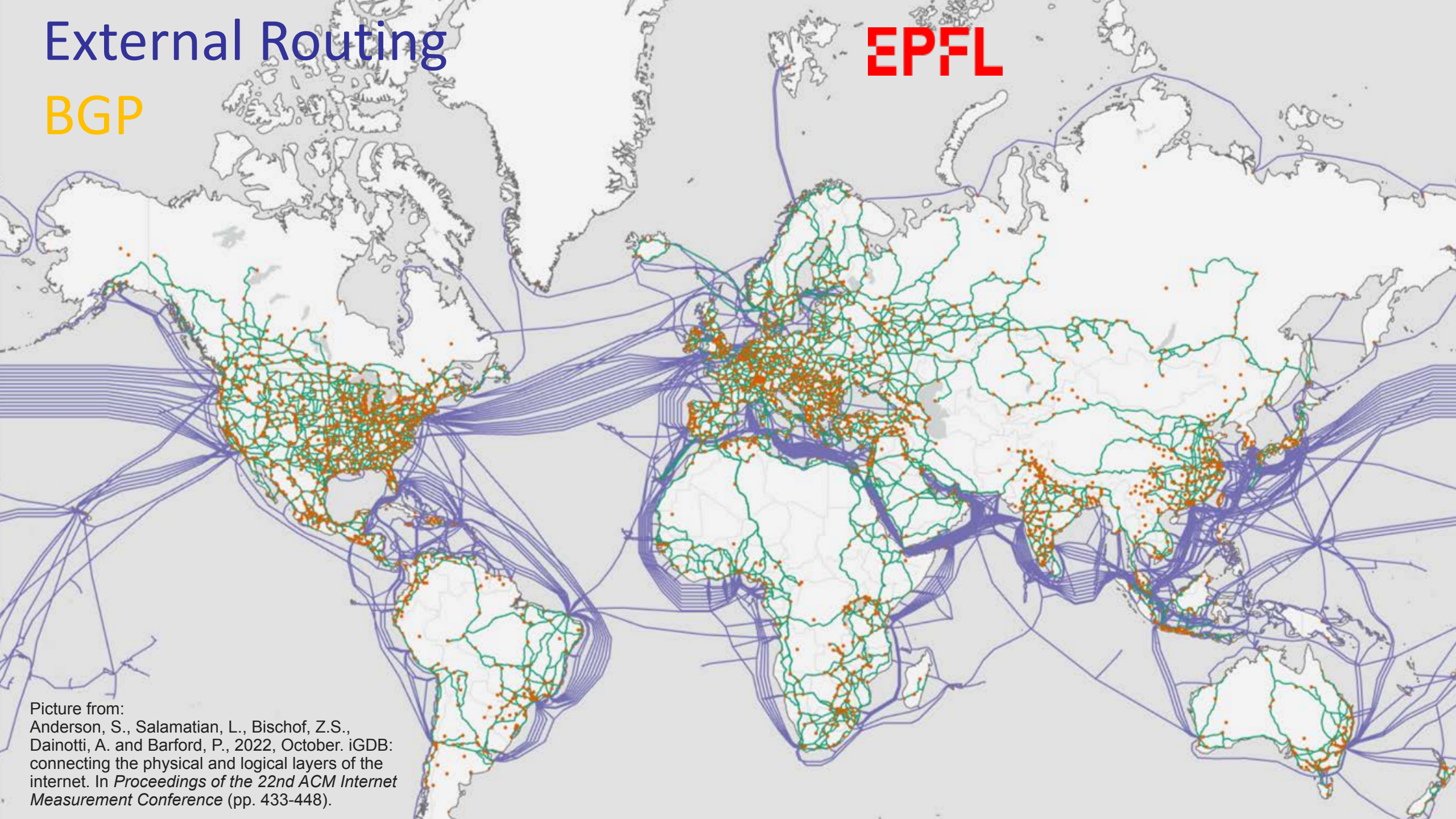


External Routing

BGP

EPFL



Picture from:
Anderson, S., Salamatian, L., Bischof, Z.S.,
Dainotti, A. and Barford, P., 2022, October. iGDB:
connecting the physical and logical layers of the
internet. In *Proceedings of the 22nd ACM Internet
Measurement Conference* (pp. 433-448).

Contents

A. BGP at a high level

1. Inter-Domain Routing
2. Policy Routing

B. BGP in detail

1. How it works
2. Aggregation
3. Interaction BGP—IGP—Packet Forwarding
4. Other Attributes
5. Bells and Whistles
6. Security of BGP

C. Illustrations and Statistics

Recall: routing algorithms differ in at least 3 aspects

Nature of “best” path — i.e. what is optimization objective of an algorithm?

- to use shortest path
- to use equal-cost multi-path
- to respect policies
- arbitrary

Scope of network — i.e. what is the underlying network? is topology info available?

- single domain —> *intra-domain* routing (main alg. is OSPF)
- multiple domains —> *inter-domain* routing (main alg. is BGP)

A *domain* is a network under the *same* administrative entity (e.g. a campus network, an enterprise network, or an ISP, etc.)

State location — i.e. where is the output (i.e. the routing information) finally stored?

- inside a local forwarding table
- directly into the packet headers

Domains — terminology

ARD = Autonomous Routing Domain = routing domain under a *single administrative entity*

AS = Autonomous System = ARD with a *number* (“AS number”), used in BGP routes

- AS number is 32 bits, written in 2-field dotted decimal notation: e.g. 23.3456, and leading zeros may be omitted: e.g. 0.559 means 559
- Private AS numbers are: 0.64512 – 0.65535
- Real examples: AS1942 - CIGG-GