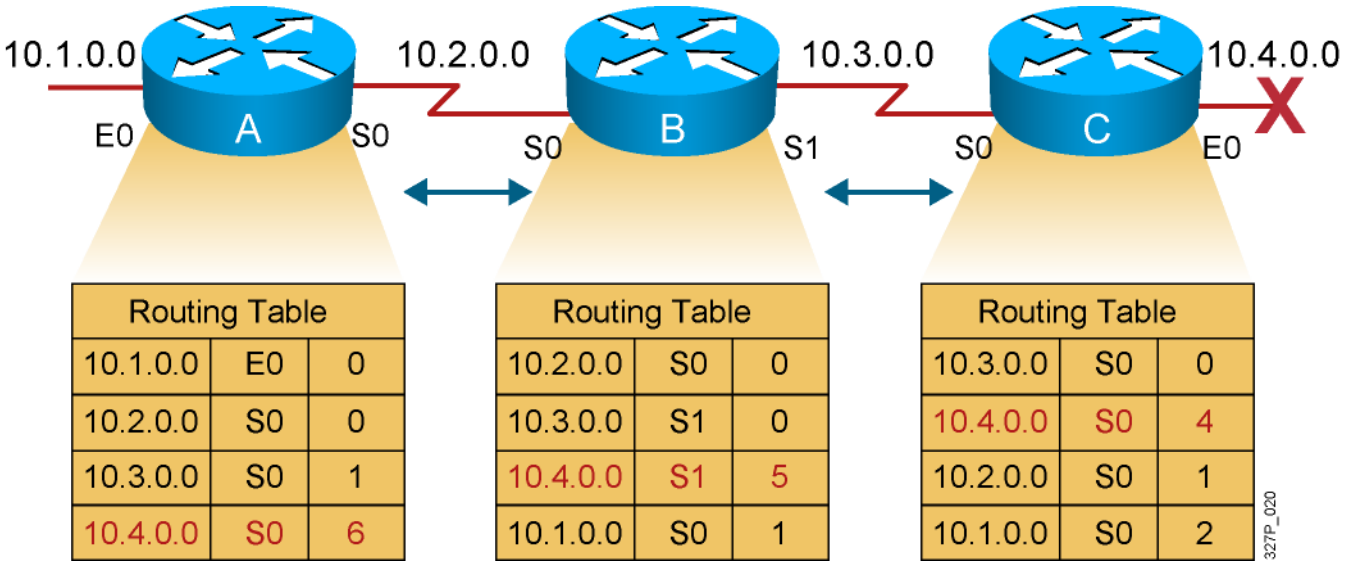
Router C concludes that the best path to network 10.4.0.0 is through router B.

RIP: Counting to Infinity (Cont.)



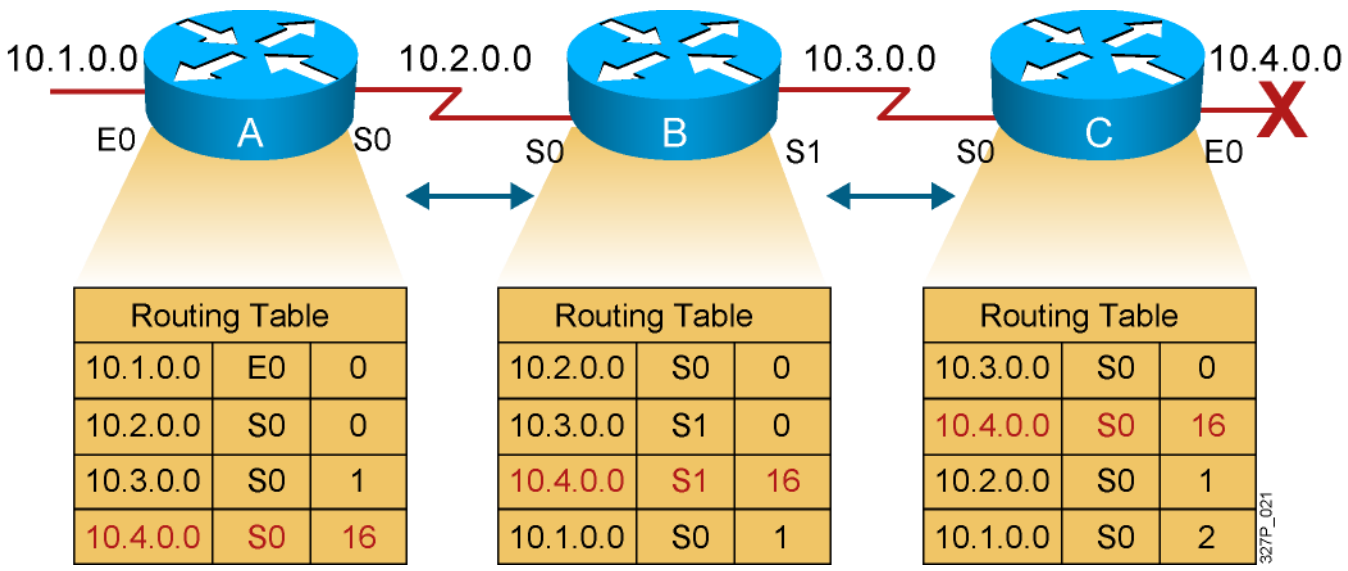
Router A updates its table to reflect the new but erroneous hop count.

RIP: Counting to Infinity (Cont.)



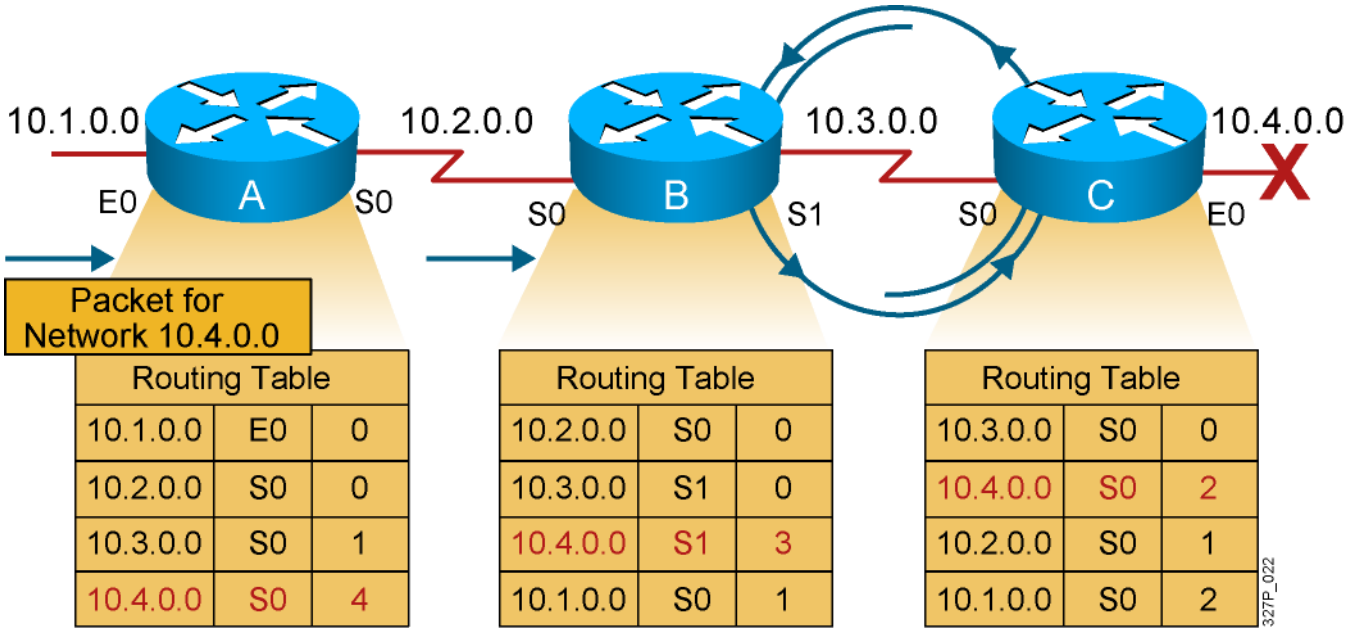
The hop count for network 10.4.0.0 counts to infinity.

RIP: Defining a Maximum



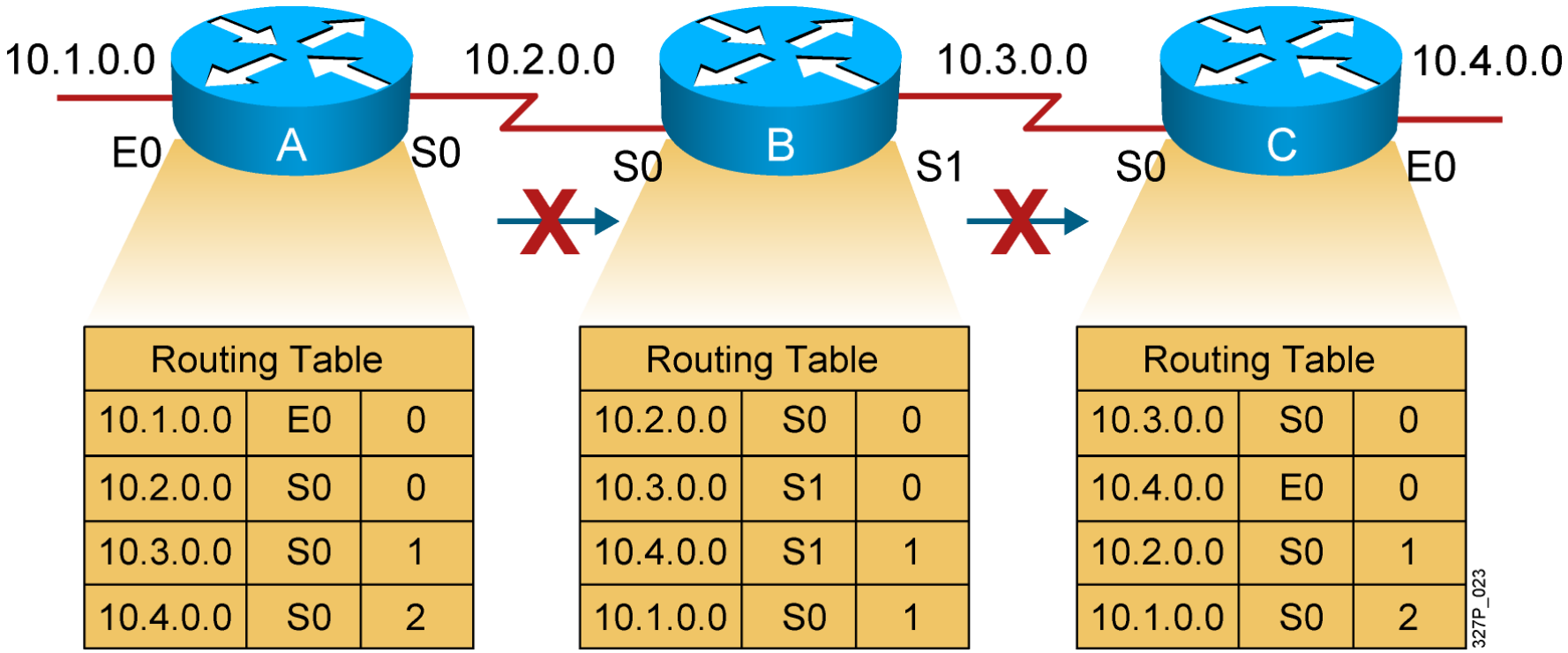
A limit is set on the number of hops to prevent infinite loops.

RIP: Routing Loops



Packets for network 10.4.0.0 bounce (loop) between routers B and C.

RIP: Split Horizon



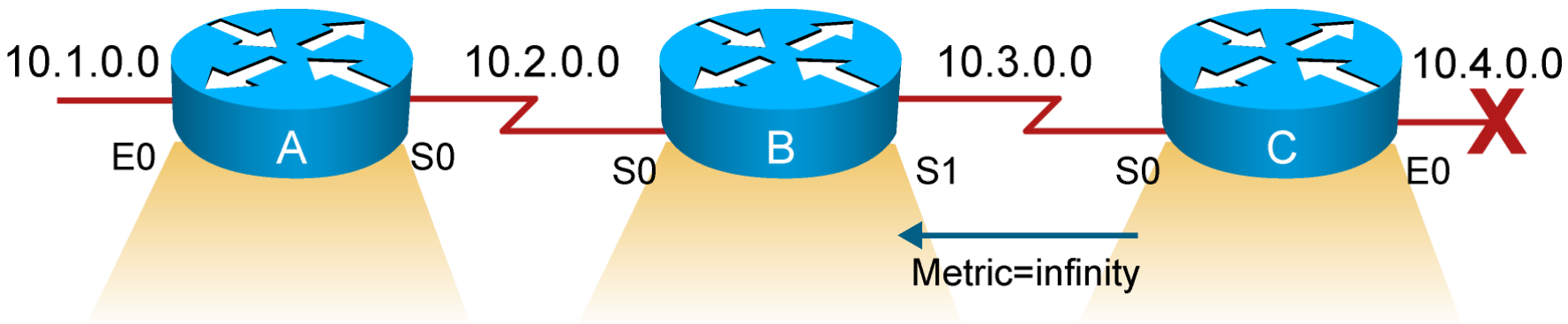
Routing Table		
10.1.0.0	E0	0
10.2.0.0	S0	0
10.3.0.0	S0	1
10.4.0.0	S0	2

Routing Table		
10.2.0.0	S0	0
10.3.0.0	S1	0
10.4.0.0	S1	1
10.1.0.0	S0	1

Routing Table		
10.3.0.0	S0	0
10.4.0.0	E0	0
10.2.0.0	S0	1
10.1.0.0	S0	2

It is never useful to send information about a route back in the direction from which the original information came.

RIP: Poison Reverse



Routing Table		
10.1.0.0	E0	0
10.2.0.0	S0	0
10.3.0.0	S0	1
10.4.0.0	S0	2

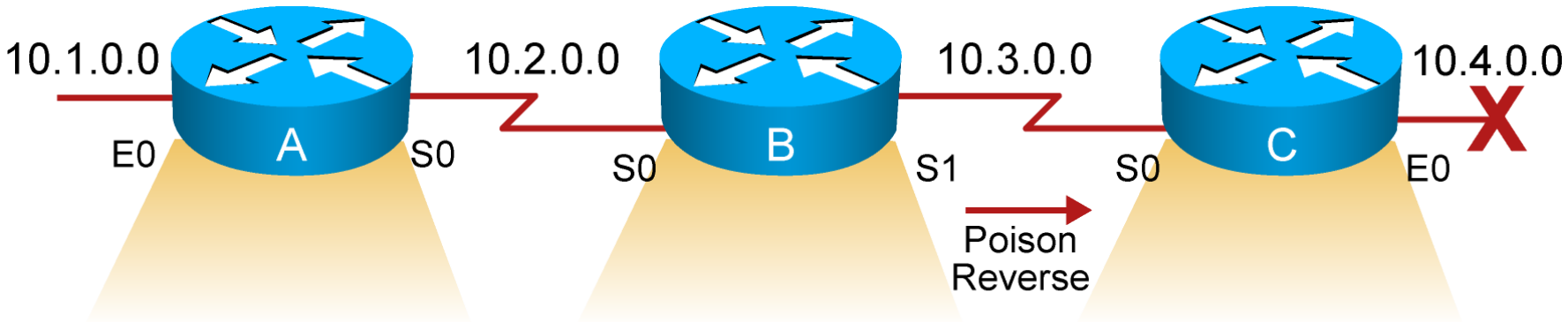
Routing Table		
10.2.0.0	S0	0
10.3.0.0	S1	0
10.4.0.0	S1	1
10.1.0.0	S0	1

Routing Table		
10.3.0.0	S0	0
10.4.0.0	E0	Down
10.2.0.0	S0	1
10.1.0.0	S0	2

327P_024

Routers advertise the distance of routes that have gone down to infinity.

RIP: Poison Reverse (Cont.)



Routing Table		
10.1.0.0	E0	0
10.2.0.0	S0	0
10.3.0.0	S0	1
10.4.0.0	S0	2

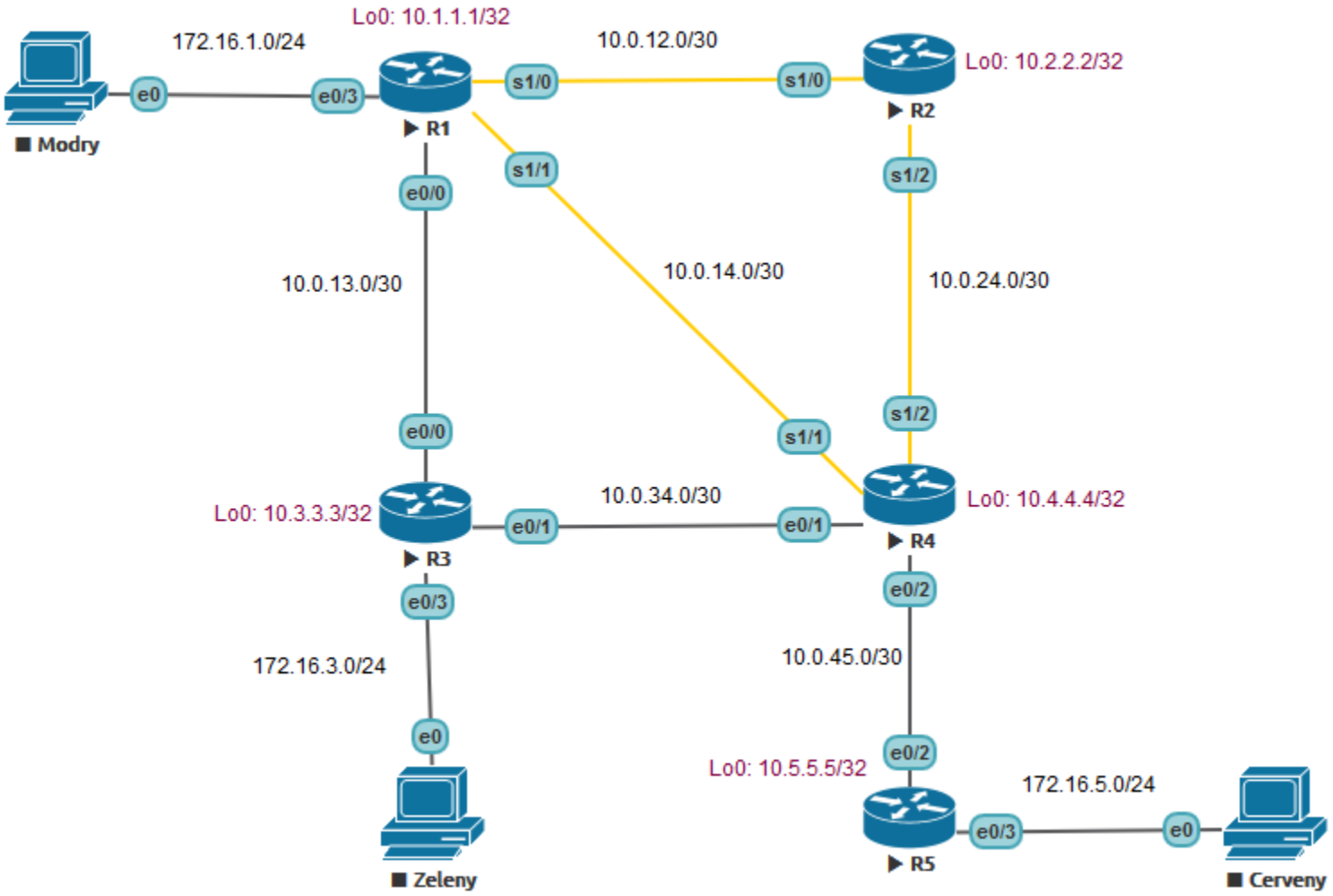
Routing Table		
10.2.0.0	S0	0
10.3.0.0	S1	0
10.4.0.0	S1	Possibly Down
10.1.0.0	S0	1

Routing Table		
10.3.0.0	S0	0
10.4.0.0	S0	Infinity
10.2.0.0	S0	1
10.1.0.0	S0	2

327P_025

Poison reverse overrides split horizon.

RIP: Network Graph



RIP: Database

R1

```
R1#sh ip rip database
10.0.0.0/8    auto-summary
10.0.12.0/30    directly connected, Serial1/0
10.0.13.0/30    directly connected, Ethernet0/0
10.0.14.0/30    directly connected, Serial1/1
10.0.24.0/30
    [1] via 10.0.12.2, 00:00:24, Serial1/0
    [1] via 10.0.14.2, 00:00:03, Serial1/1
10.0.34.0/30
    [1] via 10.0.13.2, 00:00:07, Ethernet0/0
    [1] via 10.0.14.2, 00:00:03, Serial1/1
10.0.45.0/30
    [1] via 10.0.14.2, 00:00:03, Serial1/1
10.1.1.1/32    directly connected, Loopback0
10.2.2.2/32
    [1] via 10.0.12.2, 00:00:24, Serial1/0
10.3.3.3/32
    [1] via 10.0.13.2, 00:00:07, Ethernet0/0
10.4.4.4/32
    [1] via 10.0.14.2, 00:00:03, Serial1/1
10.5.5.5/32
    [2] via 10.0.14.2, 00:00:03, Serial1/1
172.16.0.0/16    auto-summary
172.16.1.0/24    directly connected, Ethernet0/3
172.16.3.0/24
    [1] via 10.0.13.2, 00:00:07, Ethernet0/0
172.16.5.0/24
    [2] via 10.0.14.2, 00:00:03, Serial1/1
R1#
R1#
```

R4

```
R4#
R4#show ip rip database
10.0.0.0/8    auto-summary
10.0.12.0/30
    [1] via 10.0.24.1, 00:00:08, Serial1/2
    [1] via 10.0.14.1, 00:00:04, Serial1/1
10.0.13.0/30
    [1] via 10.0.34.1, 00:00:02, Ethernet0/1
    [1] via 10.0.14.1, 00:00:04, Serial1/1
10.0.14.0/30    directly connected, Serial1/1
10.0.24.0/30    directly connected, Serial1/2
10.0.34.0/30    directly connected, Ethernet0/1
10.0.45.0/30    directly connected, Ethernet0/2
10.1.1.1/32
    [1] via 10.0.14.1, 00:00:04, Serial1/1
10.2.2.2/32
    [1] via 10.0.24.1, 00:00:08, Serial1/2
10.3.3.3/32
    [1] via 10.0.34.1, 00:00:02, Ethernet0/1
10.4.4.4/32    directly connected, Loopback0
10.5.5.5/32
    [1] via 10.0.45.2, 00:00:21, Ethernet0/2
172.16.0.0/16    auto-summary
172.16.1.0/24
    [1] via 10.0.14.1, 00:00:04, Serial1/1
172.16.3.0/24
    [1] via 10.0.34.1, 00:00:02, Ethernet0/1
172.16.5.0/24
    [1] via 10.0.45.2, 00:00:21, Ethernet0/2
R4#
```

RIP: Routing Table

R1

```
       ia - IS-IS inter area, * - candidate default, U - per-user s
       o - ODR, P - periodic downloaded static route, H - NHRP, l -
       a - application route
       + - replicated route, % - next hop override

Gateway of last resort is not set

      10.0.0.0/8 is variably subnetted, 14 subnets, 2 masks
C        10.0.12.0/30 is directly connected, Serial1/0
L        10.0.12.1/32 is directly connected, Serial1/0
C        10.0.13.0/30 is directly connected, Ethernet0/0
L        10.0.13.1/32 is directly connected, Ethernet0/0
C        10.0.14.0/30 is directly connected, Serial1/1
L        10.0.14.1/32 is directly connected, Serial1/1
R        10.0.24.0/30 [120/1] via 10.0.14.2, 00:00:11, Serial1/1
                      [120/1] via 10.0.12.2, 00:00:06, Serial1/0
R        10.0.34.0/30 [120/1] via 10.0.14.2, 00:00:11, Serial1/1
                      [120/1] via 10.0.13.2, 00:00:19, Ethernet0/0
R        10.0.45.0/30 [120/1] via 10.0.14.2, 00:00:11, Serial1/1
C        10.1.1.1/32 is directly connected, Loopback0
R        10.2.2.2/32 [120/1] via 10.0.12.2, 00:00:06, Serial1/0
R        10.3.3.3/32 [120/1] via 10.0.13.2, 00:00:19, Ethernet0/0
R        10.4.4.4/32 [120/1] via 10.0.14.2, 00:00:11, Serial1/1
R        10.5.5.5/32 [120/2] via 10.0.14.2, 00:00:11, Serial1/1
      172.16.0.0/16 is variably subnetted, 4 subnets, 2 masks
C        172.16.1.0/24 is directly connected, Ethernet0/3
L        172.16.1.1/32 is directly connected, Ethernet0/3
R        172.16.3.0/24 [120/1] via 10.0.13.2, 00:00:19, Ethernet0/0
R        172.16.5.0/24 [120/2] via 10.0.14.2, 00:00:11, Serial1/1
R1#
```

R4

```
       ia - IS-IS inter area, * - candidate default, U - per-user s
       o - ODR, P - periodic downloaded static route, H - NHRP, l -
       a - application route
       + - replicated route, % - next hop override

Gateway of last resort is not set

      10.0.0.0/8 is variably subnetted, 15 subnets, 2 masks
R        10.0.12.0/30 [120/1] via 10.0.24.1, 00:00:01, Serial1/2
                      [120/1] via 10.0.14.1, 00:00:22, Serial1/1
R        10.0.13.0/30 [120/1] via 10.0.34.1, 00:00:24, Ethernet0/1
                      [120/1] via 10.0.14.1, 00:00:22, Serial1/1
C        10.0.14.0/30 is directly connected, Serial1/1
L        10.0.14.2/32 is directly connected, Serial1/1
C        10.0.24.0/30 is directly connected, Serial1/2
L        10.0.24.2/32 is directly connected, Serial1/2
C        10.0.34.0/30 is directly connected, Ethernet0/1
L        10.0.34.2/32 is directly connected, Ethernet0/1
C        10.0.45.0/30 is directly connected, Ethernet0/2
L        10.0.45.1/32 is directly connected, Ethernet0/2
R        10.1.1.1/32 [120/1] via 10.0.14.1, 00:00:22, Serial1/1
R        10.2.2.2/32 [120/1] via 10.0.24.1, 00:00:01, Serial1/2
R        10.3.3.3/32 [120/1] via 10.0.34.1, 00:00:24, Ethernet0/1
C        10.4.4.4/32 is directly connected, Loopback0
R        10.5.5.5/32 [120/1] via 10.0.45.2, 00:00:16, Ethernet0/2
      172.16.0.0/24 is subnetted, 3 subnets
R        172.16.1.0 [120/1] via 10.0.14.1, 00:00:22, Serial1/1
R        172.16.3.0 [120/1] via 10.0.34.1, 00:00:24, Ethernet0/1
R        172.16.5.0 [120/1] via 10.0.45.2, 00:00:16, Ethernet0/2
R4#
```

EIGRP

- RFC 7868
- Messages
  - EIGRP has modular structure independent on routed protocol (L3 protocol)
  - Encapsulated directly into IPv4, IPv6, IPX, AppleTalk
- Classless (VLSM)
  - Automatic and manual summarization, authentication, stub routing
- Composite metric based on multiple factors
- Neighbor Detection
  - Every router has its own neighbor table where it stores information about directly connected neighbors
- Reliable Transport Protocol (RTP)
  - Transport protocol independent on L3 protocol – protocol number 88
  - Guarantees delivery of unicast and multicast communication
- **DUAL Finite-state Automata** by J.J. Garcia-Luna-Aceves
  - It directs whole best route selection mechanism
  - Loop-free Topology Protection
    - Guarantees that each used next-hop doesn't cause routing loop in topology

EIGRP: Metric

- Composite metric consists of following factors
  - **K1 – Bandwidth** (static parameter, turned on by default)
  - **K3 – Delay** (static parameter, turned on by default)
  - **K4, K5 – Reliability** (dynamically evaluated, turned off by default)
  - **K2 – Load** (dynamically evaluated, turned off by default)
  - **K6 – Energy** (acummulative energy consumption)
    Jitter (accumulative delay variation)
  - MTU (some literature mentions it as tie-breaker, but it is in fact useless)

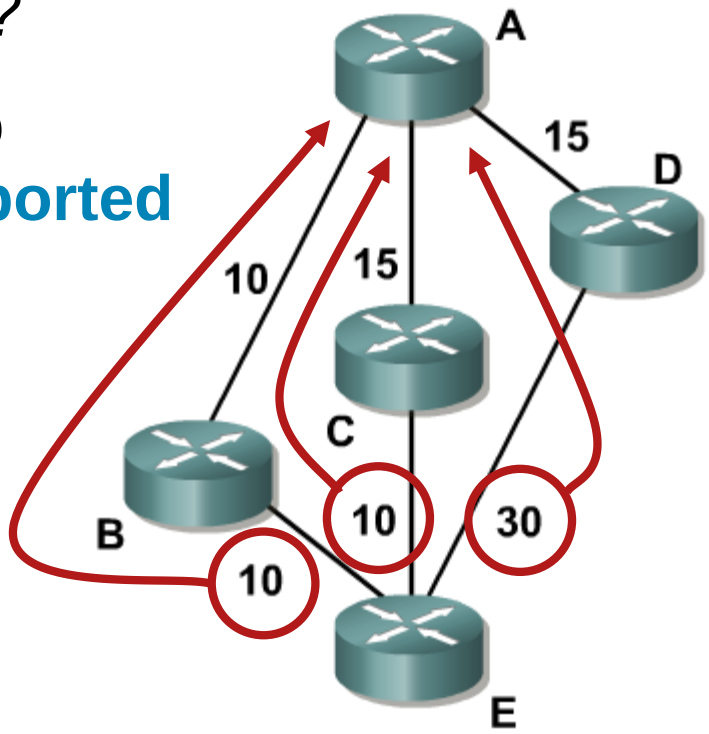
$$K_1 \cdot Bw + K_3 \cdot Dl$$

$$\left(K_1 \cdot Bw + \frac{K_2 \cdot Bw}{256 - Lo} + K_3 \cdot Dl\right) \cdot \frac{K_5}{Re + K_4}$$

$$\left(K_1 \cdot Bw + \frac{K_2 \cdot Bw}{256 - Lo} + K_3 \cdot Dl + K_6 \cdot (En + Ji)\right) \cdot \frac{K_5}{Re + K_4}$$

EIGRP: Use-Case ①

- *How does EIGRP know which routes don't cause loops in topology?*
- It uses neighbors distances to destination network called **reported distances (RD)**
- Every neighbor router of *A* is advertising its RD to *E*
  - RD(*B*) = 10
  - RD(*C*) = 10
  - RD(*D*) = 30

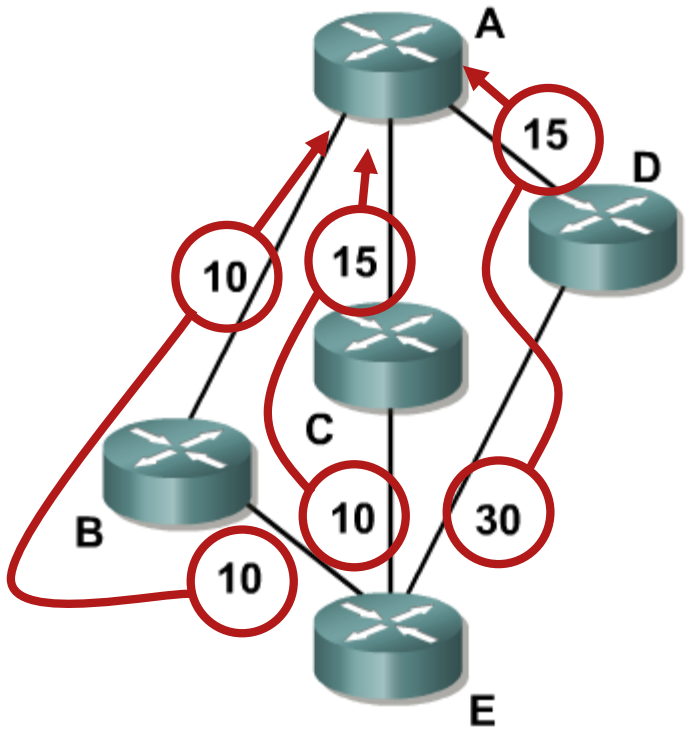


EIGRP: Use-Case ②

- For *A* is distance to *E*:
  - via(*B*) = 20
  - via(*C*) = 25
  - via(*D*) = 45

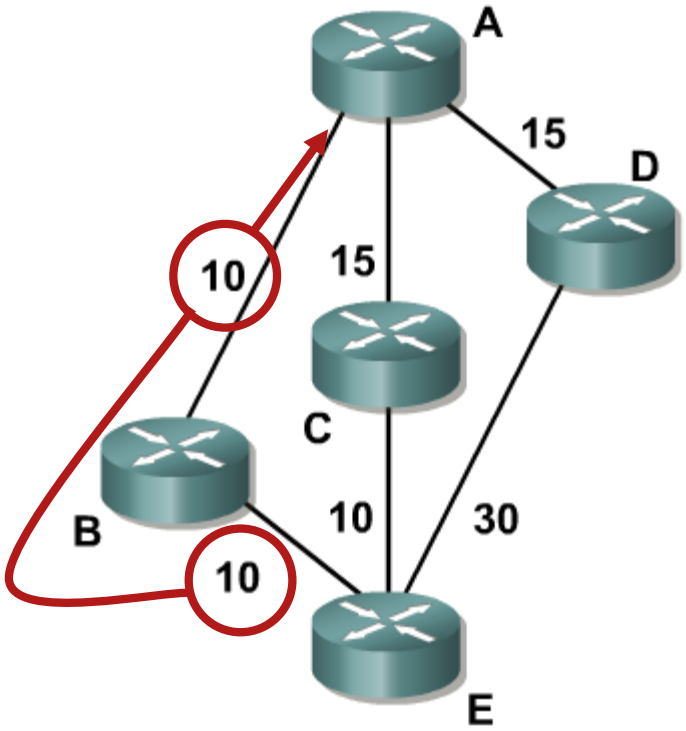
- Best route from *A* to *E* is via(*B*) and its distance is called **feasible distance (FD)**

- More precisely FD is by router *A* best known distance to destination network



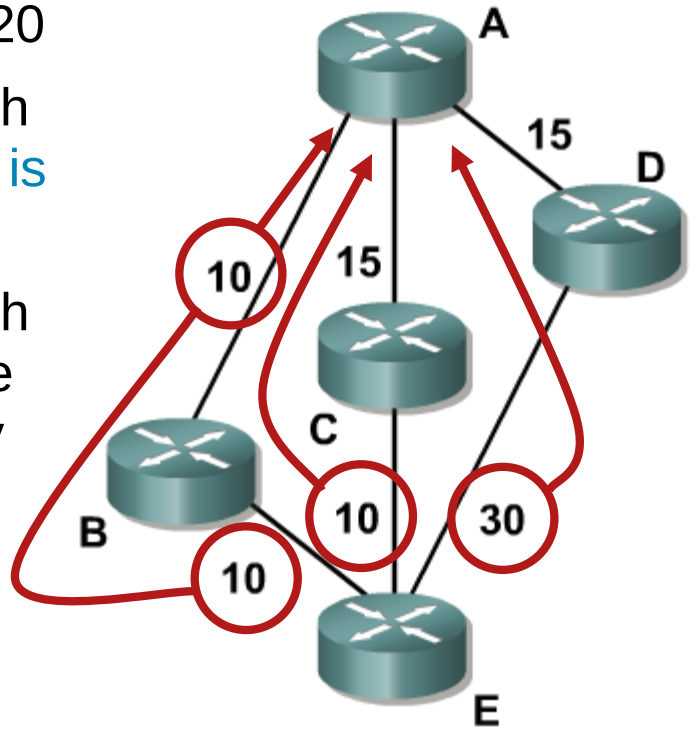
EIGRP: Use-Case ③

- Router *A* is using FD and RD to check **feasibility condition**
- FD is etalon for this validation – every route with **RD < FD** is without any doubts loopless
- Some loopless routes are (falsely) denied by this rule
- But it never accepts route which certainly cause loop in topology

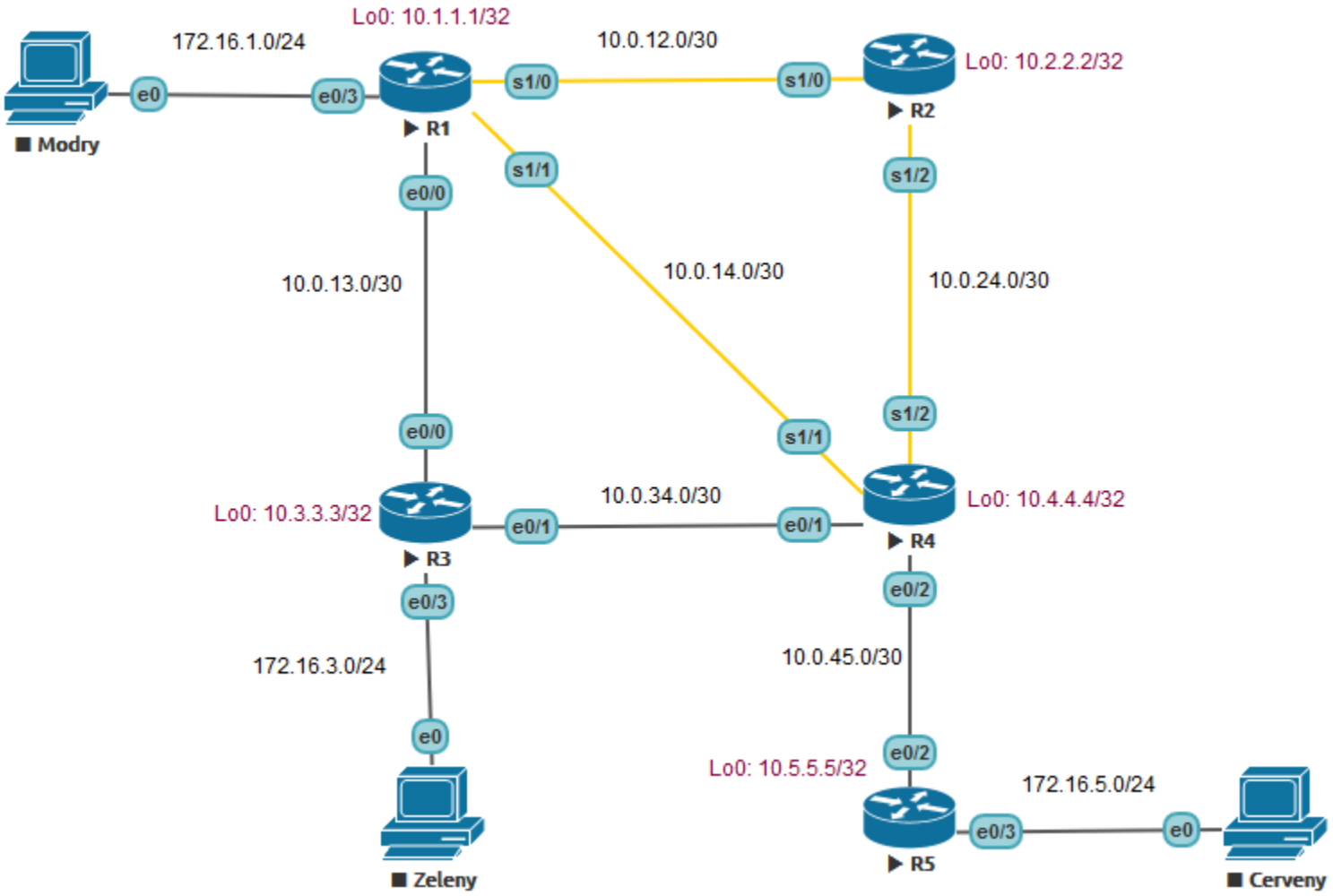


EIGRP: Use-Case ④

- Router *A*:
  - Route via *B* is best with FD(*B*) = 20
  - Route via *C* has RD(*C*) = 10 which is lower than FD hence this route is loopless
  - Route via *D* has RD(*D*) = 30 which is higher than FD hence this route "potentially" could cause topology loop



EIGRP: Network Graph



EIGRP: Topology Database

R1

```
R1#show ip eigrp topology
EIGRP-IPv4 Topology Table for AS(65000)/ID(10.1.1.1)
Codes: P - Passive, A - Active, U - Update, Q - Query,
       r - reply Status, s - sia Status

P 10.0.34.0/30, 1 successors, FD is 307200
        via 10.0.13.2 (307200/281600), Ethernet0/0
        via 10.0.14.2 (2195456/281600), Serial1/1
P 10.3.3.3/32, 1 successors, FD is 409600
        via 10.0.13.2 (409600/128256), Ethernet0/0
P 10.1.1.1/32, 1 successors, FD is 128256
        via Connected, Loopback0
P 172.16.5.0/24, 1 successors, FD is 358400
        via 10.0.13.2 (358400/332800), Ethernet0/0
        via 10.0.14.2 (2221056/307200), Serial1/1
P 10.5.5.5/32, 1 successors, FD is 460800
        via 10.0.13.2 (460800/435200), Ethernet0/0
        via 10.0.14.2 (2323456/409600), Serial1/1
P 10.0.14.0/30, 1 successors, FD is 2169856
        via Connected, Serial1/1
P 172.16.3.0/24, 1 successors, FD is 307200
        via 10.0.13.2 (307200/281600), Ethernet0/0
P 10.4.4.4/32, 1 successors, FD is 435200
        via 10.0.13.2 (435200/409600), Ethernet0/0
        via 10.0.14.2 (2297856/128256), Serial1/1
P 10.0.13.0/30, 1 successors, FD is 281600
        via Connected, Ethernet0/0
P 172.16.1.0/24, 1 successors, FD is 281600
        via Connected, Ethernet0/3
P 10.2.2.2/32, 1 successors, FD is 2297856
        via 10.0.12.2 (2297856/128256), Serial1/0
P 10.0.45.0/30, 1 successors, FD is 332800
        via 10.0.13.2 (332800/307200), Ethernet0/0
        via 10.0.14.2 (2195456/281600), Serial1/1
P 10.0.12.0/30, 1 successors, FD is 2169856
        via Connected, Serial1/0
P 10.0.24.0/30, 1 successors, FD is 2221056
        via 10.0.13.2 (2221056/2195456), Ethernet0/0
        via 10.0.12.2 (2681856/2169856), Serial1/0
        via 10.0.14.2 (2681856/2169856), Serial1/1

R1#
```

R4

```
R4#sh ip eigrp topology
EIGRP-IPv4 Topology Table for AS(65000)/ID(10.4.4.4)
Codes: P - Passive, A - Active, U - Update, Q - Query,
       r - reply Status, s - sia Status

P 10.0.34.0/30, 1 successors, FD is 281600
        via Connected, Ethernet0/1
P 10.3.3.3/32, 1 successors, FD is 409600
        via 10.0.34.1 (409600/128256), Ethernet0/1
P 10.1.1.1/32, 1 successors, FD is 435200
        via 10.0.34.1 (435200/409600), Ethernet0/1
        via 10.0.14.1 (2297856/128256), Serial1/1
P 172.16.5.0/24, 1 successors, FD is 307200
        via 10.0.45.2 (307200/281600), Ethernet0/2
P 10.5.5.5/32, 1 successors, FD is 409600
        via 10.0.45.2 (409600/128256), Ethernet0/2
P 10.0.14.0/30, 1 successors, FD is 2169856
        via Connected, Serial1/1
P 172.16.3.0/24, 1 successors, FD is 307200
        via 10.0.34.1 (307200/281600), Ethernet0/1
P 10.4.4.4/32, 1 successors, FD is 128256
        via Connected, Loopback0
P 10.0.13.0/30, 1 successors, FD is 307200
        via 10.0.34.1 (307200/281600), Ethernet0/1
        via 10.0.14.1 (2195456/281600), Serial1/1
P 172.16.1.0/24, 1 successors, FD is 332800
        via 10.0.34.1 (332800/307200), Ethernet0/1
        via 10.0.14.1 (2195456/281600), Serial1/1
P 10.2.2.2/32, 1 successors, FD is 2297856
        via 10.0.24.1 (2297856/128256), Serial1/2
P 10.0.45.0/30, 1 successors, FD is 281600
        via Connected, Ethernet0/2
P 10.0.12.0/30, 1 successors, FD is 2221056
        via 10.0.34.1 (2221056/2195456), Ethernet0/1
        via 10.0.24.1 (2681856/2169856), Serial1/2
        via 10.0.14.1 (2681856/2169856), Serial1/1
P 10.0.24.0/30, 1 successors, FD is 2169856
        via Connected, Serial1/2

R4#
```

EIGRP: Routing Table

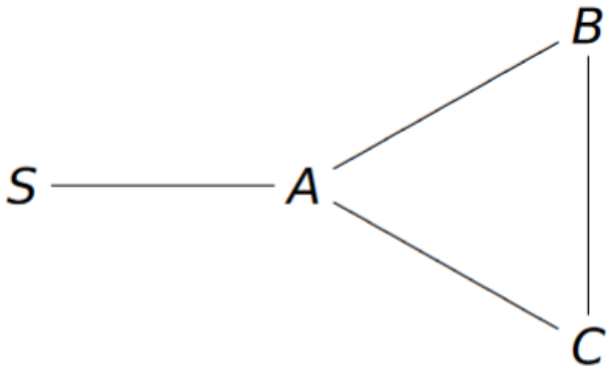
R1

```
Gateway of last resort is not set

      10.0.0.0/8 is variably subnetted, 14 subnets, 2 masks
C        10.0.12.0/30 is directly connected, Serial1/0
L        10.0.12.1/32 is directly connected, Serial1/0
C        10.0.13.0/30 is directly connected, Ethernet0/0
L        10.0.13.1/32 is directly connected, Ethernet0/0
C        10.0.14.0/30 is directly connected, Serial1/1
L        10.0.14.1/32 is directly connected, Serial1/1
D        10.0.24.0/30 [90/2221056] via 10.0.13.2, 00:05:49, Ethernet0/0
D        10.0.34.0/30 [90/307200] via 10.0.13.2, 00:05:49, Ethernet0/0
D        10.0.45.0/30 [90/332800] via 10.0.13.2, 00:05:49, Ethernet0/0
C        10.1.1.1/32 is directly connected, Loopback0
D        10.2.2.2/32 [90/2297856] via 10.0.12.2, 00:05:49, Serial1/0
D        10.3.3.3/32 [90/409600] via 10.0.13.2, 00:05:55, Ethernet0/0
D        10.4.4.4/32 [90/435200] via 10.0.13.2, 00:05:49, Ethernet0/0
D        10.5.5.5/32 [90/460800] via 10.0.13.2, 00:05:49, Ethernet0/0
      172.16.0.0/16 is variably subnetted, 4 subnets, 2 masks
C        172.16.1.0/24 is directly connected, Ethernet0/3
L        172.16.1.1/32 is directly connected, Ethernet0/3
D        172.16.3.0/24 [90/307200] via 10.0.13.2, 00:05:55, Ethernet0/0
D        172.16.5.0/24 [90/358400] via 10.0.13.2, 00:05:49, Ethernet0/0
R1#
```

R4

```
Gateway of last resort is not set

      10.0.0.0/8 is variably subnetted, 15 subnets, 2 masks
D        10.0.12.0/30 [90/2221056] via 10.0.34.1, 00:05:41, Ethernet0/1
D        10.0.13.0/30 [90/307200] via 10.0.34.1, 00:05:41, Ethernet0/1
C        10.0.14.0/30 is directly connected, Serial1/1
L        10.0.14.2/32 is directly connected, Serial1/1
C        10.0.24.0/30 is directly connected, Serial1/2
L        10.0.24.2/32 is directly connected, Serial1/2
C        10.0.34.0/30 is directly connected, Ethernet0/1
L        10.0.34.2/32 is directly connected, Ethernet0/1
C        10.0.45.0/30 is directly connected, Ethernet0/2
L        10.0.45.1/32 is directly connected, Ethernet0/2
D        10.1.1.1/32 [90/435200] via 10.0.34.1, 00:05:41, Ethernet0/1
D        10.2.2.2/32 [90/2297856] via 10.0.24.1, 00:05:41, Serial1/2
D        10.3.3.3/32 [90/409600] via 10.0.34.1, 00:05:46, Ethernet0/1
C        10.4.4.4/32 is directly connected, Loopback0
D        10.5.5.5/32 [90/409600] via 10.0.45.2, 00:05:43, Ethernet0/2
      172.16.0.0/24 is subnetted, 3 subnets
D        172.16.1.0 [90/332800] via 10.0.34.1, 00:05:41, Ethernet0/1
D        172.16.3.0 [90/307200] via 10.0.34.1, 00:05:46, Ethernet0/1
D        172.16.5.0 [90/307200] via 10.0.45.2, 00:05:43, Ethernet0/2
R4#
```

Babel



- RFC 6126
- Messages
  - operates over UDP on port 6696
- Babel is a modular protocol
  - a robust and mostly transient-free routing core
  - switchable metric computation
- Neighbor Detection
- Distributed Bellman-Ford for the best route selection
  - Source node feasibility condition just as EIGRP

$$D_A(S) = (s_A, m_A), FD_B(S) = (s_B, m_B):$$
$$D_A(S) < FD_B(S) \leftrightarrow (s_A = s_B \wedge m_A < m_B) \vee s_A > s_B$$

Babel: Metric

- Babel is metric-agnostic
  - a metric MUST be strictly monotonic

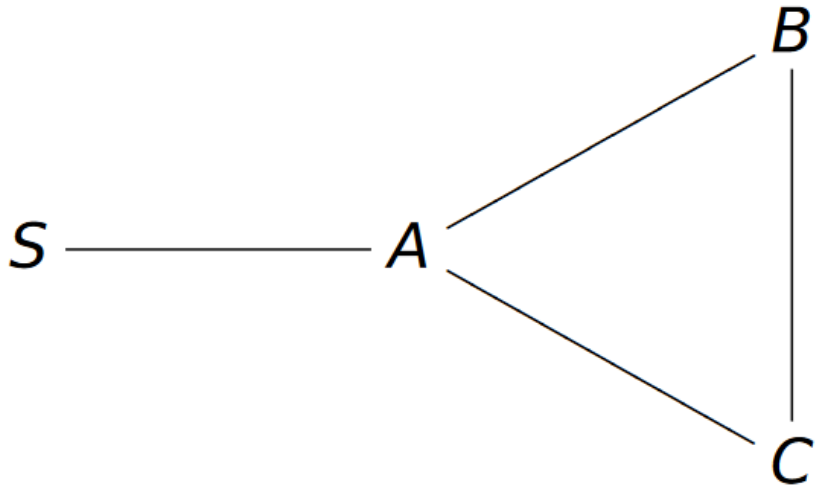
$$m < c + m$$

  - a metric SHOULD be isotonic

$$m \leq \acute{m} : c + m \leq c + \acute{m}$$

- Metric types
  - K-out-of-J reliability for wired networks
  - ETX for wireless networks, where link cost varies in time, and it is determined based on succeful hello receptions and transmissions
  - Z3 metric refines ETX takes into account radio interface
  - RTT based metric reflecting delay

Babel: Towards Distrib. Bellman-Ford ①

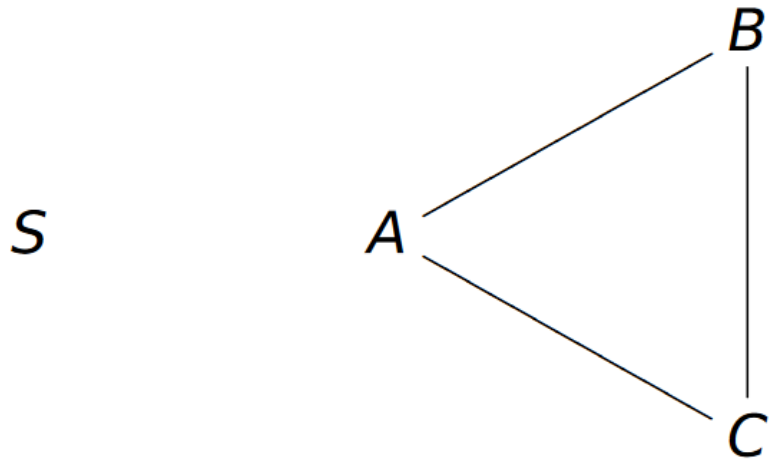


S	0	0	0	0
A	∞	1, nh = S	1, nh = S	1, nh = S
B	∞	∞	2, nh = A	2, nh = A
C	∞	∞	2, nh = A	2, nh = A

Converges in $O(\Delta)$.

- *Solution* = Periodic updates + Unsolicited updates

Babel: Towards Distrib. Bellman-Ford ②



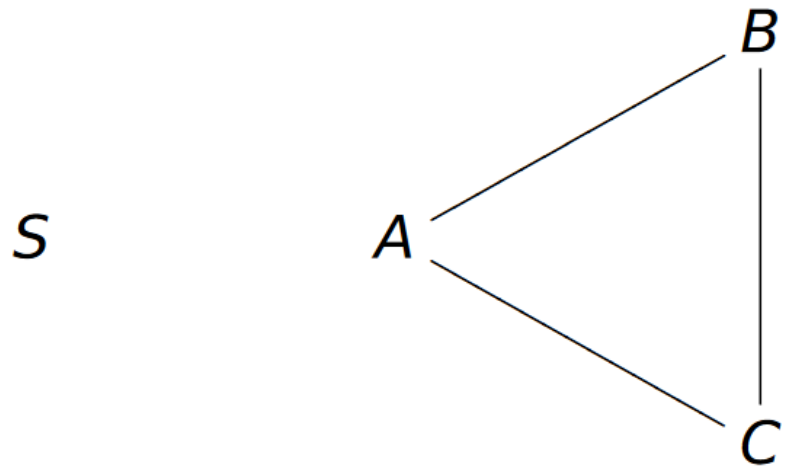
A	1, nh $= S$	3, nh $= B$	3, nh $= B$	3, nh $= B$
B	2, nh $= A$	2, nh $= A$	3, nh $= C$	3, nh $= C$
C	2, nh $= A$	2, nh $= A$	2, nh $= A$	4, nh $= A$

Converges in $O(\infty)$. (RIP: $\infty = 16$.)

Before convergence, there is a routing loop.
« Good news travel fast, bad news travel forever. »

- *Solution* = Poison Reverse + Split Horizon

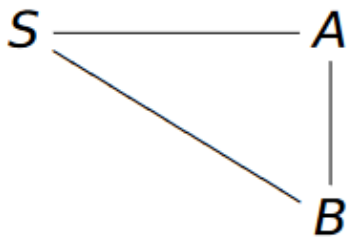
Babel: Towards Distrib. Bellman-Ford ③



A	1, fd = 1	∞, fd = 1	∞, fd = 1	∞, fd = 1
B	2, fd = 2	2, fd = 2	∞, fd = 2	∞, fd = 2
C	2, fd = 2	2, fd = 2	∞, fd = 2	∞, fd = 2

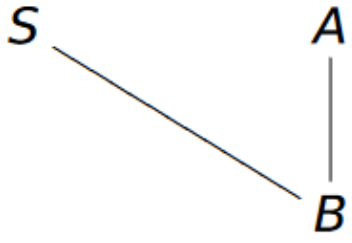
Converges in $O(\Delta)$.

SNC Causes Starvation



$\mathrm{d}(A) = 1, \mathrm{fd}(A) = 1$

$\mathrm{d}(B) = 1, \mathrm{fd}(B) = 1$



$\mathrm{fd}(A) = 1$

$\mathrm{d}(B) = 1$

S	$(1, 0)$	$(2, 0)$	$(2, 0)$
A	$\infty, \mathrm{fd} = (1, 1)$	$\infty, \mathrm{fd} = (1, 1)$	$(2, 2), \mathrm{fd} = (2, 2)$
B	$(1, 1), \mathrm{fd} = (1, 1)$	$(2, 1), \mathrm{fd} = (2, 1)$	$(2, 1), \mathrm{fd} = (2, 1)$

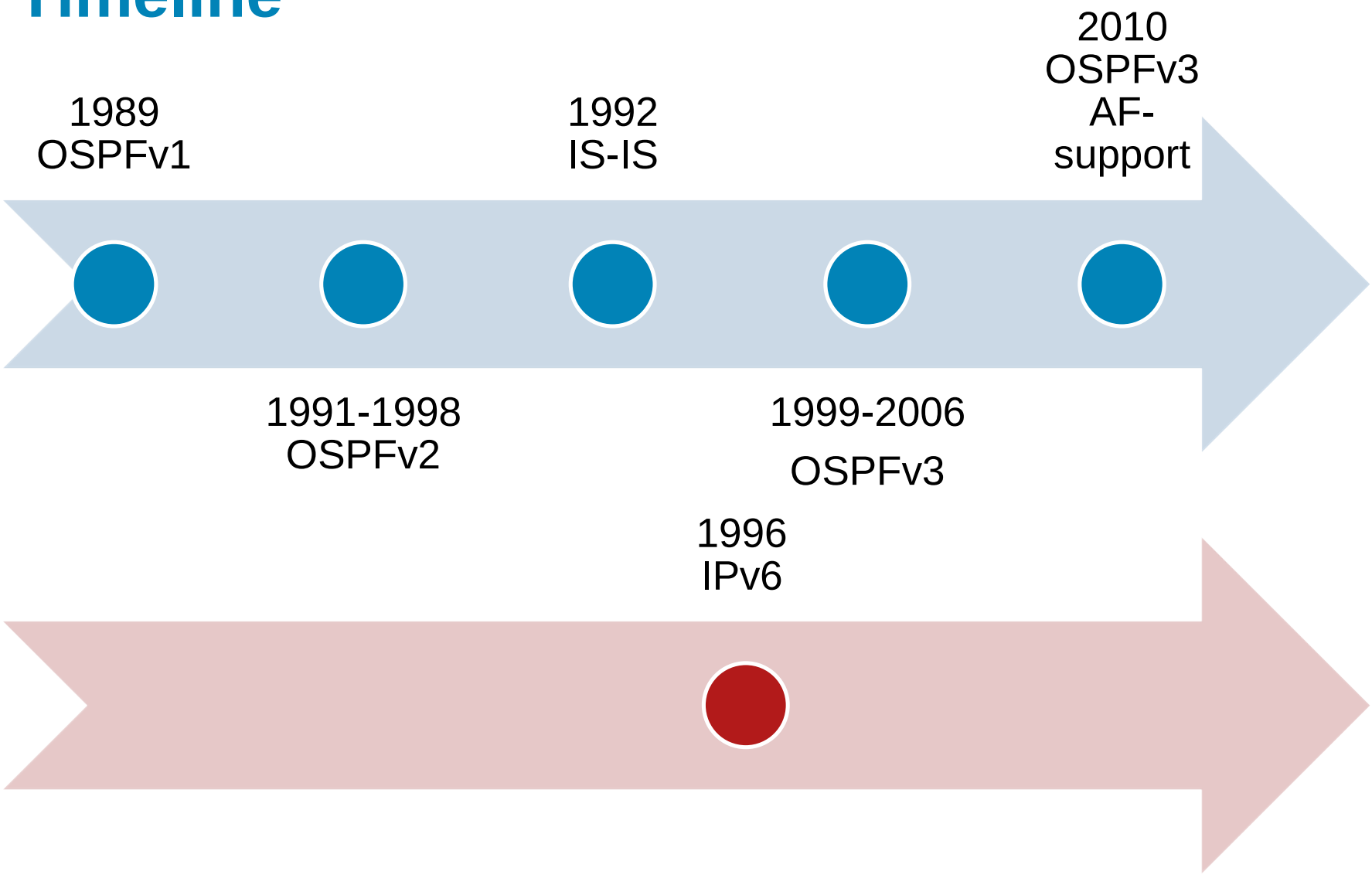
- *Solution*
  - EIGRP resets the network
  - Babel employs sequence numbers

Link-state Protocol Evolution

Q. What did the OSPF router say to the other OSPF router?
A. Hello. Hello. Hello. Hello. Hello. Hello. Hello. Hello.

Anonymous Jokers

Timeline


1989
OSPFv1
1991-1998
OSPFv2
1992
IS-IS
1996
IPv6
1999-2006
OSPFv3
2010
OSPFv3
AF-
support

OSPF

- Versions
  - OSPFv2 for IPv4 in RFC 2328
  - OSPFv3 for IPv6 in RFC 5340 (+ IPv4 in RFC 5838)
- Messages
  - Encapsulated directly into IP
  - No transport layer reliability
- Neighbor Detection
  - periodic Hellos every 5 seconds
- Routing Hierarchy
  - Faster Convergence and less consumption of network resources
  - Division into areas to achieve star topologies
  - Separate internal and external routes

OSPF: Link-State Database

- Distributed, replicated database model
  - describes complete routing topology
  - identical for all the routers
- Link-state advertisements
  - Carry local piece of routing topology
  - Distribution of LSAs using reliable flooding
- LSA has lifetime
  - Max. 60 minutes
  - Refreshed every 30 minutes

OSPF: Metric

- OSPF metric is called **cost** and lower cost is better/preferred

$$Cost = \frac{100\ \text{Mbps}}{Bandwidth}$$

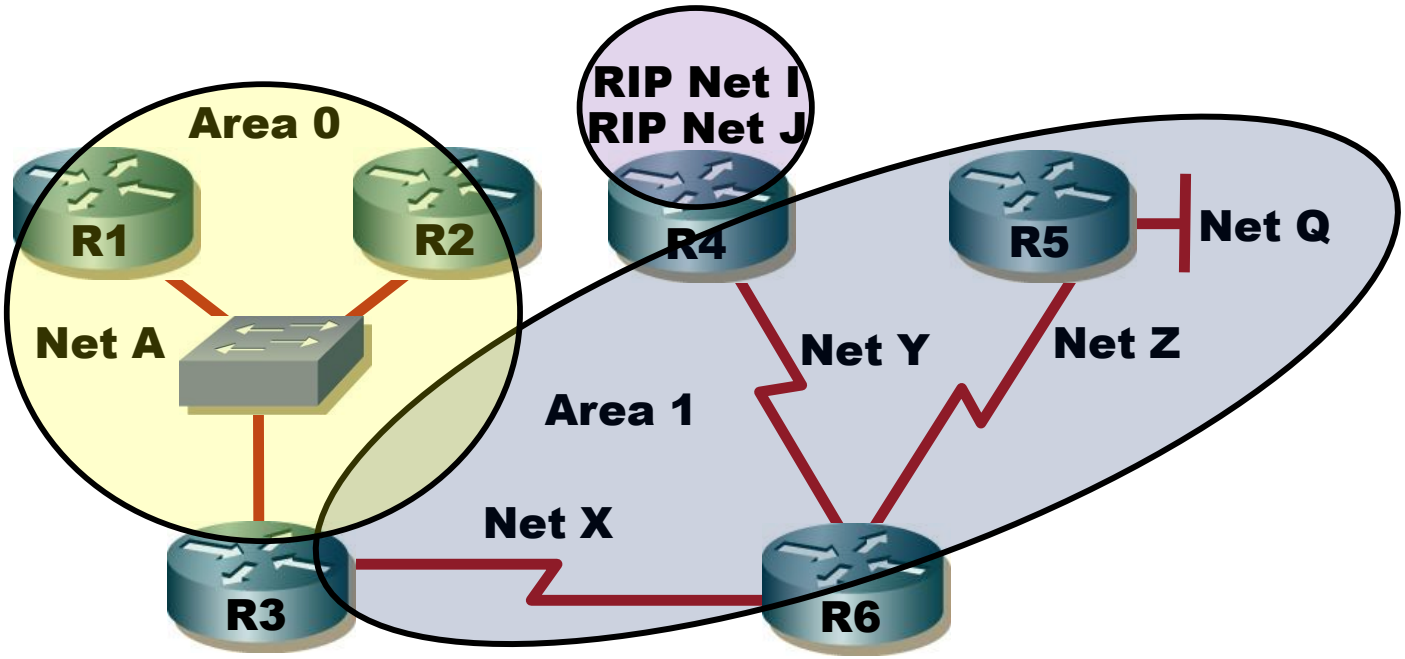
Interface Type	10^8/bps = Cost
Fast Ethernet and faster	10^8/100,000,000 bps = 1
Ethernet	10^8/10,000,000 bps = 10
E1	10^8/2,048,000 bps = 48
T1	10^8/1,544,000 bps = 64
128 kbps	10^8/128,000 bps = 781
64 kbps	10^8/64,000 bps = 1562
56 kbps	10^8/56,000 bps = 1785

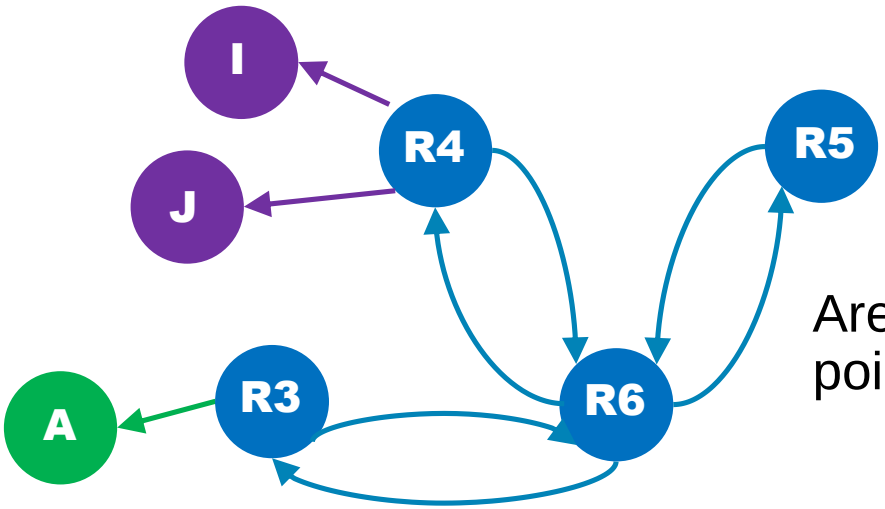
OSPF: LSAs

LSA Type	Description
1	Router LSAs
2	Network LSAs
3 or 4	Summary LSAs
5	Autonomous System External LSAs
6	Multicast OSPF LSAs
7	Defined for Not-So-Stubby Areas
8	External Attributes LSA for Border Gateway Protocol (BGP)
9, 10, 11	Opaque LSAs

	LSA Function Code	LSA Type
Router-LSA	1	0x2001
Network-LSA	2	0x2002
Inter-Area-Prefix-LSA	3	0x2003
Inter-Area-Router-LSA	4	0x2004
AS-External-LSA	5	0x4005
Group-Membership-LSA	6	0x2006
Type-7-LSA	7	0x2007
Link-LSA	8	0x2008
Intra-Area-Prefix-LSA	9	0x2009

OSPF: Modelling Network Area 1





Area 1 routers point of view

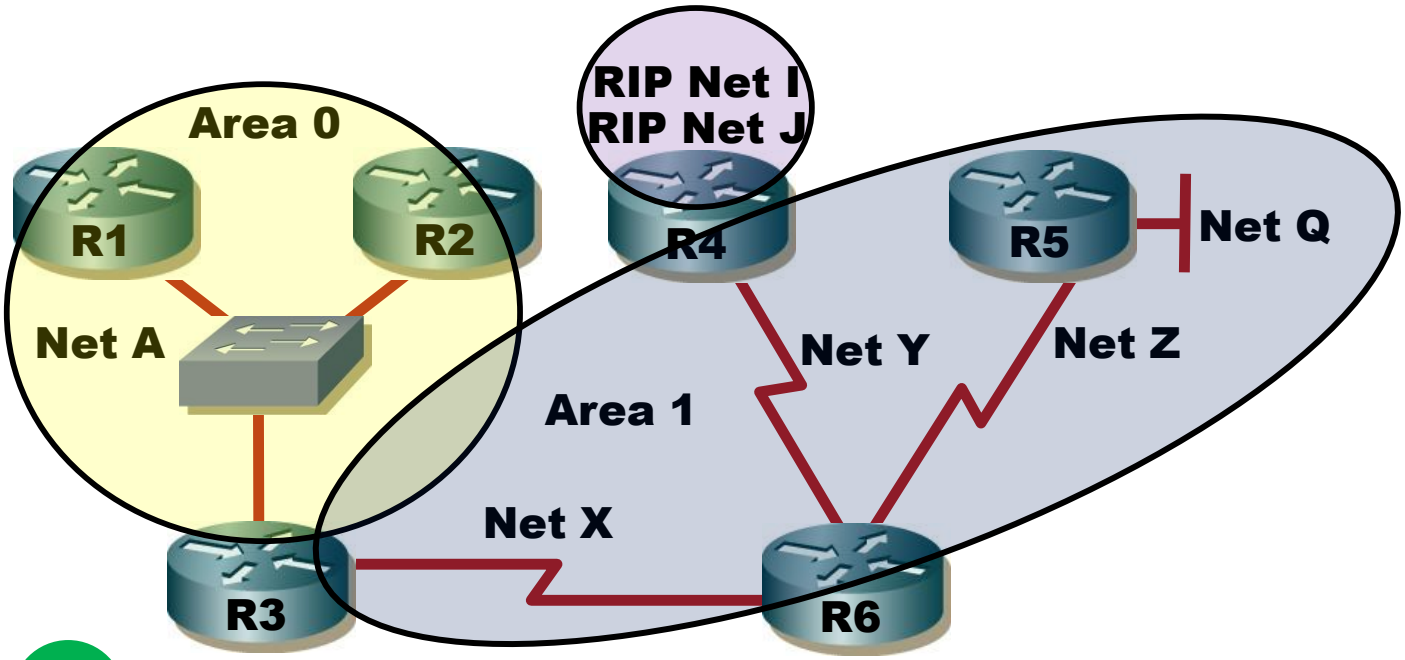
=LSA1

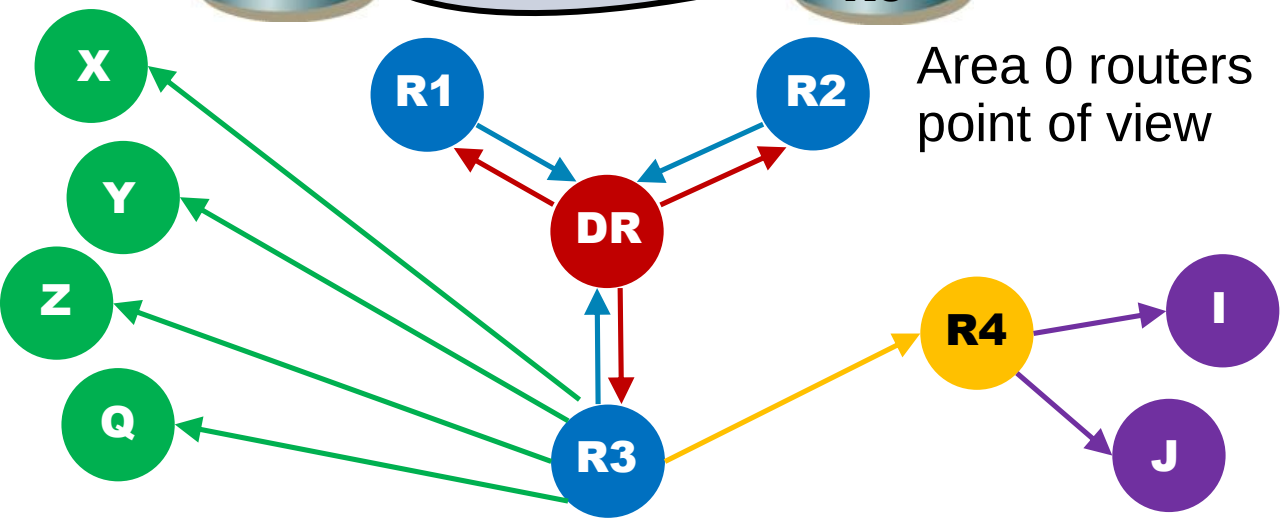
=LSA2

=LSA3

=LSA5

OSPF: Modelling Network Area 0





Area 0 routers point of view

=LSA1

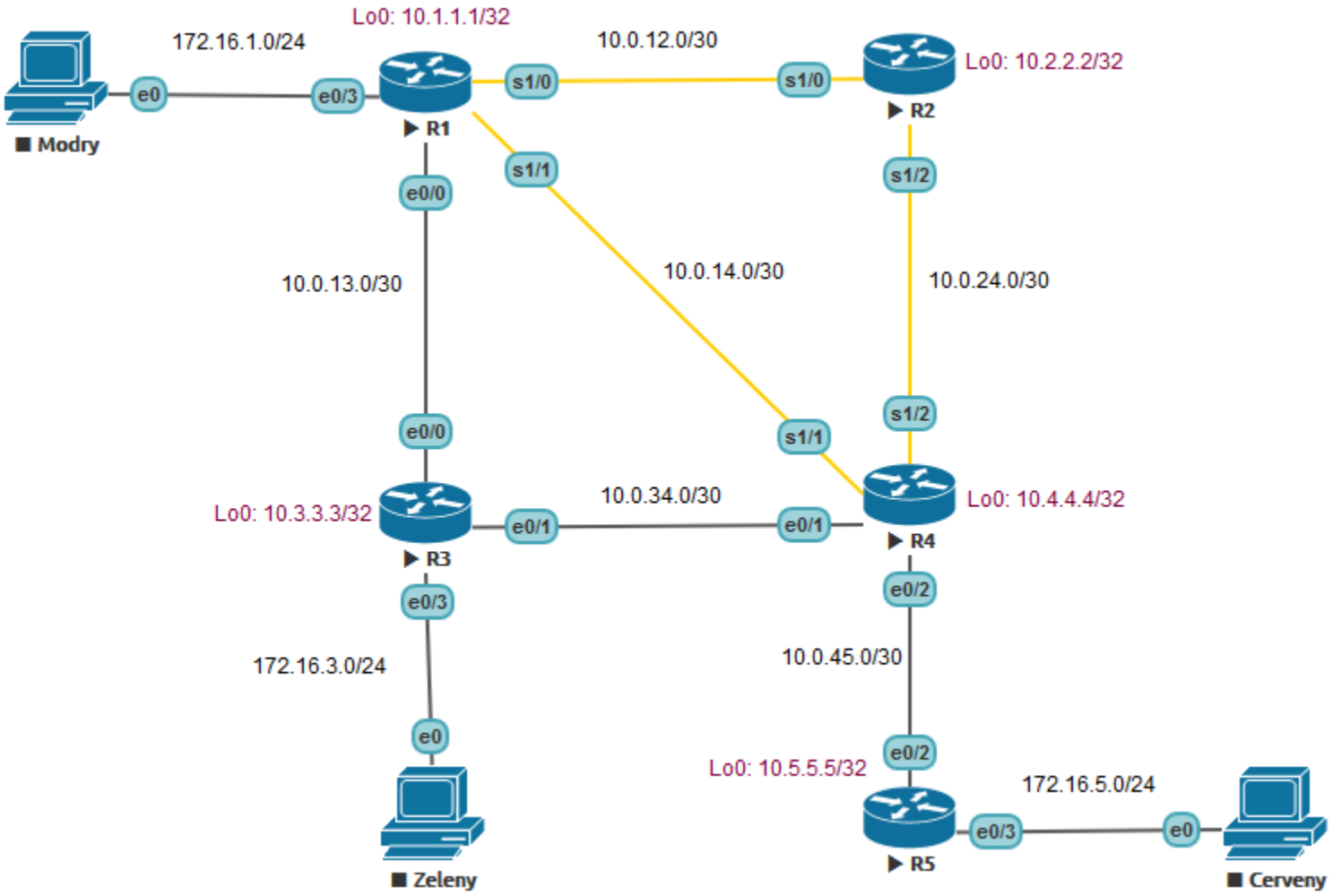
=LSA2

=LSA3

=LSA4

=LSA5

OSPF: Network Graph



OSPF: Link-State Database

R1

```
R1#show ip ospf database

            OSPF Router with ID (10.1.1.1) (Process ID 1)

                Router Link States (Area 0)

Link ID         ADV Router      Age         Seq#       Checksum Link count
10.1.1.1        10.1.1.1        26          0x80000003 0x009199 7
10.2.2.2        10.2.2.2        62          0x80000001 0x002D07 5
10.3.3.3        10.3.3.3        22          0x80000003 0x001B42 4
10.4.4.4        10.4.4.4        25          0x80000005 0x004ED1 7
10.5.5.5        10.5.5.5        59          0x80000001 0x00661A 3

                Net Link States (Area 0)

Link ID         ADV Router      Age         Seq#       Checksum
10.0.13.2       10.3.3.3        27          0x80000001 0x00FAF5
10.0.34.2       10.4.4.4        25          0x80000001 0x006768
10.0.45.1       10.4.4.4        58          0x80000001 0x004679
R1#
```

R4

```
R4#show ip ospf database

            OSPF Router with ID (10.4.4.4) (Process ID 1)

                Router Link States (Area 0)

Link ID         ADV Router      Age         Seq#       Checksum Link count
10.1.1.1        10.1.1.1        81          0x80000003 0x009199 7
10.2.2.2        10.2.2.2        115         0x80000001 0x002D07 5
10.3.3.3        10.3.3.3        76          0x80000003 0x001B42 4
10.4.4.4        10.4.4.4        77          0x80000005 0x004ED1 7
10.5.5.5        10.5.5.5        111         0x80000001 0x00661A 3

                Net Link States (Area 0)

Link ID         ADV Router      Age         Seq#       Checksum
10.0.13.2       10.3.3.3        82          0x80000001 0x00FAF5
10.0.34.2       10.4.4.4        77          0x80000001 0x006768
10.0.45.1       10.4.4.4        111         0x80000001 0x004679
R4#
```

OSPF: Routing Tables

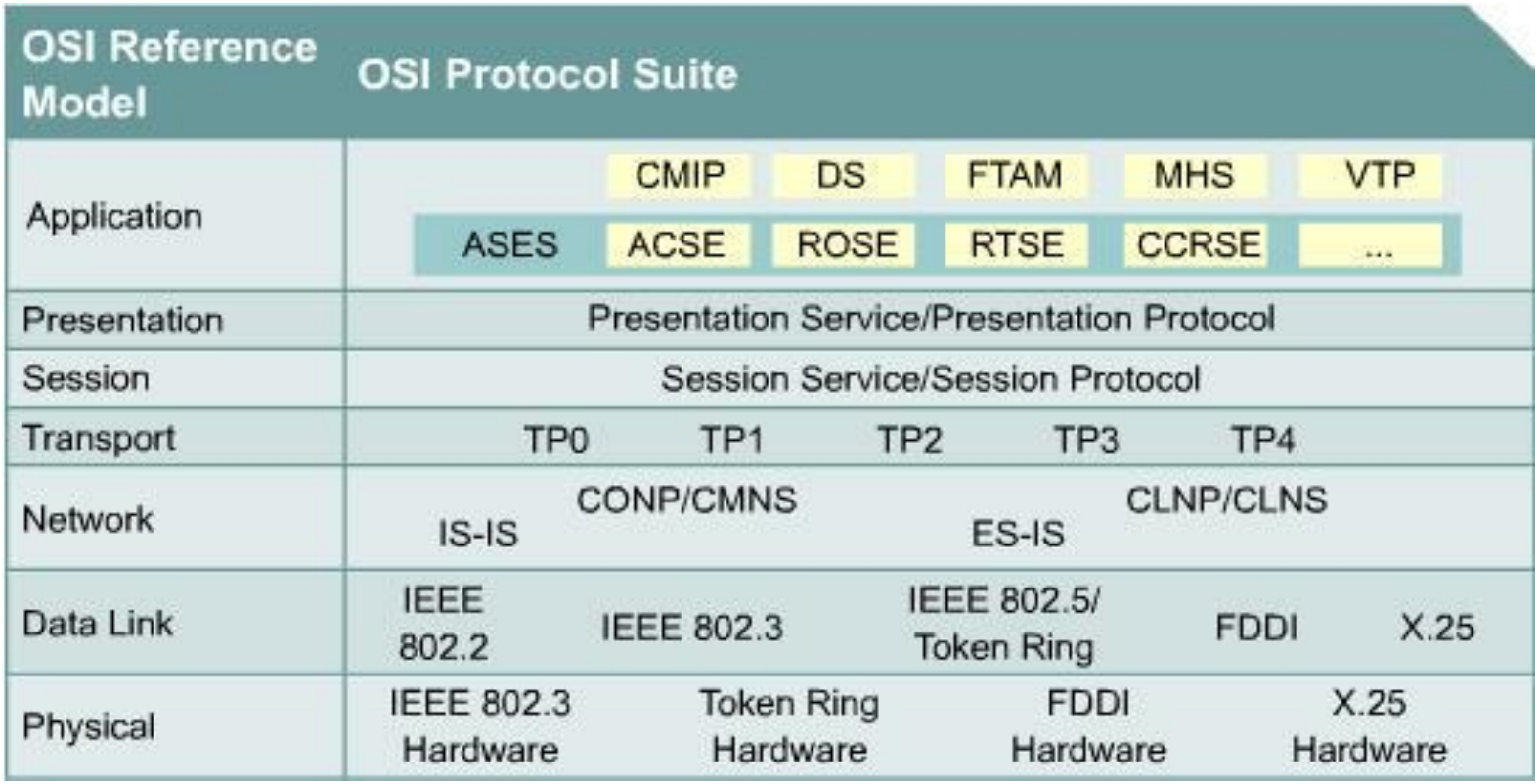
R1

```
Gateway of last resort is not set

      10.0.0.0/8 is variably subnetted, 14 subnets, 2 masks
C        10.0.12.0/30 is directly connected, Serial1/0
L        10.0.12.1/32 is directly connected, Serial1/0
C        10.0.13.0/30 is directly connected, Ethernet0/0
L        10.0.13.1/32 is directly connected, Ethernet0/0
C        10.0.14.0/30 is directly connected, Serial1/1
L        10.0.14.1/32 is directly connected, Serial1/1
O        10.0.24.0/30 [110/84] via 10.0.13.2, 00:04:05, Ethernet0/0
O        10.0.34.0/30 [110/20] via 10.0.13.2, 00:04:05, Ethernet0/0
O        10.0.45.0/30 [110/30] via 10.0.13.2, 00:04:05, Ethernet0/0
C        10.1.1.1/32 is directly connected, Loopback0
O        10.2.2.2/32 [110/65] via 10.0.12.2, 00:04:49, Serial1/0
O        10.3.3.3/32 [110/11] via 10.0.13.2, 00:04:15, Ethernet0/0
O        10.4.4.4/32 [110/21] via 10.0.13.2, 00:04:05, Ethernet0/0
O        10.5.5.5/32 [110/31] via 10.0.13.2, 00:04:05, Ethernet0/0
      172.16.0.0/16 is variably subnetted, 4 subnets, 2 masks
C        172.16.1.0/24 is directly connected, Ethernet0/3
L        172.16.1.1/32 is directly connected, Ethernet0/3
O        172.16.3.0/24 [110/20] via 10.0.13.2, 00:04:15, Ethernet0/0
O        172.16.5.0/24 [110/40] via 10.0.13.2, 00:04:05, Ethernet0/0
R1#
```

R4

```
Gateway of last resort is not set

      10.0.0.0/8 is variably subnetted, 15 subnets, 2 masks
O        10.0.12.0/30 [110/84] via 10.0.34.1, 00:04:25, Ethernet0/1
O        10.0.13.0/30 [110/20] via 10.0.34.1, 00:04:25, Ethernet0/1
C        10.0.14.0/30 is directly connected, Serial1/1
L        10.0.14.2/32 is directly connected, Serial1/1
C        10.0.24.0/30 is directly connected, Serial1/2
L        10.0.24.2/32 is directly connected, Serial1/2
C        10.0.34.0/30 is directly connected, Ethernet0/1
L        10.0.34.2/32 is directly connected, Ethernet0/1
C        10.0.45.0/30 is directly connected, Ethernet0/2
L        10.0.45.1/32 is directly connected, Ethernet0/2
O        10.1.1.1/32 [110/21] via 10.0.34.1, 00:04:25, Ethernet0/1
O        10.2.2.2/32 [110/65] via 10.0.24.1, 00:05:09, Serial1/2
O        10.3.3.3/32 [110/11] via 10.0.34.1, 00:04:25, Ethernet0/1
C        10.4.4.4/32 is directly connected, Loopback0
O        10.5.5.5/32 [110/11] via 10.0.45.2, 00:04:59, Ethernet0/2
      172.16.0.0/24 is subnetted, 3 subnets
O        172.16.1.0 [110/30] via 10.0.34.1, 00:04:25, Ethernet0/1
O        172.16.3.0 [110/20] via 10.0.34.1, 00:04:25, Ethernet0/1
O        172.16.5.0 [110/20] via 10.0.45.2, 00:04:59, Ethernet0/2
R4#
```

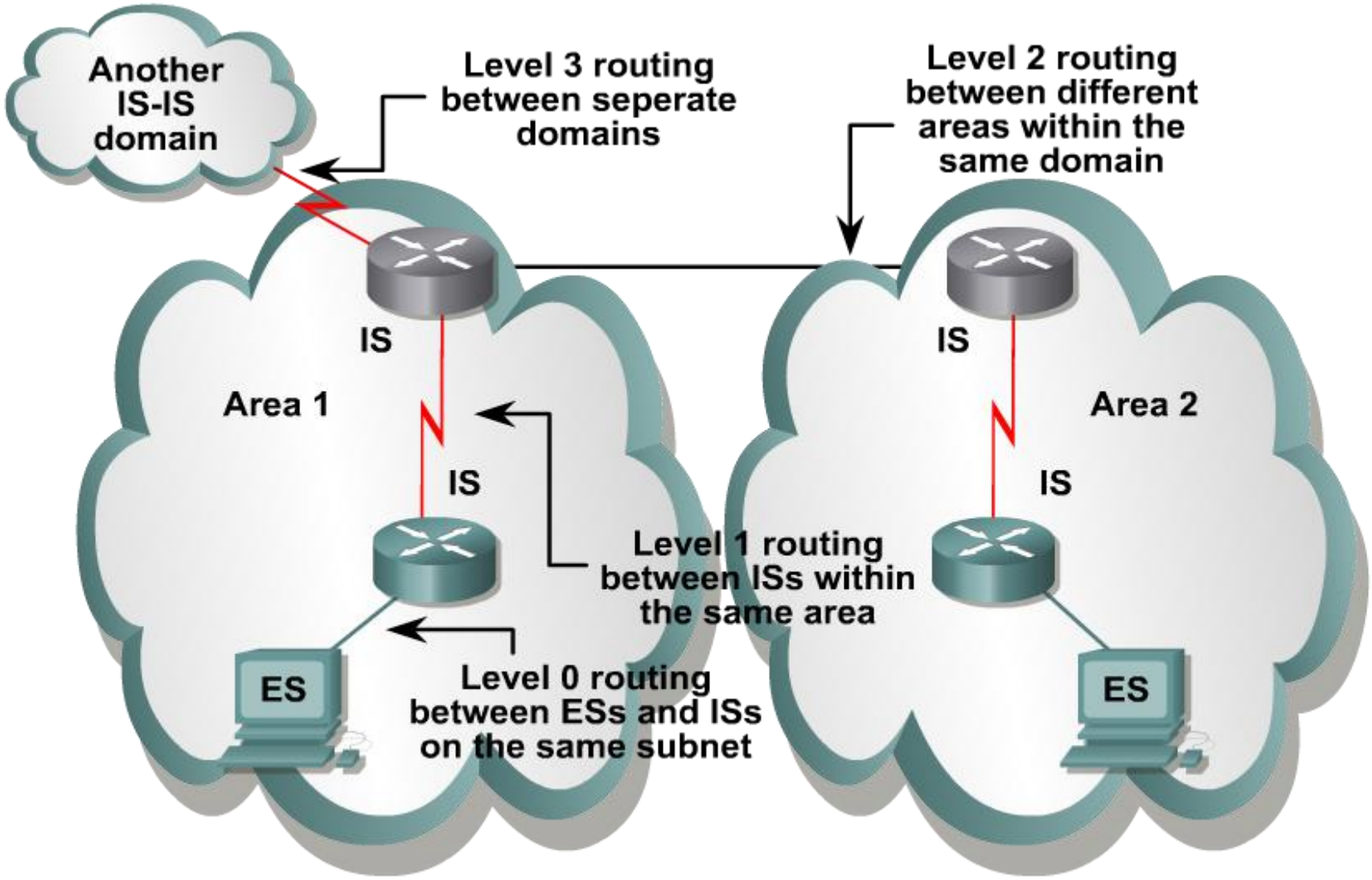
OSI Protocols



OSI Routing

- Multiple routing protocols were proposed for OSI
  - ES-IS: routing between end-station and its gateway (Level 0)
  - IS-IS: routing between routers in one AS a.k.a. domain in ISO terminology (Level 1 and Level 2)
  - IDRP: routing between domains (analogous to BGP) (Level 3)
- Properties of IS-IS turn out to be sophisticated and flexible
  - IS-IS was proposed and functional before OSPF, OSPF started as just a lite version of IS-IS
  - During migration from OSI to IP it was suitable to have routing protocol capable of using both stacks
  - RFC 1195 integrated extension to cooperate with IP without redefining basic structure of protocol

OSI: Routing Levels



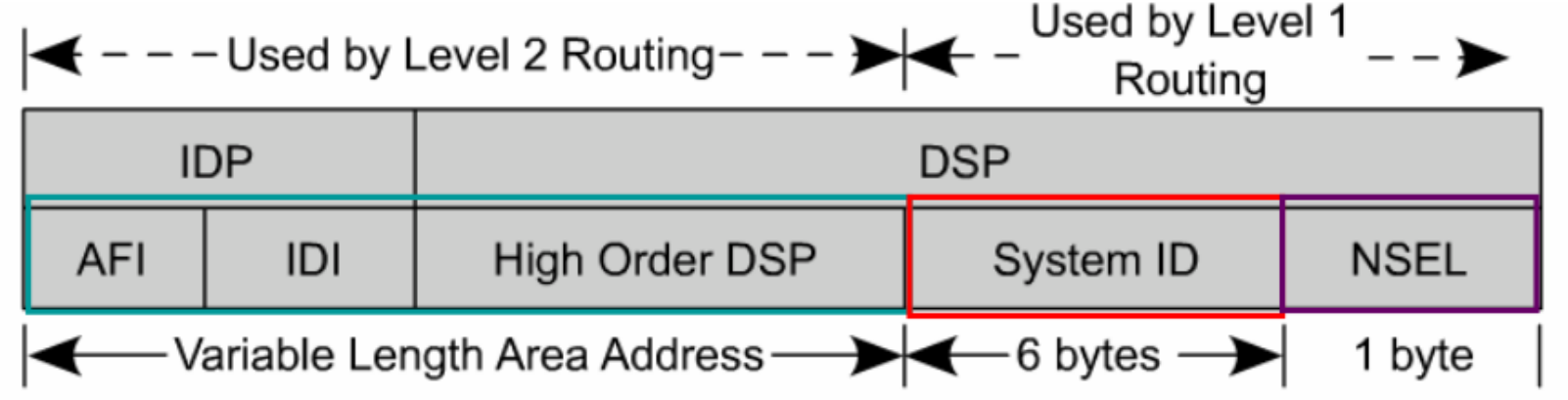
The biggest IPv4/IPv6 vs CLNS Difference

- L3 address
  - IPv4/IPv6 address identifies interface
  - CLNS address identifies node

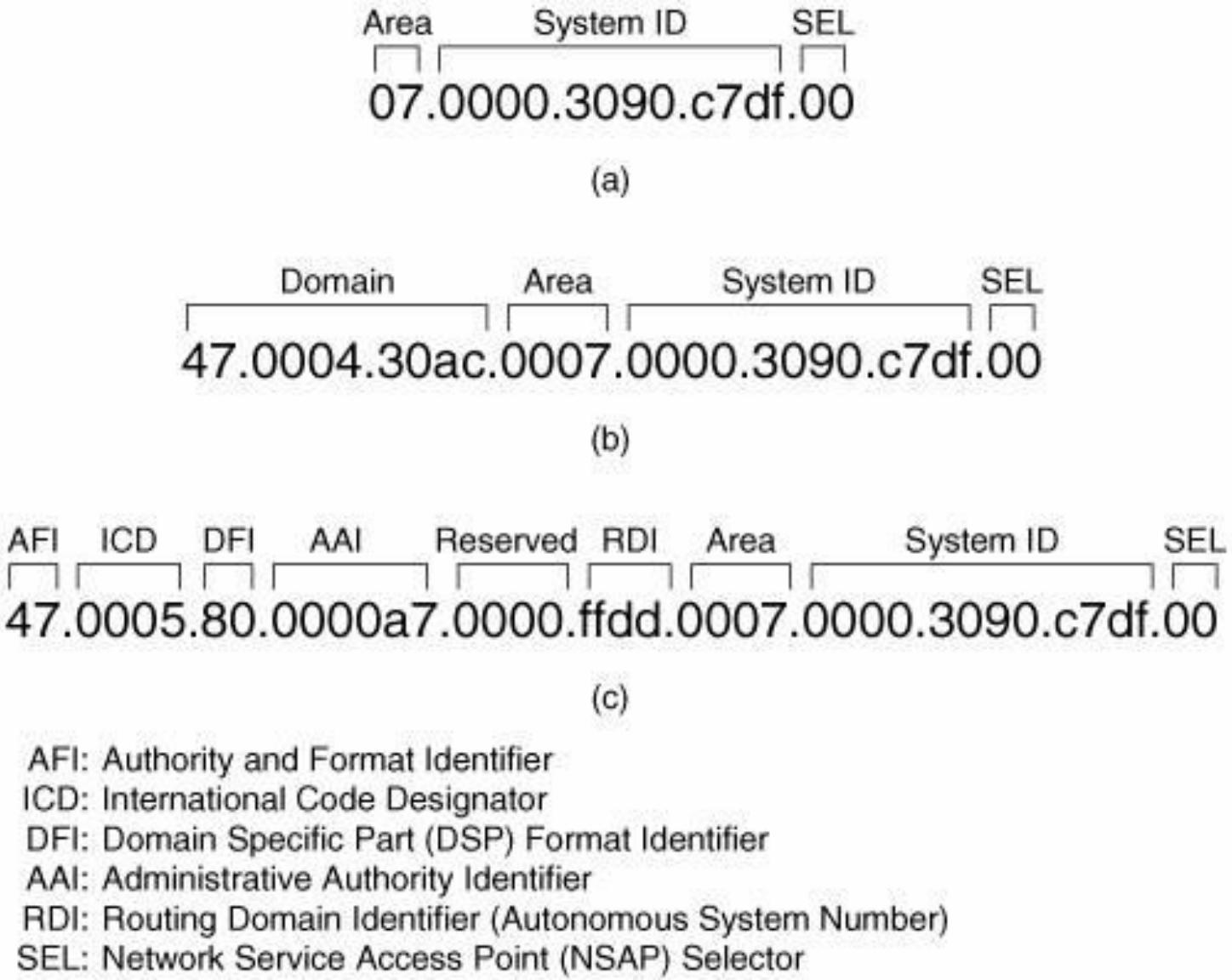
- *CLNS node has L3 address as a unit not as for each interface!!!*
- *Which has huge implications!*

Network Service Access Point

- Node address in OSI networks contains domain number, area number, node identifier and particular service on it
- NSAP/NET addresses should be read from right to left
  49.0001.1234.5678.9012.00
  - The most right byte: **NSEL**
  - The next 6 bytes: **System ID**
  - The remaining bytes except the last one: **HO-DSP, IDI** – their length and semantics is specified by AFI
  - The most left byte: **AFI**, for private domains reserved 49

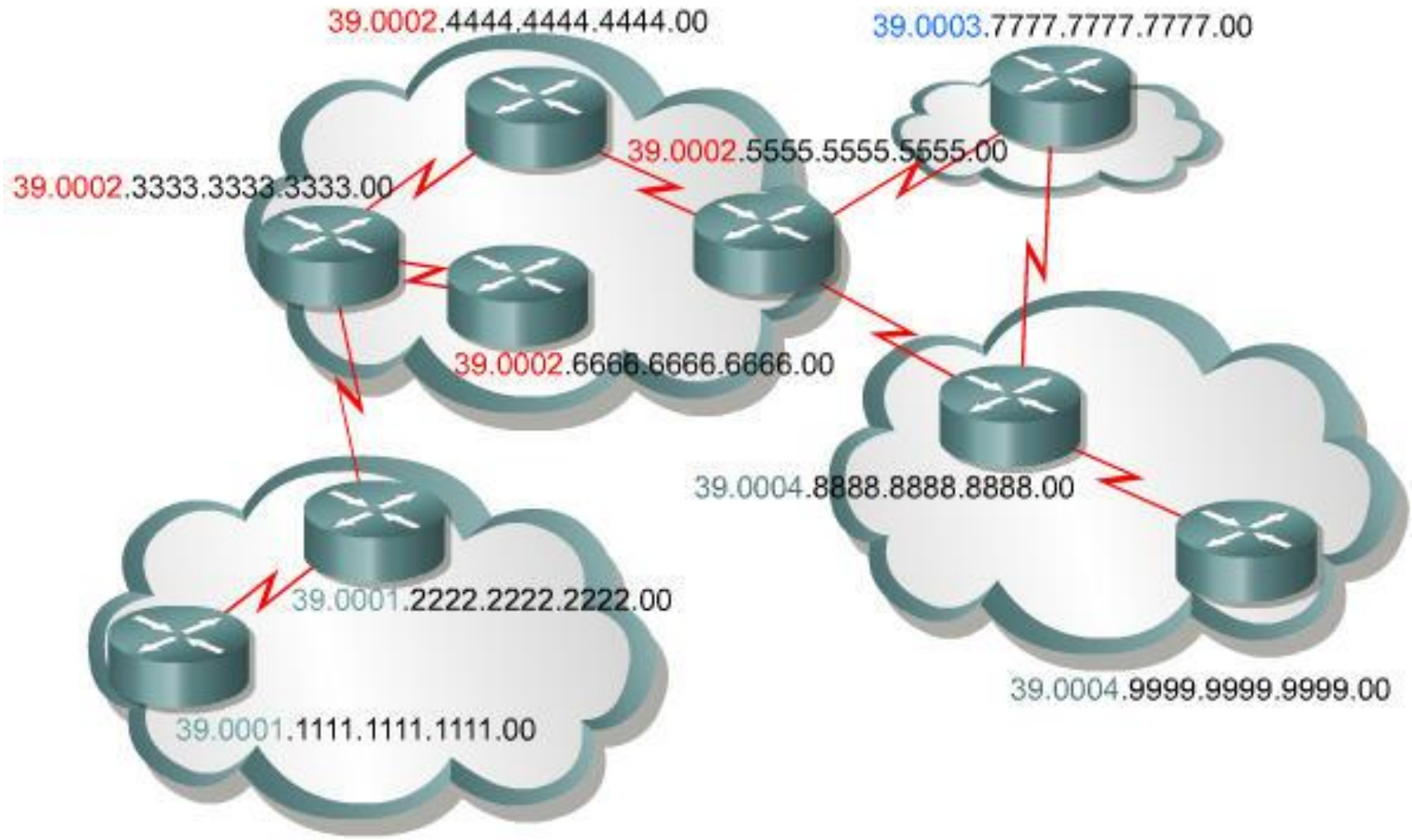


Identifier Examples



http://flylib.com/books/2/296/1/html/2/images/10fig05.jpg

Identifier Assignment



Node Interface

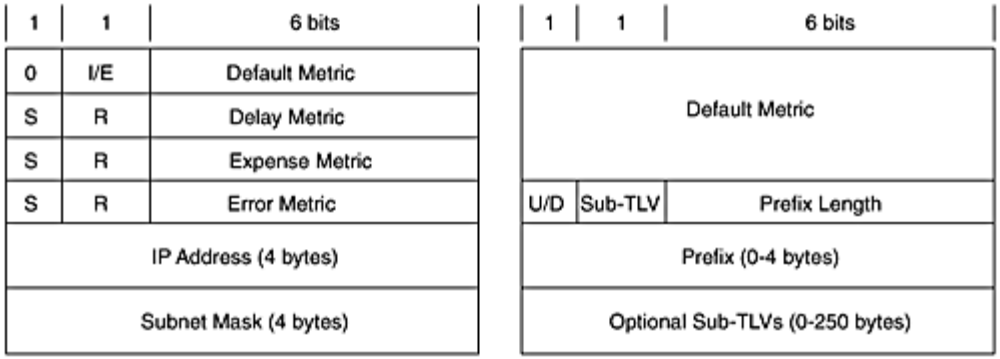
- Each and every node interface is identified by **SubNetwork Point of Attachment (SNPA)**
  - L2 identity of interface
  - Ethernet: MAC address
  - Frame Relay, ATM, X.25: DLCI
  - HDLC and PPP
- Router mark each interface with **Circuit ID** for internal purposes
  - 1 B long number (some IS-IS extensions uses larger space)
  - Assigned by system itself automatically and it CAN NOT be changed via any configuration
  - Multiaccess segment has number that is composed of System ID of Designated IS (analogy of OSPF DR) and its Circuit ID for this segment

IS-IS

- *IS-IS is link-state protocol just like je OSPF, except it is completely different than OSPF*
- Messages
  - IS-IS was originally designed for OSI networks
  - Later it was integrated with IPv4 and IPv6 support
  - IS-IS is encapsulated directly into L2 frames
  - Current IS-IS implementation supporting multiple address families are called Integrated IS-IS (classless, multi-af)
- Neighbor discovery
  - Periodic Hellos just as OSPF
- Simultaneous routing tables with different metrics

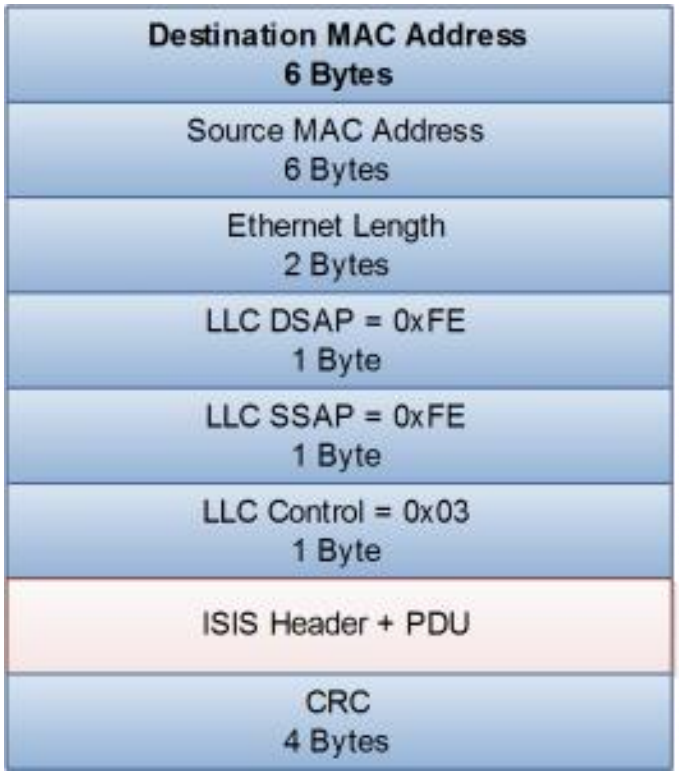
IS-IS: Metric

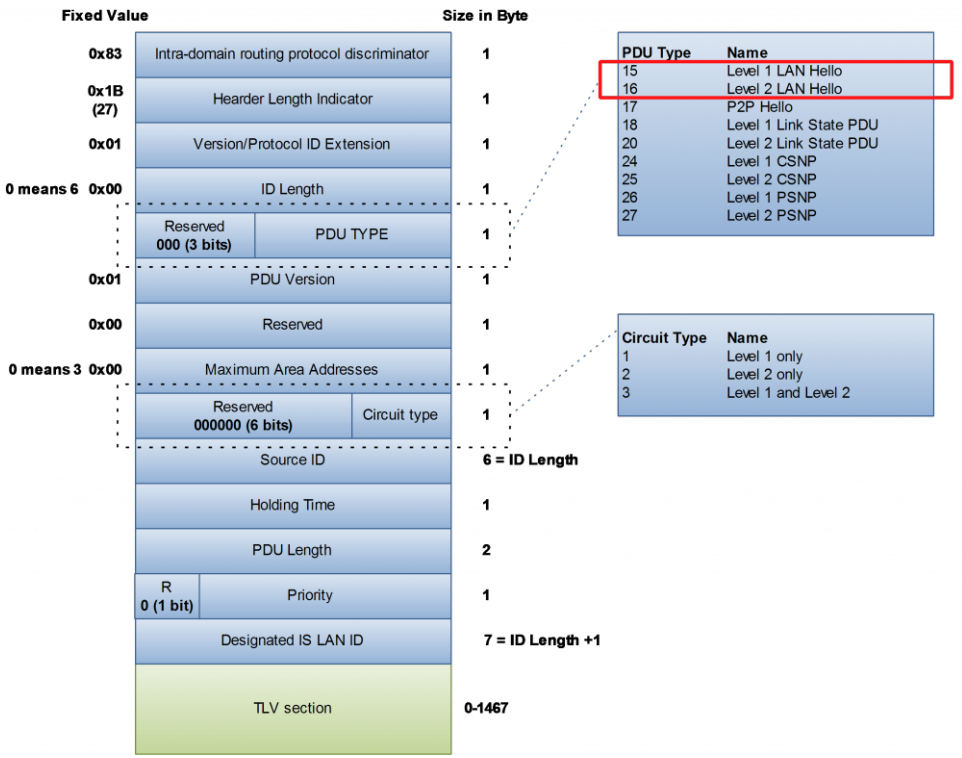
- IS-IS defines 4 different types of metric
  - Default
  - Expense (financial costs for data transfer across the link)
  - Delay
  - Error (error rate on the link)



http://flylib.com/books/2/788/1/html/2/images/1578702208/graphics/05fig10.gif

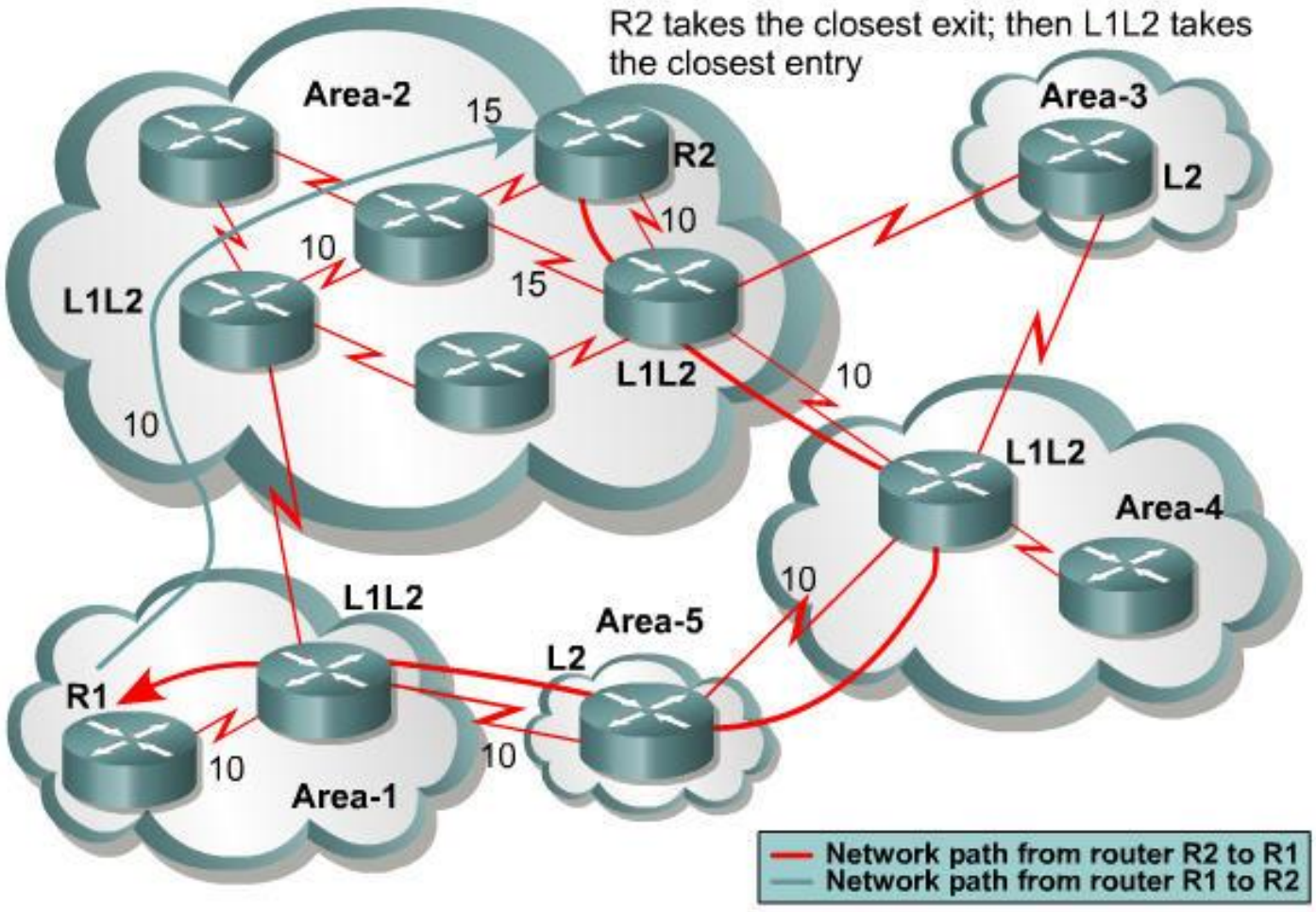
IS-IS: Message Encapsulation



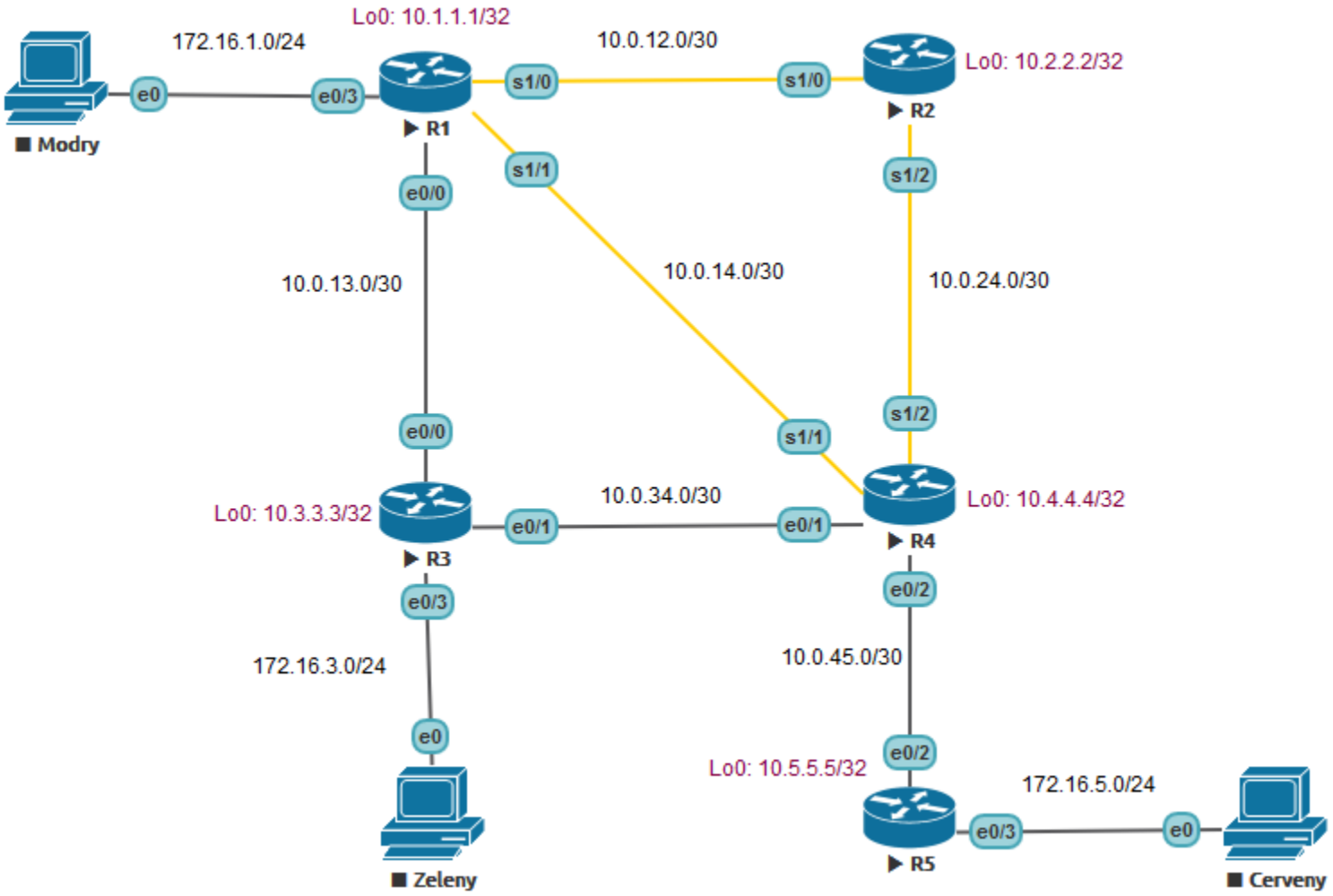


Name	Dst MAC	Description
AllL1ISs	01-80-C2-00-00-14	The multi-destination address “All L1 Intermediate Systems”
AllL2ISs	01-80-C2-00-00-15	The multi-destination address “All L2 Intermediate Systems”
AllIS	09-00-2B-00-00-05	The multi-destination address “All Intermediate Systems” used by ISO

Threat of Suboptimal Routing



IS-IS: Network Graph



IS-IS: Link-State Database

R1

```
R1#show isis database

IS-IS Level-1 Link State Database:
LSPID                 LSP Seq Num  LSP Checksum  LSP Holdtime      ATT/P/OL
R1.00-00            * 0x0000000A   0xF254        904               0/0/0
R2.00-00              0x0000026D   0x61A6        895               0/0/0
R2.01-00              0x00000133   0x3E64        0 (883)           0/0/0
R3.00-00              0x00000005   0x735D        913               0/0/0
R3.01-00              0x00000001   0xECE7        898               0/0/0
R4.00-00              0x0000000A   0x6F8D        913               0/0/0
R4.01-00              0x00000002   0x92F8        909               0/0/0
R5.00-00              0x00000003   0x08F3        788               0/0/0
R5.01-00              0x00000001   0x8BCA        784               0/0/0
IS-IS Level-2 Link State Database:
LSPID                 LSP Seq Num  LSP Checksum  LSP Holdtime      ATT/P/OL
R1.00-00            * 0x00000017   0xF432        906               0/0/0
R2.00-00              0x00000272   0xF58C        903               0/0/0
R2.01-00              0x00000133   0x3E64        0 (883)           0/0/0
R3.00-00              0x00000007   0x487F        923               0/0/0
R3.01-00              0x00000001   0x7CE0        893               0/0/0
R4.00-00              0x00000015   0x053D        920               0/0/0
R4.01-00              0x00000002   0x22F1        919               0/0/0
R5.00-00              0x0000000E   0x1C0D        917               0/0/0
R5.01-00              0x00000001   0x1BC3        784               0/0/0
R1#
```

R4

```
R4#show isis database

IS-IS Level-1 Link State Database:
LSPID                 LSP Seq Num  LSP Checksum  LSP Holdtime      ATT/P/OL
R1.00-00              0x0000000A   0xF254        894               0/0/0
R2.00-00              0x0000026D   0x61A6        887               0/0/0
R2.01-00              0x00000133   0x3E64        0 (875)           0/0/0
R3.00-00              0x00000005   0x735D        905               0/0/0
R3.01-00              0x00000001   0xECE7        890               0/0/0
R4.00-00            * 0x0000000A   0x6F8D        907               0/0/0
R4.01-00            * 0x00000002   0x92F8        903               0/0/0
R5.00-00              0x00000003   0x08F3        782               0/0/0
R5.01-00              0x00000001   0x8BCA        778               0/0/0
IS-IS Level-2 Link State Database:
LSPID                 LSP Seq Num  LSP Checksum  LSP Holdtime      ATT/P/OL
R1.00-00              0x00000017   0xF432        896               0/0/0
R2.00-00              0x00000272   0xF58C        896               0/0/0
R2.01-00              0x00000133   0x3E64        0 (875)           0/0/0
R3.00-00              0x00000007   0x487F        915               0/0/0
R3.01-00              0x00000001   0x7CE0        885               0/0/0
R4.00-00            * 0x00000015   0x053D        913               0/0/0
R4.01-00            * 0x00000002   0x22F1        913               0/0/0
R5.00-00              0x0000000E   0x1C0D        911               0/0/0
R5.01-00              0x00000001   0x1BC3        778               0/0/0
R4#
```

IS-IS: Routing Tables

R1

```
Gateway of last resort is not set

      10.0.0.0/8 is variably subnetted, 14 subnets, 2 masks
C        10.0.12.0/30 is directly connected, Serial1/0
L        10.0.12.1/32 is directly connected, Serial1/0
C        10.0.13.0/30 is directly connected, Ethernet0/0
L        10.0.13.1/32 is directly connected, Ethernet0/0
C        10.0.14.0/30 is directly connected, Serial1/1
L        10.0.14.1/32 is directly connected, Serial1/1
i L1     10.0.24.0/30 [115/20] via 10.0.14.2, 00:10:28, Serial1/1
                      [115/20] via 10.0.12.2, 00:10:28, Serial1/0
i L1     10.0.34.0/30 [115/20] via 10.0.14.2, 00:09:58, Serial1/1
                      [115/20] via 10.0.13.2, 00:09:58, Ethernet0/0
i L1     10.0.45.0/30 [115/20] via 10.0.14.2, 00:13:20, Serial1/1
C        10.1.1.1/32 is directly connected, Loopback0
i L1     10.2.2.2/32 [115/20] via 10.0.12.2, 00:10:28, Serial1/0
i L1     10.3.3.3/32 [115/20] via 10.0.13.2, 00:09:58, Ethernet0/0
i L1     10.4.4.4/32 [115/20] via 10.0.14.2, 00:13:20, Serial1/1
i L1     10.5.5.5/32 [115/30] via 10.0.14.2, 00:11:51, Serial1/1
      172.16.0.0/16 is variably subnetted, 4 subnets, 2 masks
C        172.16.1.0/24 is directly connected, Ethernet0/3
L        172.16.1.1/32 is directly connected, Ethernet0/3
i L1     172.16.3.0/24 [115/20] via 10.0.13.2, 00:09:58, Ethernet0/0
i L1     172.16.5.0/24 [115/30] via 10.0.14.2, 00:11:51, Serial1/1
R1#
```

R4

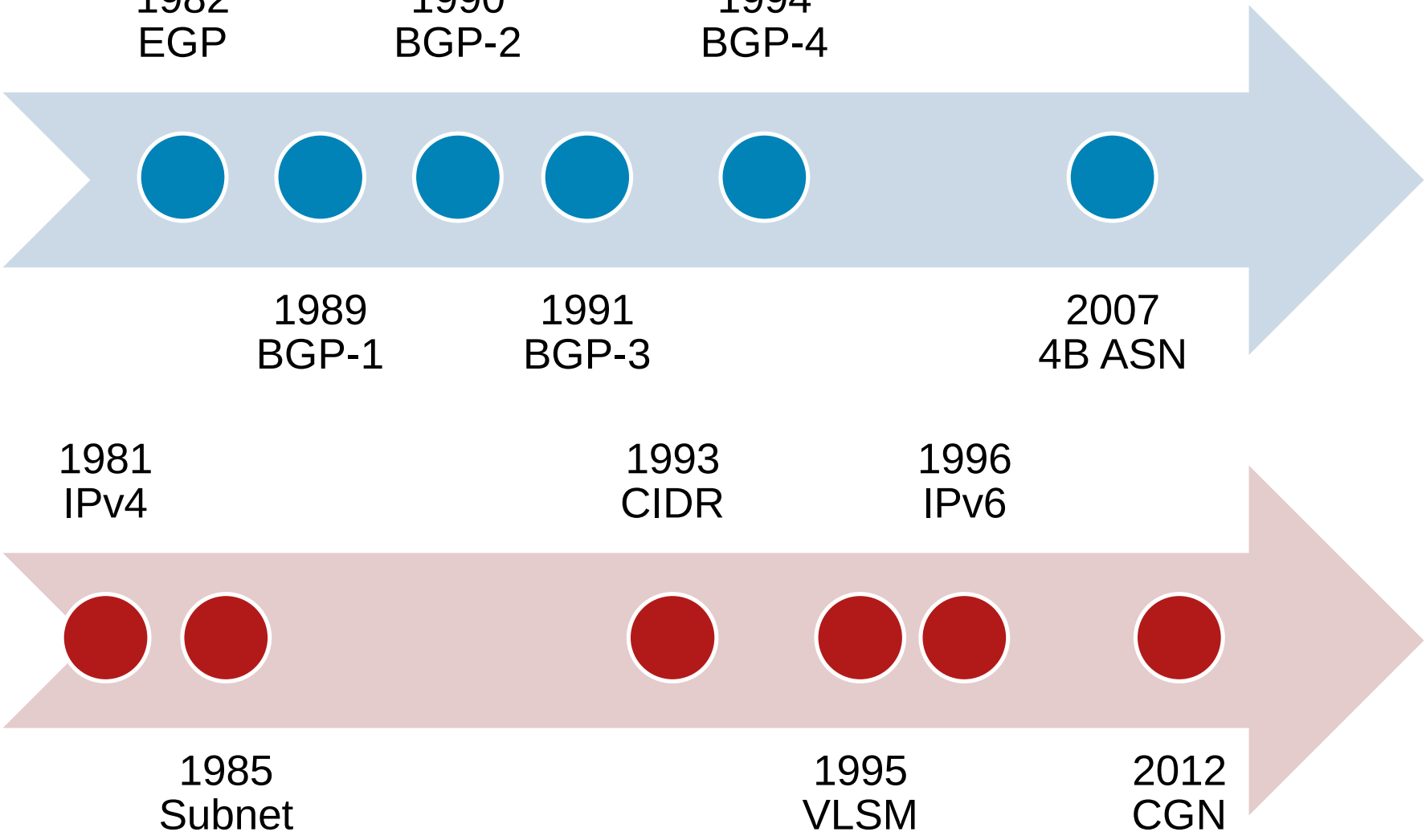
```
Gateway of last resort is not set

      10.0.0.0/8 is variably subnetted, 14 subnets, 2 masks
i L1     10.0.12.0/30 [115/20] via 10.0.24.1, 00:10:25, Serial1/2
                      [115/20] via 10.0.14.1, 00:10:25, Serial1/1
i L1     10.0.13.0/30 [115/20] via 10.0.34.1, 00:09:55, Ethernet0/1
                      [115/20] via 10.0.14.1, 00:09:55, Serial1/1
C        10.0.14.0/30 is directly connected, Serial1/1
L        10.0.14.2/32 is directly connected, Serial1/1
C        10.0.24.0/30 is directly connected, Serial1/2
L        10.0.24.2/32 is directly connected, Serial1/2
C        10.0.34.0/30 is directly connected, Ethernet0/1
L        10.0.34.2/32 is directly connected, Ethernet0/1
C        10.0.45.0/30 is directly connected, Ethernet0/2
L        10.0.45.1/32 is directly connected, Ethernet0/2
i L1     10.2.2.2/32 [115/20] via 10.0.24.1, 00:10:25, Serial1/2
i L1     10.3.3.3/32 [115/20] via 10.0.34.1, 00:09:55, Ethernet0/1
C        10.4.4.4/32 is directly connected, Loopback0
i L1     10.5.5.5/32 [115/20] via 10.0.45.2, 00:11:59, Ethernet0/2
      172.16.0.0/24 is subnetted, 3 subnets
i L1     172.16.1.0 [115/20] via 10.0.14.1, 00:13:41, Serial1/1
i L1     172.16.3.0 [115/20] via 10.0.34.1, 00:09:55, Ethernet0/1
i L1     172.16.5.0 [115/20] via 10.0.45.2, 00:11:59, Ethernet0/2
R4#
```

Path-Vector Protocol Evolution

“Addressing can follow topology or topology can follow addressing. Choose one!”

Yakov Rekhter

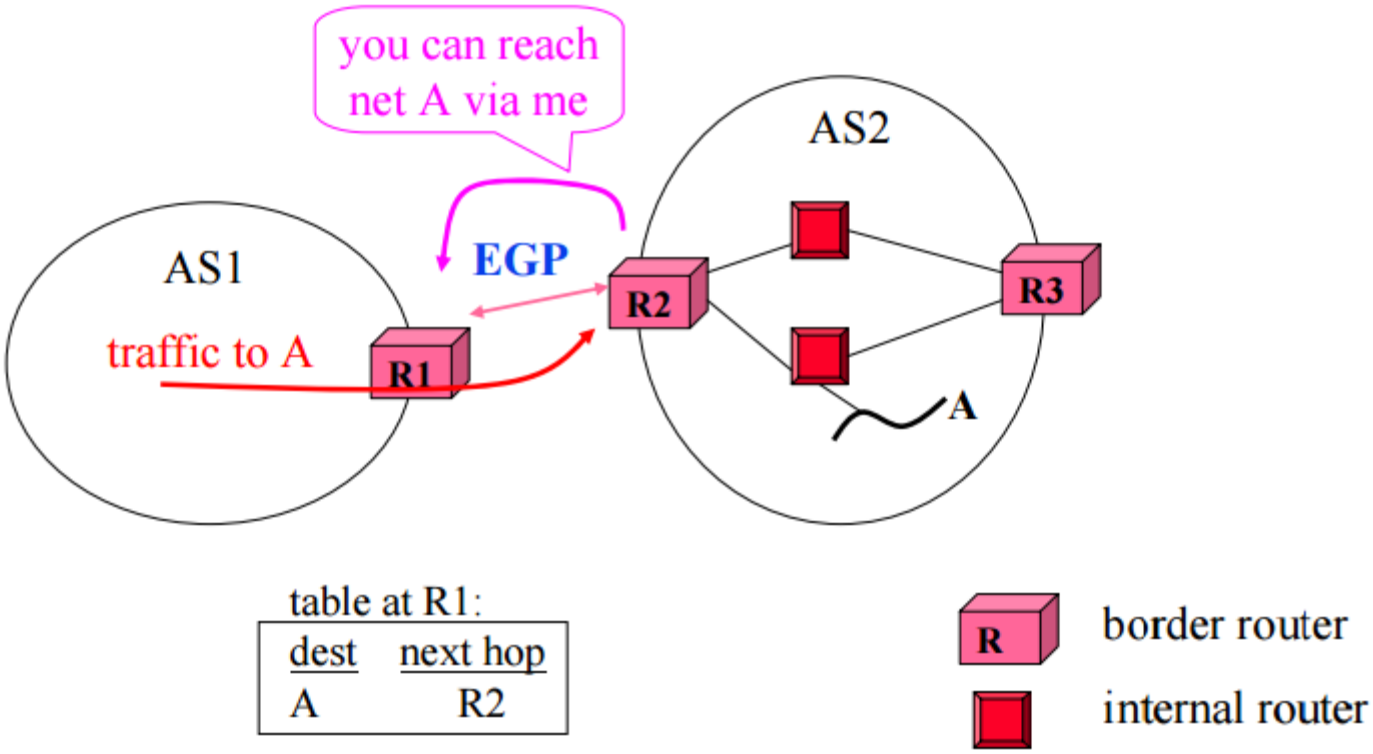
Timeline


1982
EGP
1990
BGP-2
1994
BGP-4
1989
BGP-1
1991
BGP-3
2007
4B ASN
1981
IPv4
1993
CIDR
1996
IPv6
1985
Subnet
1995
VLSM
2012
CGN

EGP

- RFC 827 a RFC 904
- Autonomous System (AS) has the responsibility of advertising reachability info to other ASs.
- A mechanism that allows non-core routers to learn routes from core (external routes) routers so that they can choose optimal backbone routes
- A mechanism for non-core routers to inform core routers about hidden networks (internal routes)
- Classful
- Unreliable transfer
  - Sequence numbers
  - Polling

EGP: Concept


you can reach
net A via me
AS2
AS1
EGP
R2
R3
traffic to A
R1
A
table at R1:
dest
next hop
A
R2
R
border router
internal router

EGP: Operation

- Neighbor Acquisition
  - Reliable 2-way handshake
- Neighbor Reachability
  - Hellos
    - K-out-of-J received hellos OK => Neighbor UP
    - K-out-of-N hellos NOT OK => Neighbor DOWN
- Updates/Queries
  - EGP is an incremental protocol. New info => send updates
  - Each router can query neighbors as well
  - Reachability advertized; metrics ignored
  - Requires a tree topology of ASes to avoid loops

EGP's Weakness

- EGP does not interpret the distance metrics in routing update messages => cannot compute shorter of two routes
- As a result it restricts the topology to a tree structure, with the core as the root
  - Rapid growth => many networks may be temporarily unreachable
  - Only one path to destination => no load sharing

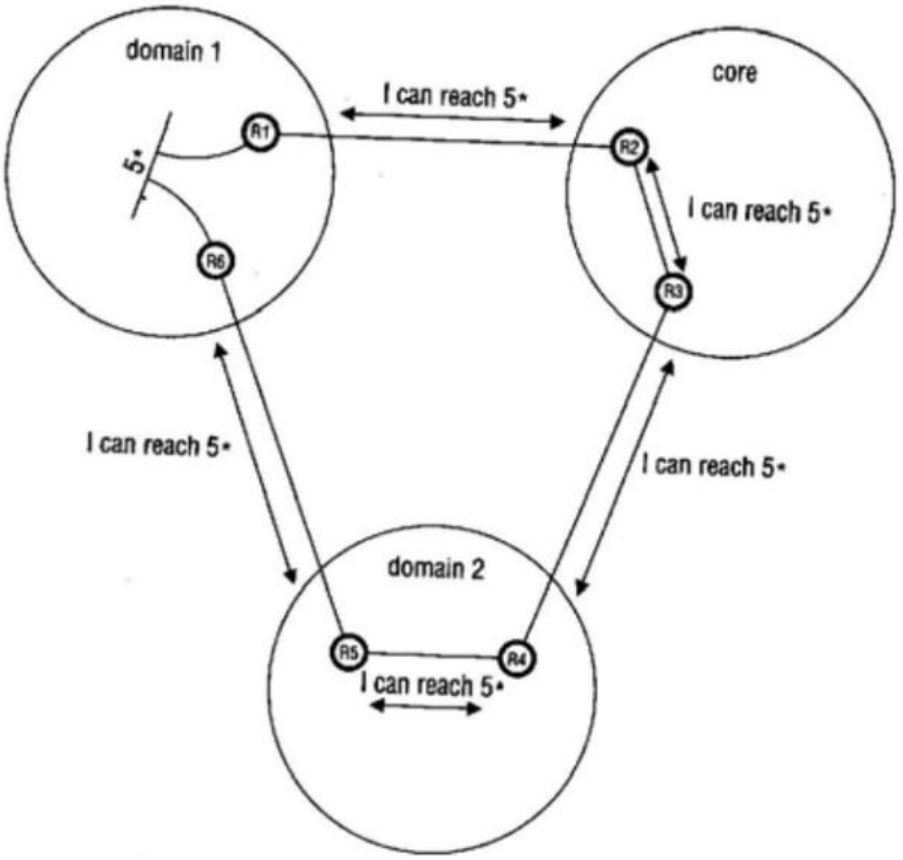


Figure 14.61 Topology in which EGP would not work

BGP

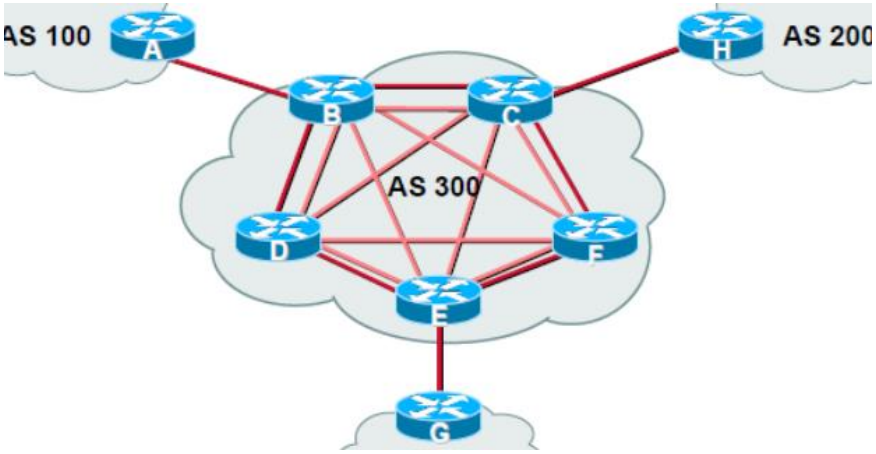
- RFC 1771
- Allows multiple cores and arbitrary topologies of AS interconnection.
  - Uses a path-vector concept which enables loop prevention in complex topologies
- Is a policy-based routing protocol
  - In AS-level, shortest path may not be preferred for policy, security, cost reasons.
  - Different routers have different preferences => as packet goes thru network it will encounter different policies
  - *Bellman-Ford/Dijkstra don't work due to the way how topology change is propagated (iterations/flooding)!*
- Is the de facto EGP of today's global Internet
  - Relatively simple protocol, but configuration is complex and the entire world can see, and be impacted by, your mistakes

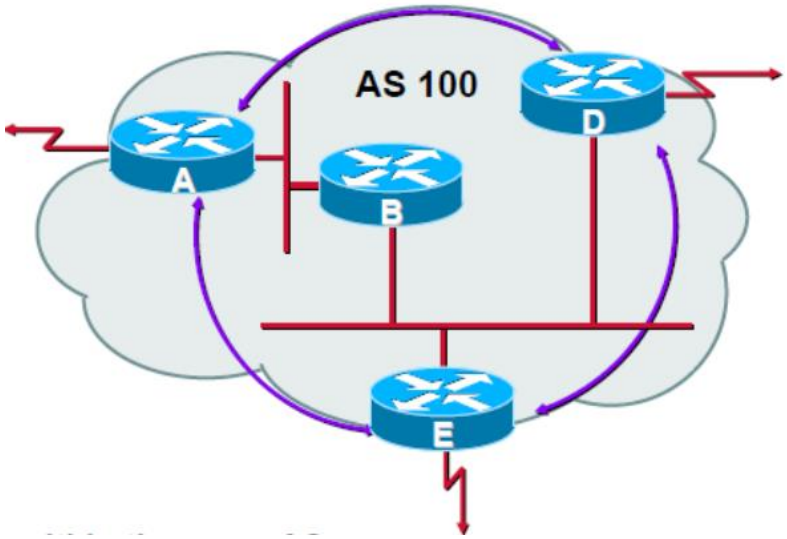
BGP

- Operates over TCP on port 179 between 2 peers (reliability)
  - Exchange entire BGP table first
  - Later exchanges only incremental updates
  - Application (BGP)-level keepalive messages
  - Hold-down timer (at least 3 sec) locally config
- Interior and exterior peers
  - need to Exchange reachability information among interior peers before updating intra-AS forwarding table
  - When a BGP Speaker A advertises a prefix to its B that it has a path to IP prefix C, B can be certain that A is actively using that AS path to reach that destination

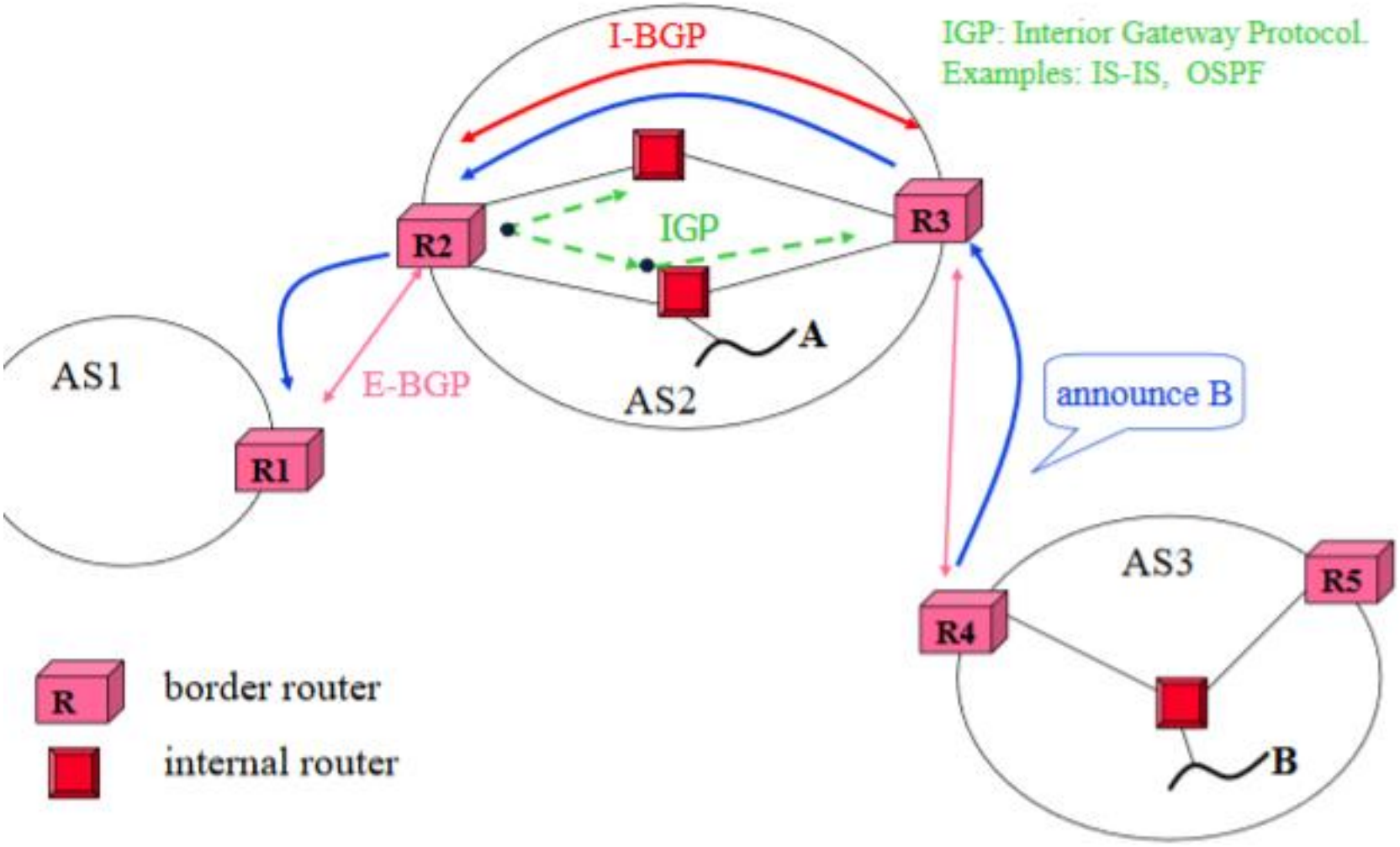
BGP: Two Types of Neighbors

- External BGP
  - Routers from different ASes
  - Direct connection
  - Filtering, policing
- Internal BGP
  - Routers from the same AS
  - Non-adjacent connection

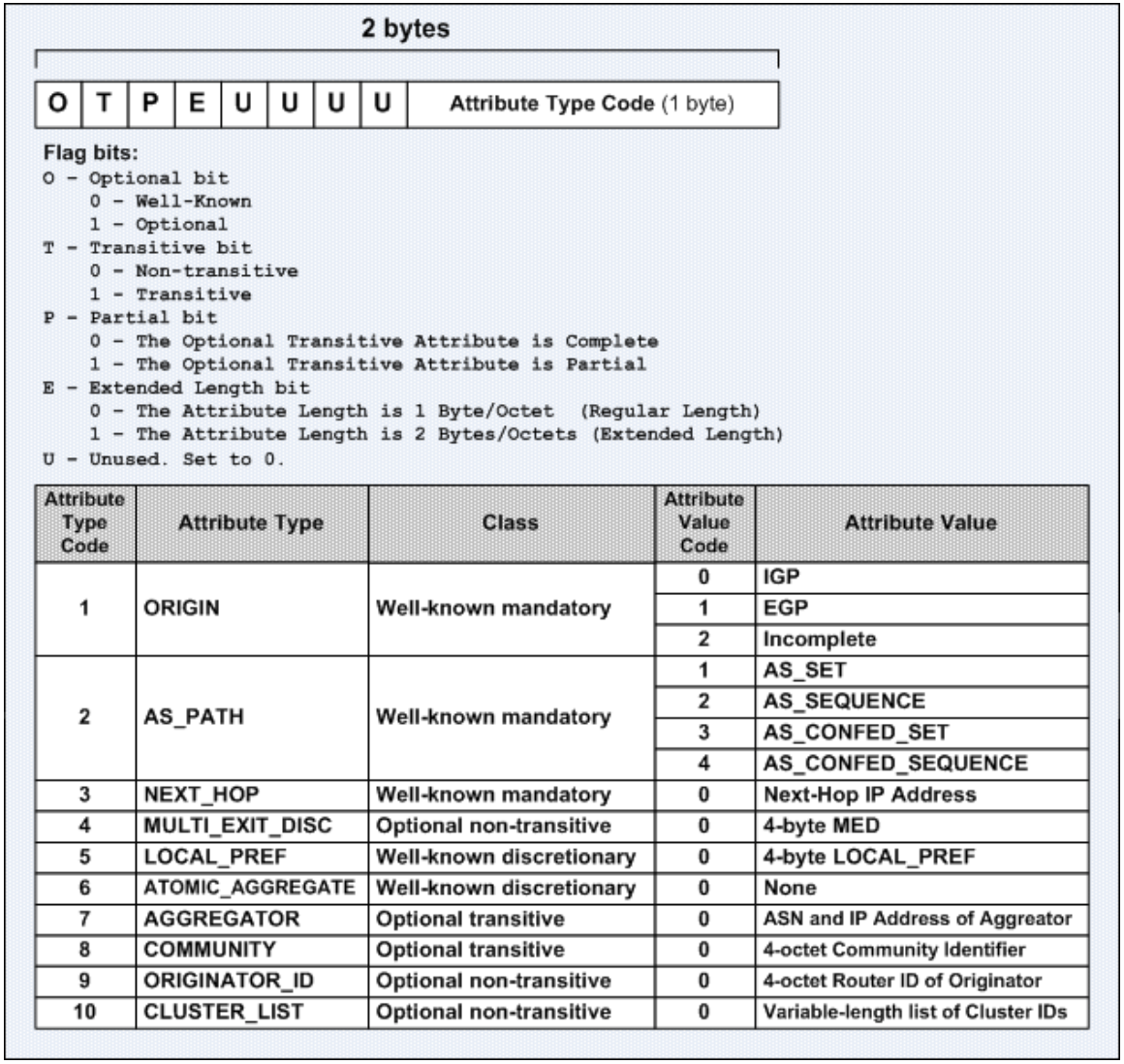




iBGP and eBGP Coexistence



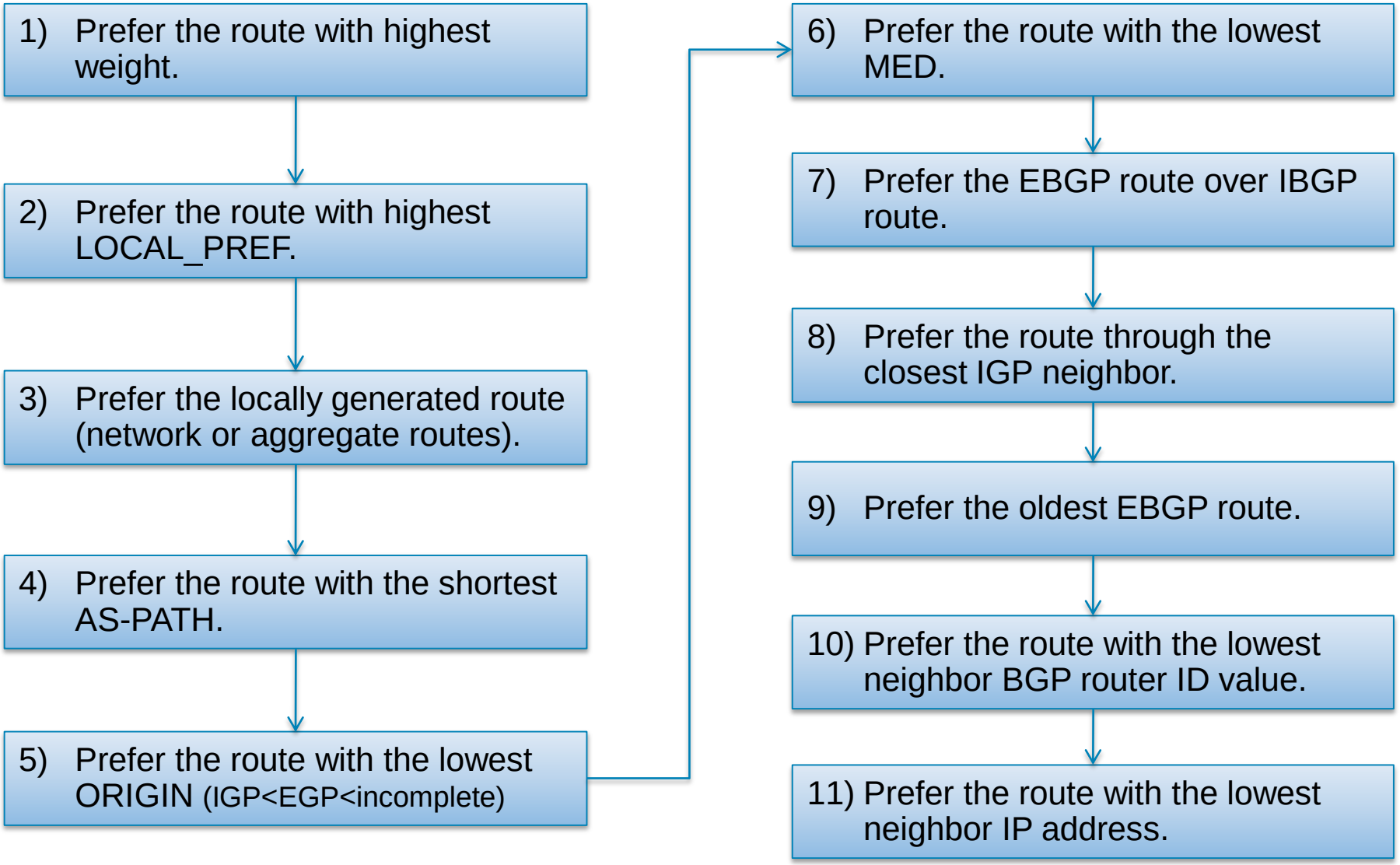
BGP: Attributes



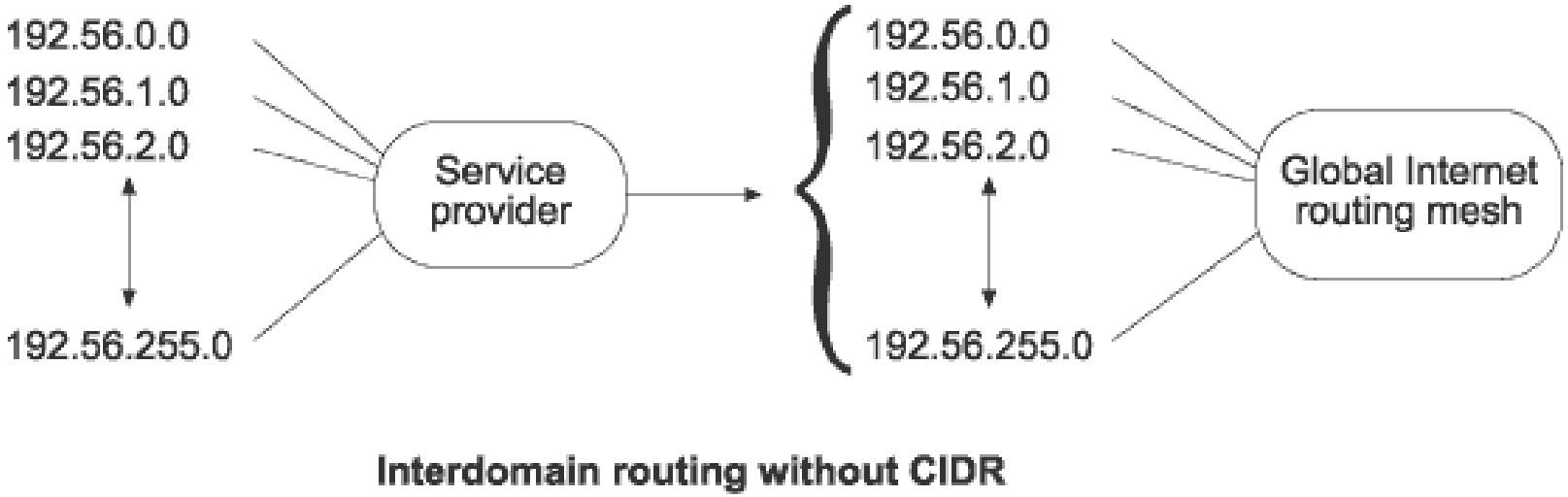
Attribute Type Code	Attribute Type	Class	Attribute Value Code	Attribute Value
1	ORIGIN	Well-known mandatory	0	IGP
			1	EGP
			2	Incomplete
2	AS_PATH	Well-known mandatory	1	AS_SET
			2	AS_SEQUENCE
			3	AS_CONFED_SET
			4	AS_CONFED_SEQUENCE
3	NEXT_HOP	Well-known mandatory	0	Next-Hop IP Address
4	MULTI_EXIT_DISC	Optional non-transitive	0	4-byte MED
5	LOCAL_PREF	Well-known discretionary	0	4-byte LOCAL_PREF
6	ATOMIC_AGGREGATE	Well-known discretionary	0	None
7	AGGREGATOR	Optional transitive	0	ASN and IP Address of Aggreator
8	COMMUNITY	Optional transitive	0	4-octet Community Identifier
9	ORIGINATOR_ID	Optional non-transitive	0	4-octet Router ID of Originator
10	CLUSTER_LIST	Optional non-transitive	0	Variable-length list of Cluster IDs

http://2.bp.blogspot.com/-UdXNfK9ERw/TyJAHgHS5YI/AAAAAAAAAIM/D4at2KGkhOU/s1600/bgp+attributes.PNG

BGP: Route Selection



CIDR

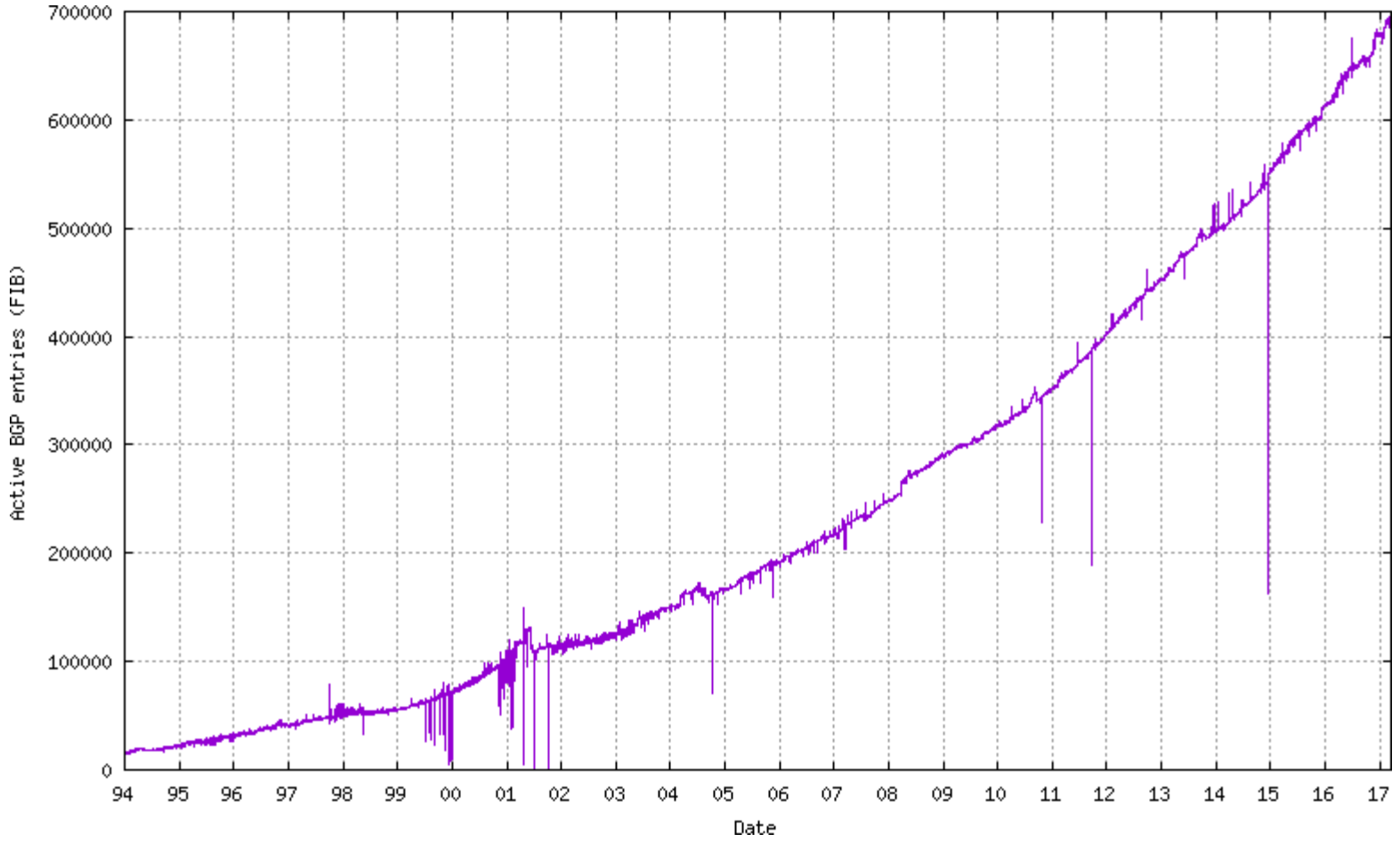


https://www.juniper.net/techpubs/images/g013302.gif

How Many Network Are on Internet?

- https://bgp.potaroo.net/as6447/

FIB / RIB Table Reports (plots)	Data Sets(txt)
Active BGP entries (FIB)	692487
All BGP entries (RIB)	29712125



How many Autonomous System do exist?

- https://bgp.potaroo.net/as6447/

AS Reports (plots)	Data Sets(txt)		Additional Reports (plots)	Data Sets(txt)
Unique ASes	57297		ASes visible in only one AS path	35915
Origin only ASes	48277		Origin ASs announced via a single AS path	34939
Transit only ASes	279		Originating AS ATOM count	57018
Mixed ASes	8741		Originating AS ATOM compression	0.0820
Multi-Origin Prefixes	6954			
ASes originating a single prefix	21534			
Average entries per origin AS	12.1940		Maximum entries for an origin AS (AS262589)	53
Average address range span for an origin AS	49787.7147		Maximum address range for an origin AS (AS4134)	120809728

How BIG is Internet for Router?

- http://www.routeviews.org/

route-views.routeviews.org - PuTTY

```
route-views>sh ip bgp sum
BGP router identifier 128.223.51.103, local AS number 6447
BGP table version is 40390544, main routing table version 40390544
726191 network entries using 180095368 bytes of memory
27887133 path entries using 3346455960 bytes of memory
4298699/127692 BGP path/bestpath attribute entries using 1066077352 bytes of memory
3972925 BGP AS-PATH entries using 196599780 bytes of memory
2 BGP ATTR_SET entries using 80 bytes of memory
153644 BGP community entries using 18775584 bytes of memory
1126 BGP extended community entries using 50112 bytes of memory
0 BGP route-map cache entries using 0 bytes of memory
0 BGP filter-list cache entries using 0 bytes of memory
BGP using 4808054156 total bytes of memory
BGP activity 1948481/1168981 prefixes, 227696216/198009168 paths, scan interval 60 secs

Neighbor        V           AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
4.69.184.193    4         3356 207050992   72582 40390529    0    0 6w1d       674780
5.101.110.2     4       202018 18929785  137834 40390529    0    0 6w1d       677567
12.0.1.63       4         7018 11214262   69578 40390529    0    0 6w1d       675984
37.139.139.0    4        57866 15522154  126879 40390529    0    0 5w5d       676746
64.71.137.241   4         6939 1515946    8586 40390529    0    0 5d10h      697737
66.59.190.221   4         6539       0       0        1    0    0 6d14h    Idle
66.110.0.86     4         6453       0       0        1    0    0 never    Idle
66.185.128.48   4         1668       0       0        1    0    0 never    Active
69.31.111.244   4         4436       0       0        1    0    0 never    Idle
80.241.176.31   4        20771       0       0        1    0    0 never    Idle
89.149.178.10   4         3257 6363372   14034 40390529    0    0 6w1d       676137
91.218.184.60   4        49788 18819274   52314 40390529    0    0 6w1d       678125
93.104.209.174  4        58901  106803    1772 40390529    0    0 1d02h      296664
94.142.247.3    4         8283 47175775   52257 40390529    0    0 6w1d       678517
```

Multicast Routing Protocol Evolution

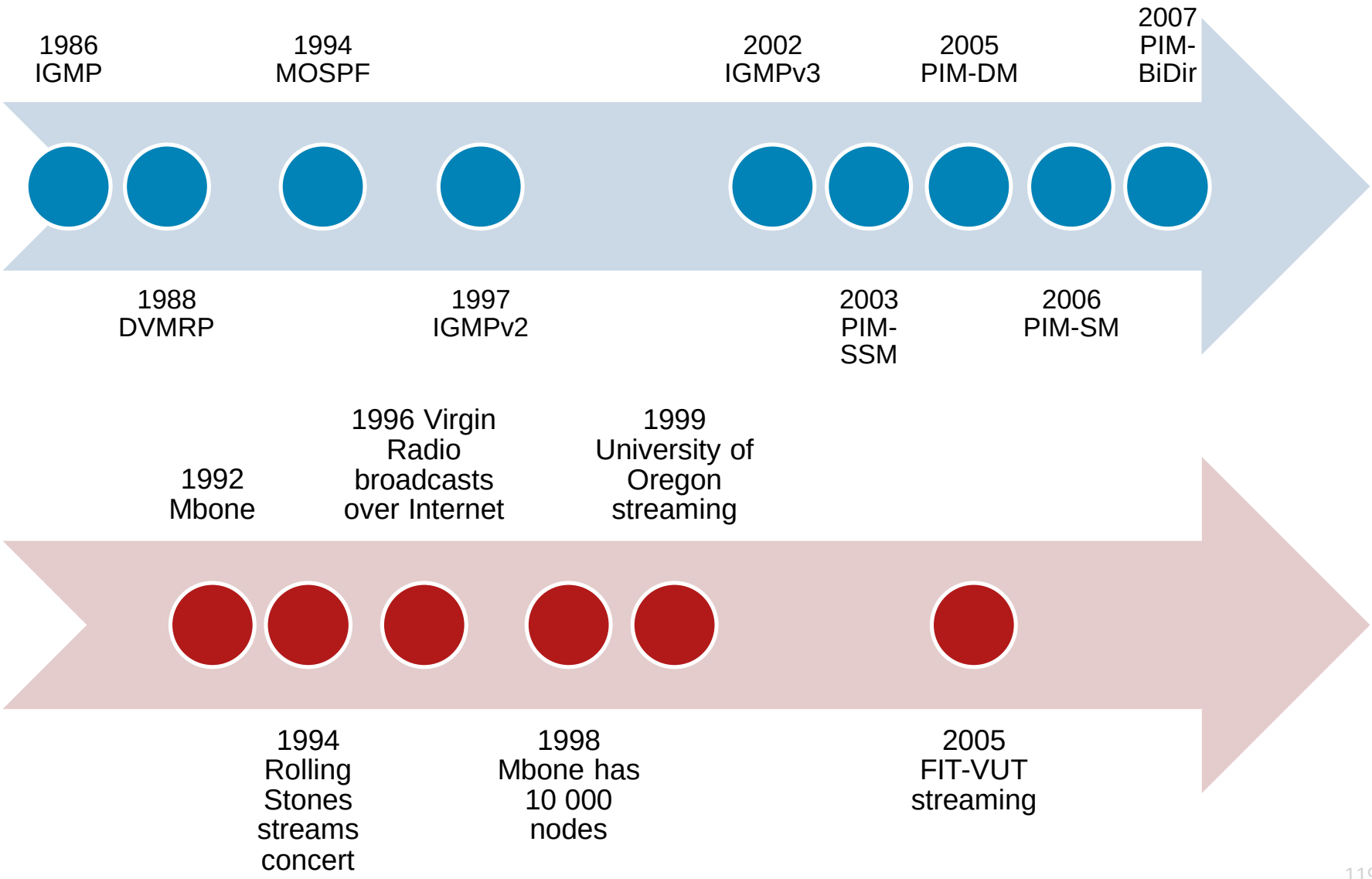
A multicast packet walks into a bar and leaves by four different exits at the same time.

A multicast packet walks into 100 bars at one time.

Multicast jokes are good, but you can only get them if you bother to listen.

Anonymous Jokers

Timeline


1986
IGMP
1988
DVMRP
1992
Mbone
1994
MOSPF
1994
Rolling
Stones
streams
concert
1996 Virgin
Radio
broadcasts
over Internet
1997
IGMPv2
1998
Mbone has
10 000
nodes
1999
University of
Oregon
streaming
2002
IGMPv3
2003
PIM-
SSM
2005
PIM-DM
2005
FIT-VUT
streaming
2006
PIM-SM
2007
PIM-
BiDir

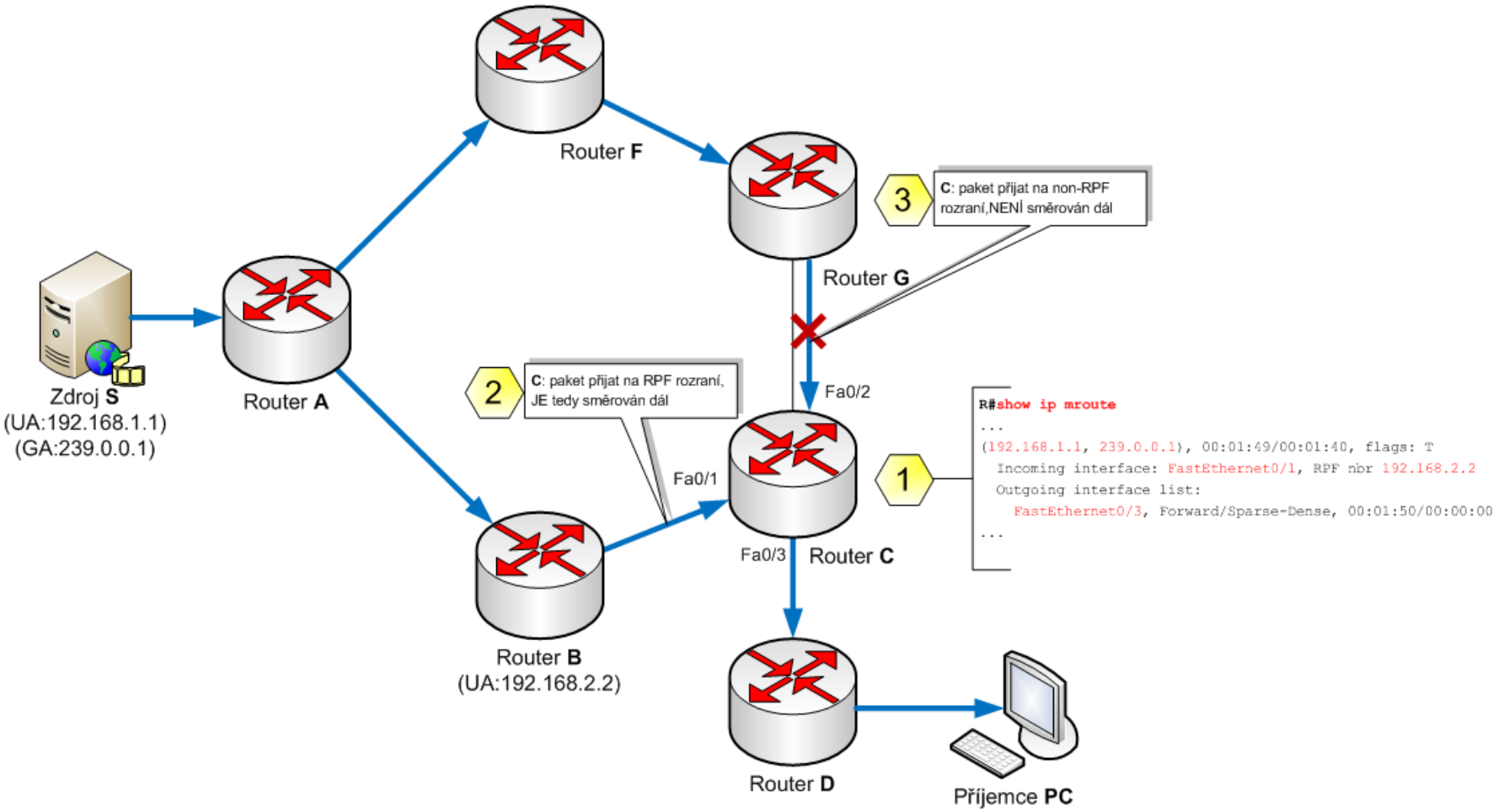
DVMRP

- RFC 1075
- *Helped MBONE to grow*
- Based on RIP, extends IGMPv1 with new messages
  - Hop count towards multicast destinations
  - Scaling issues
- Floods traffic
  - Receivers need to explicitly prune them from multicast tree
  - Periodically forgets about pruned branches

Multicast Reverse Path Forwarding ①

- Algorithm
  1) Upon multicast packet reception, IP address of multicast source is validated against unicast routing table.
  2) Unicast route towards source IP network is looked up in RT. Is multicast source reachable via searched route next-hop interface?
  3) Do following based on the answer:
     a) Yes – packet is forwarded
     b) No – packet is discarded (potential loop)

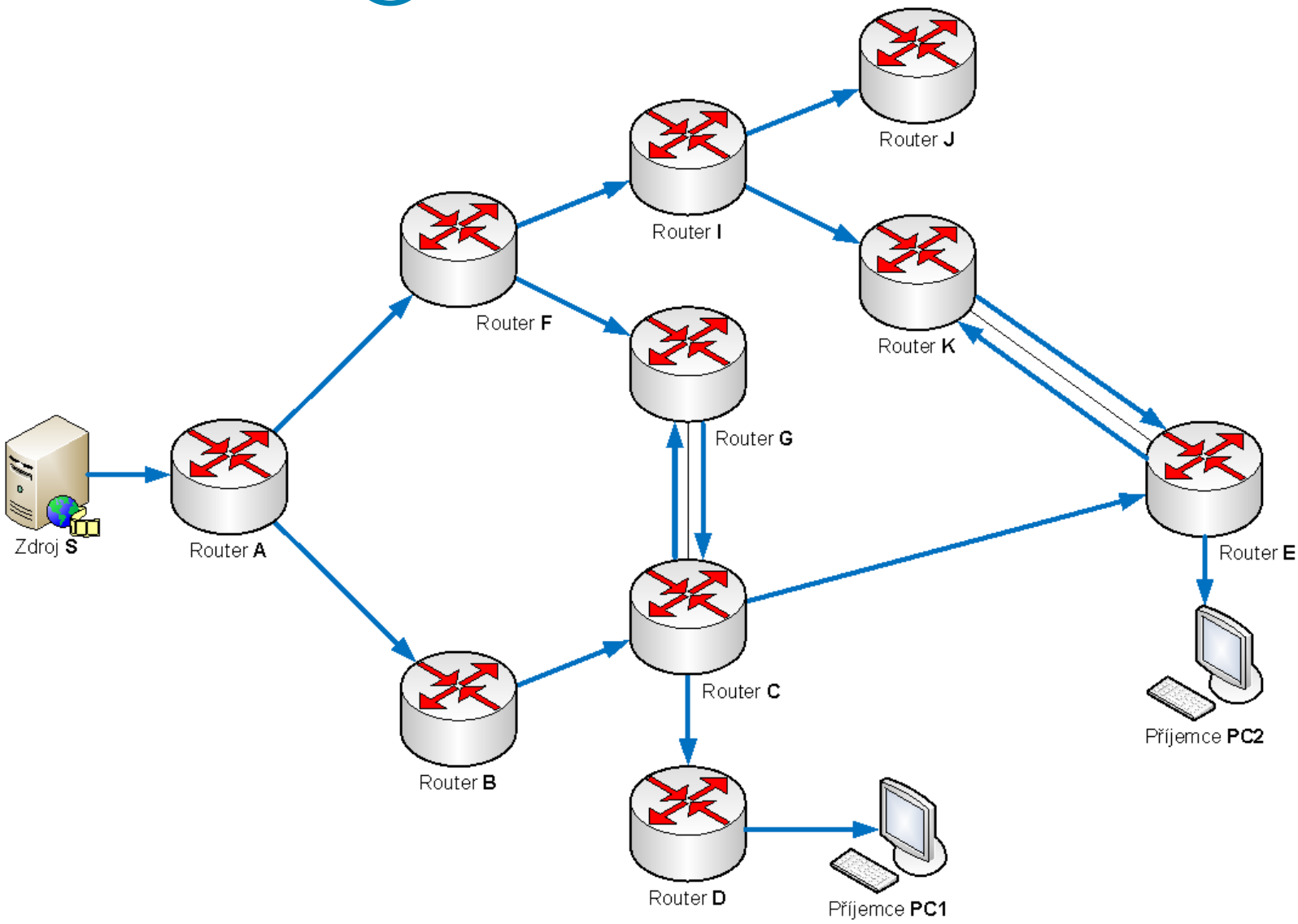
Reverse Path Forwarding ②



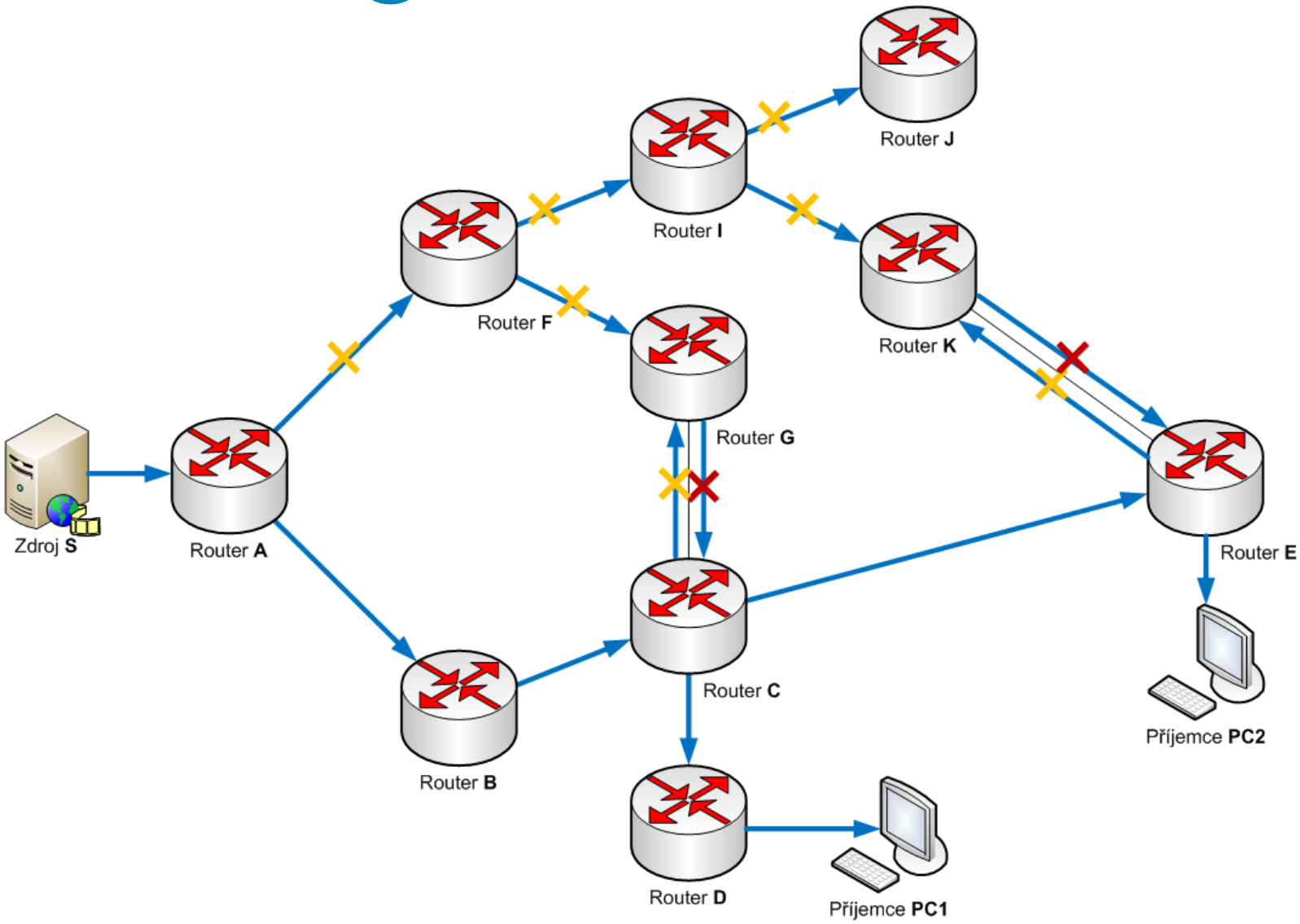
PIM

- PIM is actually not a routing protocol which would carried IP prefixes and their metrics – *more less it is signaling protocol*
  - L3 protocol with IP protocol number 103
- PIM needs another unicast routing protocol to be active but it is independent on it – *it doesn't matter whether it is RIP, OSPF or EIGRP*
- PIM routers create multicast routing table to forward multicast datagrams based on unicast RIB
- PIM works in two different regimes by initial design
  - **Dense mode**: Multicast traffic is spread across whole topology. `IF` a router has no multicast members on the some of its segments `THEN` the router prune itself from multicast distribution tree for target multicast group – a.k.a. periodic flood-and-prune (RFC 3973)
  - **Sparse mode**: Multicast traffic is sent via distribution trees which are created based on receiving clients requests (RFC 7761)
  - **Source Specific Multicast (PIM-SSM)**: Benefiting from IGMPv3 capable of specifying particular multicast source to receive data from
  - **BiDirectional (BiDir PIM)**: Senders and receivers communicate

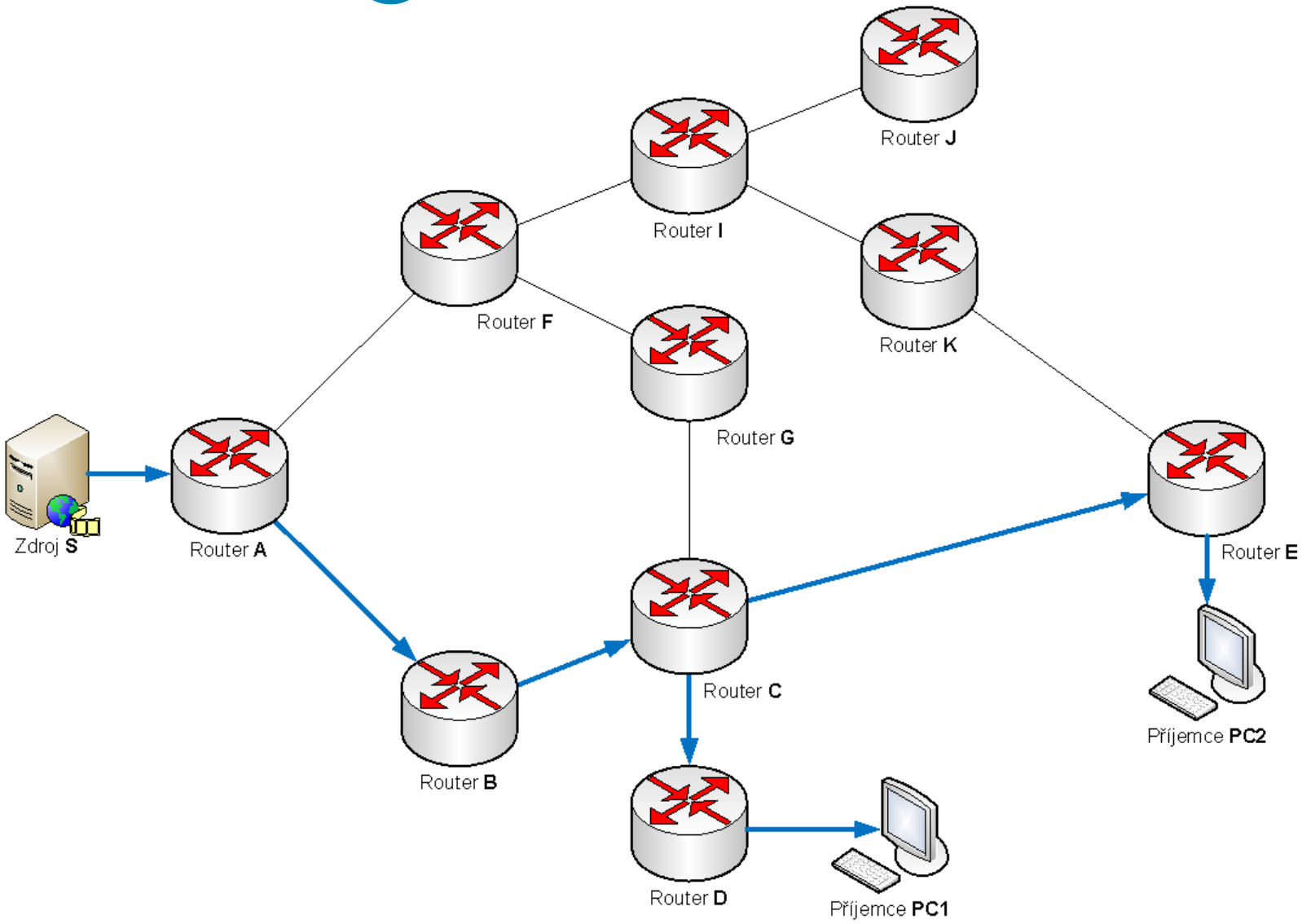
PIM-DM: Demo ①


Router J
Router I
Router F
Router K
Router G
Zdroj S
Router A
Router E
Router C
Příjemce PC2
Router B
Router D
Příjemce PC1

PIM-DM: Demo ②

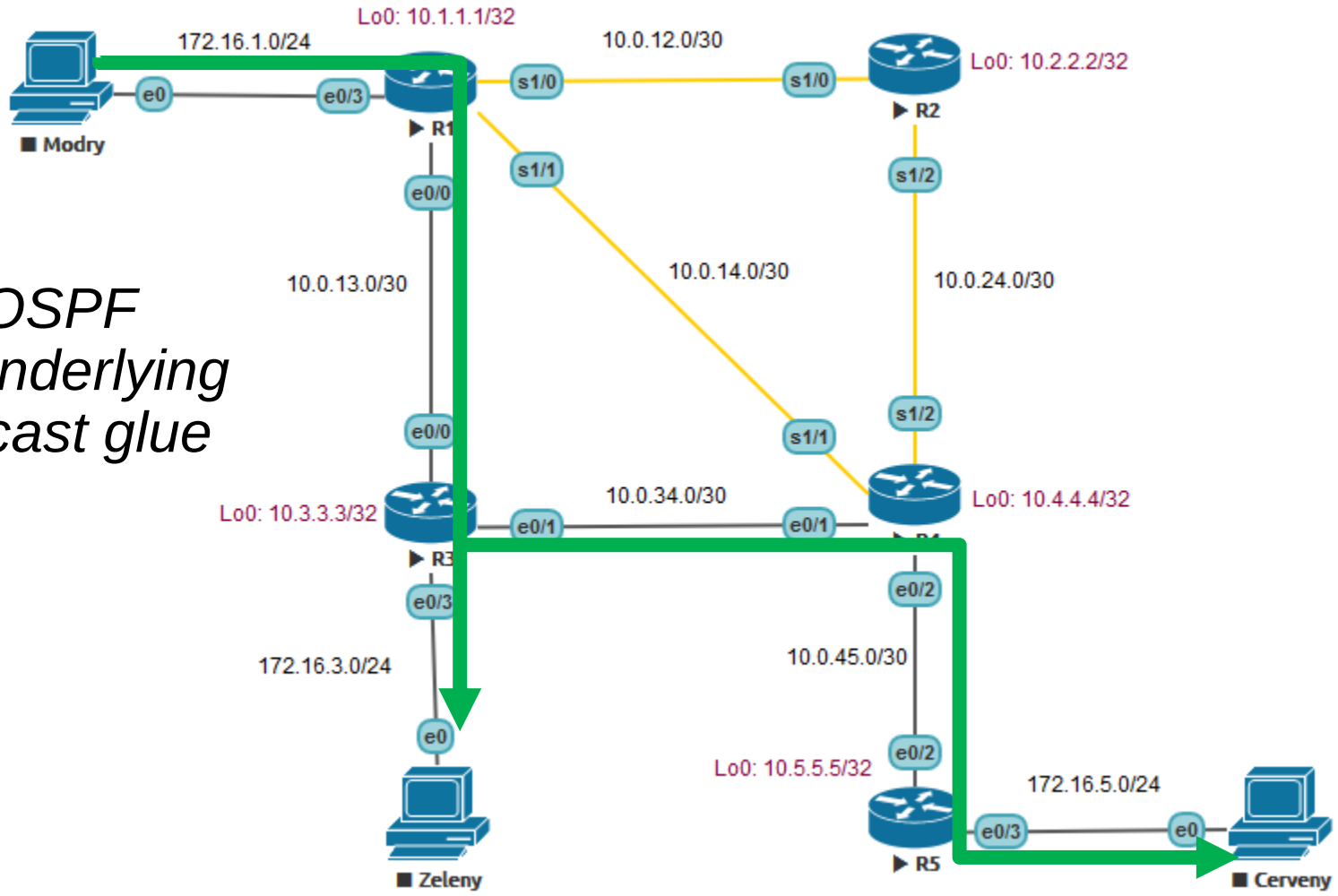

Router J
Router I
Router F
Router K
Router G
Zdroj S
Router A
Router E
Router C
Příjemce PC2
Router B
Router D
Příjemce PC1

PIM-DM: Demo ③


Router J
Router I
Router F
Router K
Router G
Zdroj S
Router A
Router E
Router C
Příjemce PC2
Router B
Router D
Příjemce PC1

PIM-DM: Network Graph

OSPF as underlying unicast glue



PIM-DM: Routing Table

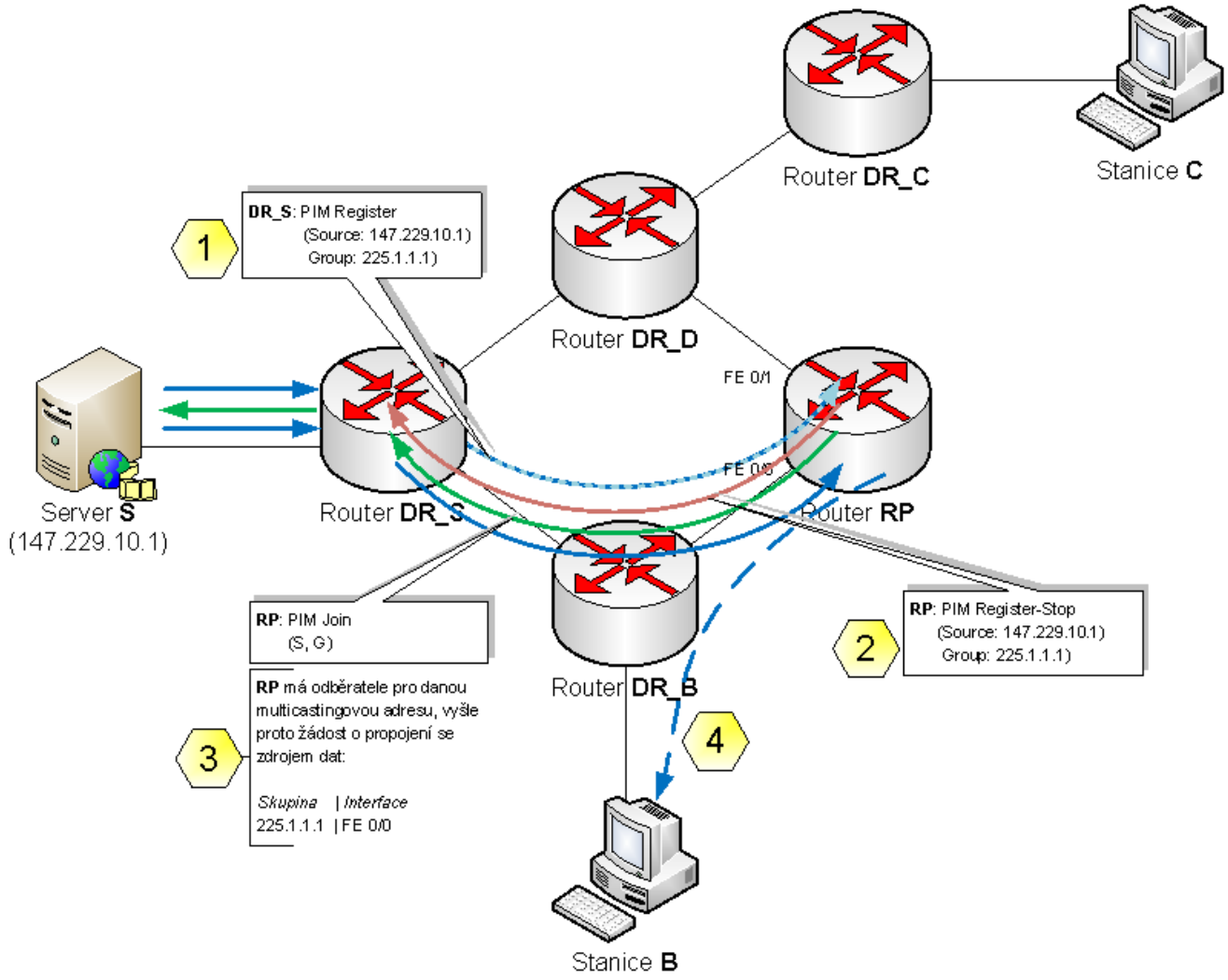
R1

```
R1#show ip mroute
IP Multicast Routing Table
Flags: D - Dense, S - Sparse, B - Bidir Group, s - SSM Group, C - Connected,
       L - Local, P - Pruned, R - RP-bit set, F - Register flag,
       T - SPT-bit set, J - Join SPT, M - MSDP created entry, E - Extranet,
       X - Proxy Join Timer Running, A - Candidate for MSDP Advertisement,
       U - URD, I - Received Source Specific Host Report,
       Z - Multicast Tunnel, z - MDT-data group sender,
       Y - Joined MDT-data group, y - Sending to MDT-data group,
       G - Received BGP C-Mroute, g - Sent BGP C-Mroute,
       N - Received BGP Shared-Tree Prune, n - BGP C-Mroute suppressed,
       Q - Received BGP S-A Route, q - Sent BGP S-A Route,
       V - RD & Vector, v - Vector, p - PIM Joins on route,
       x - VxLAN group
Outgoing interface flags: H - Hardware switched, A - Assert winner, p - PIM
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(*, 239.0.0.1), 00:07:29/stopped, RP 0.0.0.0, flags: D
  Incoming interface: Null, RPF nbr 0.0.0.0
  Outgoing interface list:
    Serial1/1, Forward/Sparse-Dense, 00:07:29/stopped
    Serial1/0, Forward/Sparse-Dense, 00:07:29/stopped
    Ethernet0/0, Forward/Sparse-Dense, 00:07:29/stopped

(172.16.1.1, 239.0.0.1), 00:00:56/00:02:45, flags: T
  Incoming interface: Ethernet0/3, RPF nbr 0.0.0.0
  Outgoing interface list:
    Ethernet0/0, Forward/Sparse-Dense, 00:00:56/stopped
    Serial1/0, Prune/Sparse-Dense, 00:00:56/00:02:03
    Serial1/1, Prune/Sparse-Dense, 00:00:56/00:02:03, A
```

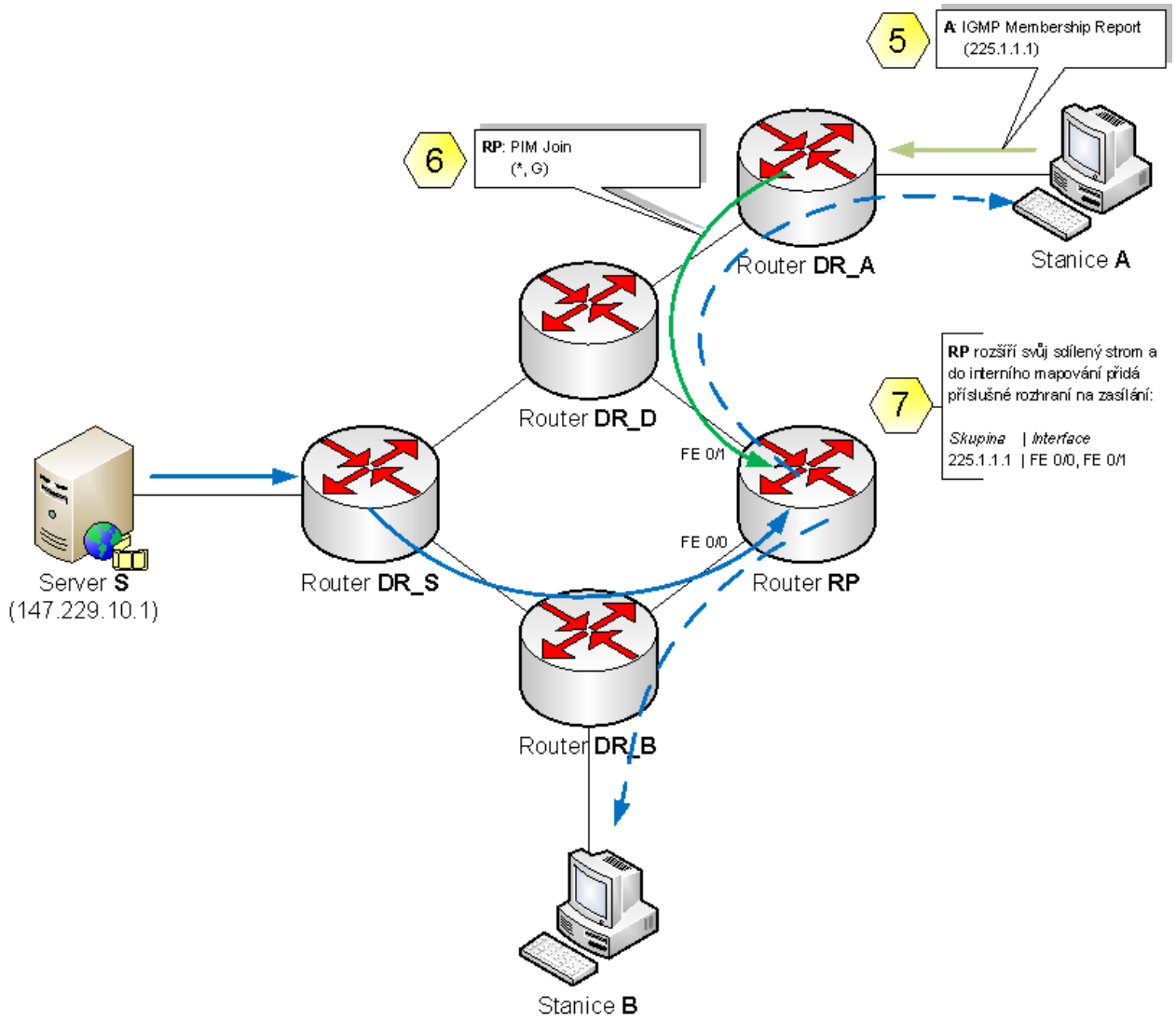
R4

```
R4#show ip mroute
IP Multicast Routing Table
Flags: D - Dense, S - Sparse, B - Bidir Group, s - SSM Group, C - Connected,
       L - Local, P - Pruned, R - RP-bit set, F - Register flag,
       T - SPT-bit set, J - Join SPT, M - MSDP created entry, E - Extranet,
       X - Proxy Join Timer Running, A - Candidate for MSDP Advertisement,
       U - URD, I - Received Source Specific Host Report,
       Z - Multicast Tunnel, z - MDT-data group sender,
       Y - Joined MDT-data group, y - Sending to MDT-data group,
       G - Received BGP C-Mroute, g - Sent BGP C-Mroute,
       N - Received BGP Shared-Tree Prune, n - BGP C-Mroute suppressed,
       Q - Received BGP S-A Route, q - Sent BGP S-A Route,
       V - RD & Vector, v - Vector, p - PIM Joins on route,
       x - VxLAN group
Outgoing interface flags: H - Hardware switched, A - Assert winner, p - PIM
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(*, 239.0.0.1), 00:07:24/stopped, RP 0.0.0.0, flags: D
  Incoming interface: Null, RPF nbr 0.0.0.0
  Outgoing interface list:
    Serial1/2, Forward/Sparse-Dense, 00:07:24/stopped
    Serial1/1, Forward/Sparse-Dense, 00:07:24/stopped
    Ethernet0/2, Forward/Sparse-Dense, 00:07:24/stopped
    Ethernet0/1, Forward/Sparse-Dense, 00:07:24/stopped

(172.16.1.1, 239.0.0.1), 00:00:52/00:02:07, flags: T
  Incoming interface: Ethernet0/1, RPF nbr 10.0.34.1
  Outgoing interface list:
    Ethernet0/2, Forward/Sparse-Dense, 00:00:52/stopped
    Serial1/1, Prune/Sparse-Dense, 00:00:52/00:02:07, A
    Serial1/2, Prune/Sparse-Dense, 00:00:52/00:02:07, A
```

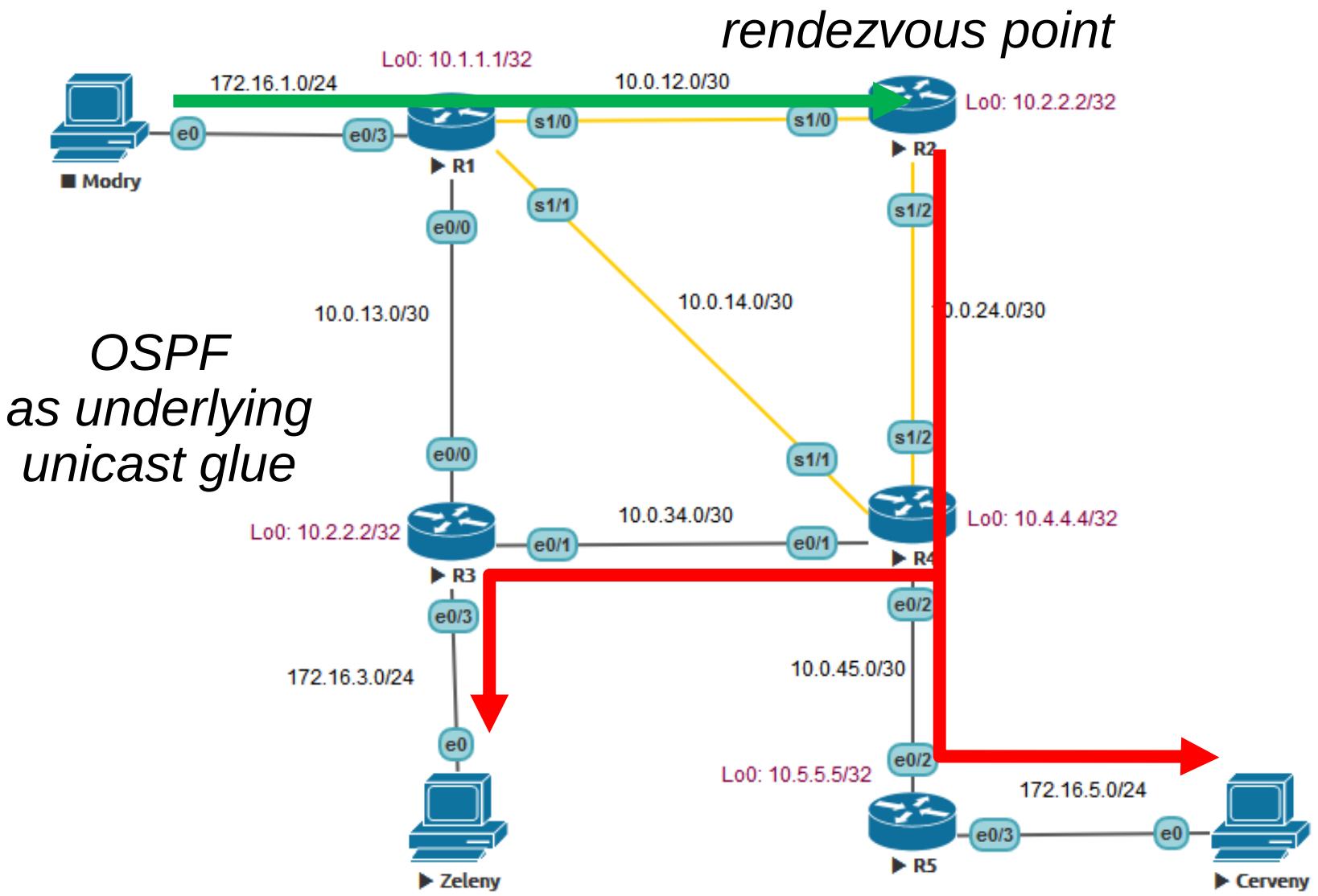
PIM-SM: Demo ①


Router DR_C
Stanice C
1
DR_S: PIM Register
(Source: 147.229.10.1)
Group: 225.1.1.1)
Router DR_D
FE 0/1
FE 0/0
Server S
(147.229.10.1)
Router DR_S
Router RP
RP: PIM Join
(S, G)
RP má odběratele pro danou multicastingovou adresu, vyše proto žádost o propojení se zdrojem dat:
Skupina | Interface
225.1.1.1 | FE 0/0
3
Router DR_B
2
RP: PIM Register-Stop
(Source: 147.229.10.1)
Group: 225.1.1.1)
4
Stanice B

PIM-SM: Demo ②


5
A: IGMP Membership Report
(225.1.1.1)
6
RP: PIM Join
(*, G)
Router DR_A
Stanice A
Router DR_D
7
RP rozšíří svůj sdílený strom a
do interního mapování přidá
příslušné rozhraní na zasílání:
Skupina | Interface
225.1.1.1 | FE 0/0, FE 0/1
FE 0/1
FE 0/0
Server S
(147.229.10.1)
Router DR_S
Router RP
Router DR_B
Stanice B

PIM-SM: Network Graph



L3: Multicast IPv4 RT

R1

```
R1#sh ip mroute
IP Multicast Routing Table
Flags: D - Dense, S - Sparse, B - Bidir Group, s - SSM Group, C - Connected,
       L - Local, P - Pruned, R - RP-bit set, F - Register flag,
       T - SPT-bit set, J - Join SPT, M - MSDP created entry, E - Extranet,
       X - Proxy Join Timer Running, A - Candidate for MSDP Advertisement,
       U - URD, I - Received Source Specific Host Report,
       Z - Multicast Tunnel, z - MDT-data group sender,
       Y - Joined MDT-data group, y - Sending to MDT-data group,
       G - Received BGP C-Mroute, g - Sent BGP C-Mroute,
       N - Received BGP Shared-Tree Prune, n - BGP C-Mroute suppressed,
       Q - Received BGP S-A Route, q - Sent BGP S-A Route,
       V - RD & Vector, v - Vector, p - PIM Joins on route,
       x - VxLAN group
Outgoing interface flags: H - Hardware switched, A - Assert winner, p - PIM
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(*, 239.0.0.1), 00:02:44/stopped, RP 10.2.2.2, flags: SPF
  Incoming interface: Serial1/0, RPF nbr 10.0.12.2
  Outgoing interface list: Null

(172.16.1.1, 239.0.0.1), 00:02:44/00:03:25, flags: FT
  Incoming interface: Ethernet0/3, RPF nbr 0.0.0.0
  Outgoing interface list:
    Serial1/0, Forward/Sparse-Dense, 00:02:02/00:02:44
```

R4

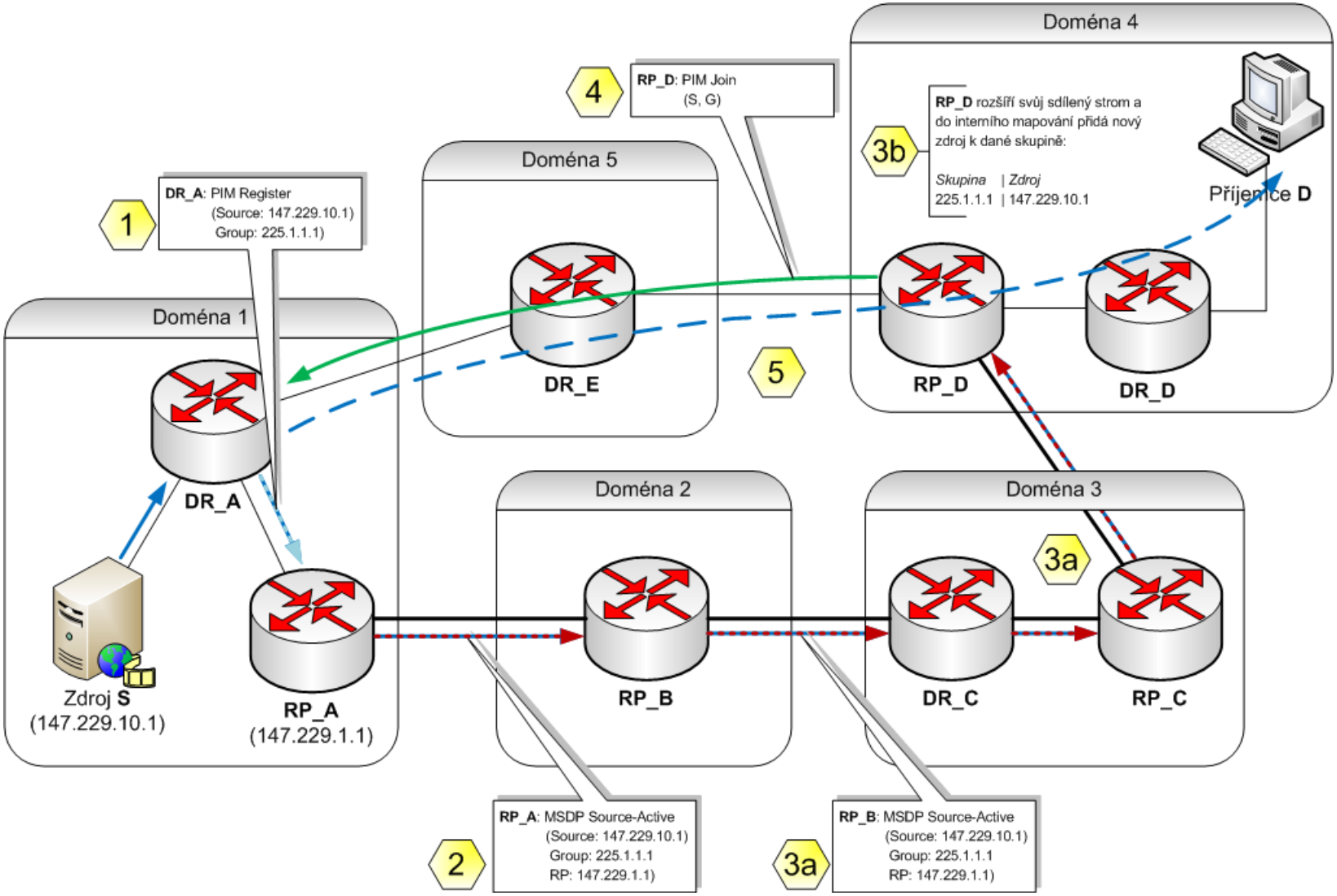
```
R4#sh ip mroute
IP Multicast Routing Table
Flags: D - Dense, S - Sparse, B - Bidir Group, s - SSM Group, C - Connected,
       L - Local, P - Pruned, R - RP-bit set, F - Register flag,
       T - SPT-bit set, J - Join SPT, M - MSDP created entry, E - Extranet,
       X - Proxy Join Timer Running, A - Candidate for MSDP Advertisement,
       U - URD, I - Received Source Specific Host Report,
       Z - Multicast Tunnel, z - MDT-data group sender,
       Y - Joined MDT-data group, y - Sending to MDT-data group,
       G - Received BGP C-Mroute, g - Sent BGP C-Mroute,
       N - Received BGP Shared-Tree Prune, n - BGP C-Mroute suppressed,
       Q - Received BGP S-A Route, q - Sent BGP S-A Route,
       V - RD & Vector, v - Vector, p - PIM Joins on route,
       x - VxLAN group
Outgoing interface flags: H - Hardware switched, A - Assert winner, p - PIM
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(*, 239.0.0.1), 00:02:00/00:03:28, RP 10.2.2.2, flags: S
  Incoming interface: Serial1/2, RPF nbr 10.0.24.1
  Outgoing interface list:
    Ethernet0/1, Forward/Sparse-Dense, 00:01:55/00:02:33
    Ethernet0/2, Forward/Sparse-Dense, 00:02:00/00:03:28
```

Multicast Source Discovery Protocol

- RFC 3618
- Addresses need to pass multicast between different AS
- MSDP notifies about existing multicast sources
- MSDP peering interconnects domain RP

- *However, global multicast is unfulfilled dream despite all technologies involved!*

MSDP: Demo

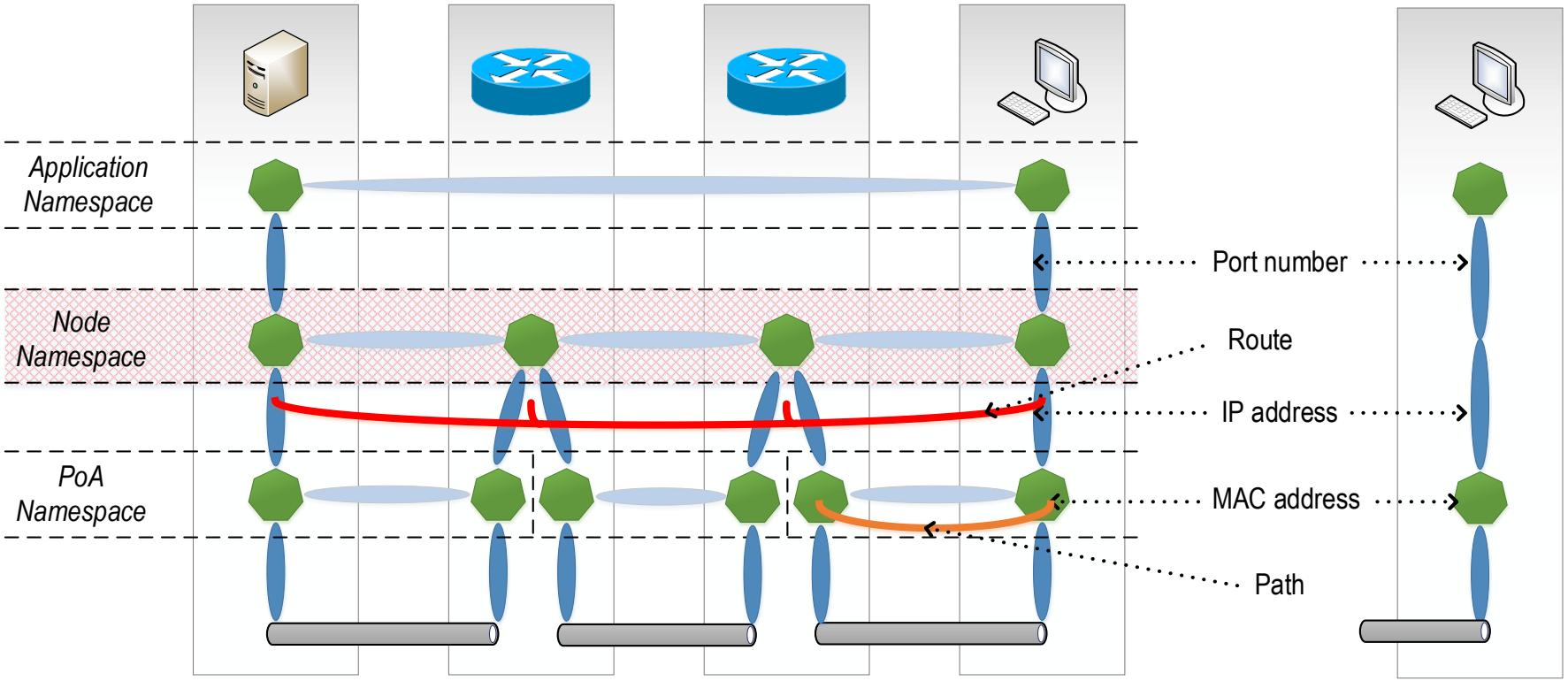

Doména 4
RP_D rozšíří svůj sdílený strom a do interního mapování přidá nový zdroj k dané skupině:
Skupina | Zdroj
225.1.1.1 | 147.229.10.1
3b
4
RP_D: PIM Join
(S, G)
Příjemce D
Doména 5
1
DR_A: PIM Register
(Source: 147.229.10.1)
Group: 225.1.1.1)
DR_E
5
RP_D
DR_D
Doména 1
DR_A
Doména 2
Doména 3
3a
Zdroj S
(147.229.10.1)
RP_A
(147.229.1.1)
RP_B
DR_C
RP_C
2
RP_A: MSDP Source-Active
(Source: 147.229.10.1)
Group: 225.1.1.1
RP: 147.229.1.1)
3a
RP_B: MSDP Source-Active
(Source: 147.229.10.1)
Group: 225.1.1.1
RP: 147.229.1.1)

Summary

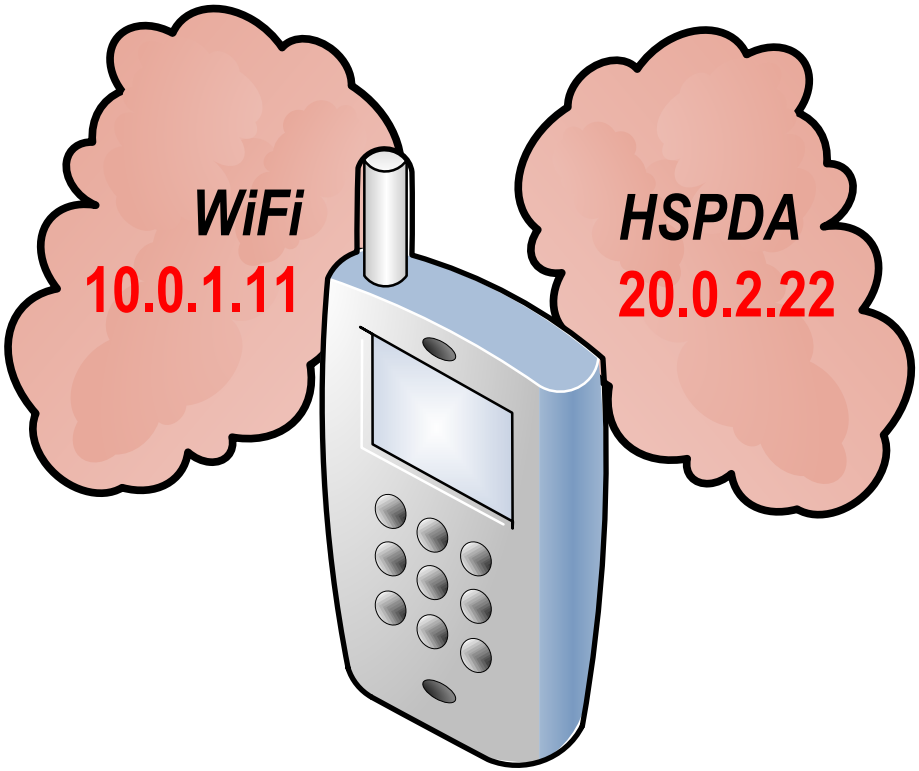
“The Internet is at its core an unfinished demo.”

John Day

Broken Addressing Model



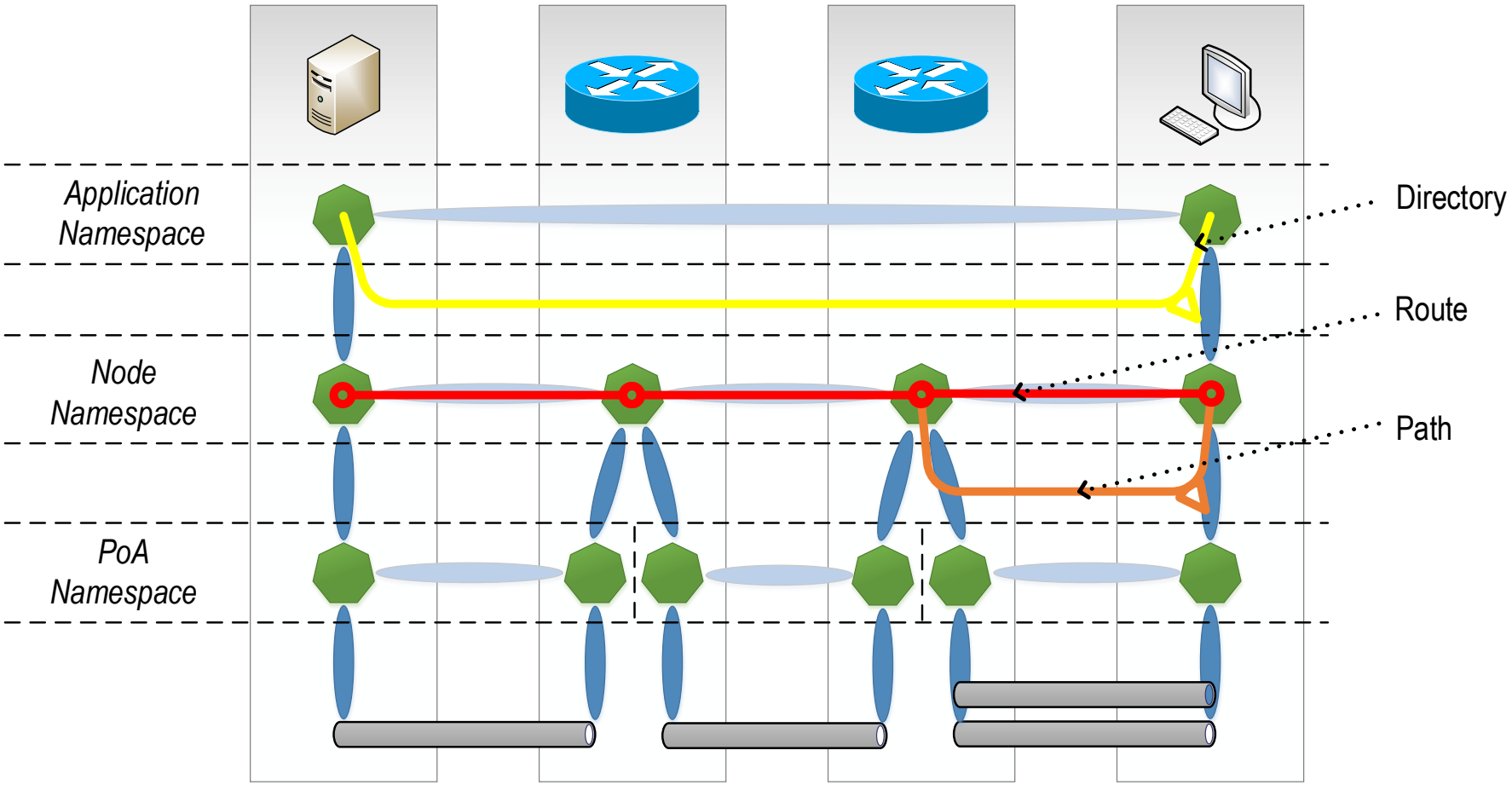
Implications



What is unique address???

Complete Addressing

- RFC 1498



Self-Check

Self-Check

- What is the advantage of using hierarchical addresses over flat ones?
- Does IP address identifies node or node interface?
- How distance-vector routing protocols models network topology?
- How link-state routing protocols models network topology?
- What is a difference between distance-vector and path-vector protocols?
- What is a difference between distance-vector and link-state routing protocols?
- What is a difference between unicast and multicast IP routing?
- Illustrate Bellman-Ford algorithm on example!
- Illustrate Dijkstra's algorithm on example!
- Compare RIP with EIGRP and Babel!
- Compare OSPF with IS-IS!

Bibliography

References

- EWD 687a: Termination detection for diffusing, Information Processing Letters 11, 1980 http://www.cs.utexas.edu/users/EWD/ewd06xx/EWD687a.PDF
- Garcia-Lunes-Aceves, J. J.: Loop-Free Routing Using Diffusing Computations. IEEE/ACM Transactions on Networking Vol. I(No. 1), 130-141 (1993)
- Albrightson, R., Garcia-Luna-Aceves, J., Boyle, J.: EIGRP a fast routing protocol based on distance vectors. Proceedings Networld/Interop Vol. XCIV, 136-147 (May 1994)
- Tomáš Fidler, Směrování Internetu, Archiv předmětu PDS rok 2011, pds-07-bgp.pdf
- Shivkumar Kalyanaraman, Rensselaer Polytechnic Institute, http://deptinfo.cnam.fr/Enseignement/CycleSpecialisation/IRE/cours%20bgp.pdf
- *Computer Networking: A Top Down Approach Featuring the Internet*. Jim Kurose, Keith Ross Addison-Wesley.
- https://www.inetzero.com/isis-training-and-junos-configuratio/
- Julius Chroboczek, Babel A flexible routing protocol, https://www.irif.fr/~jch/software/babel/babel-20140311.pdf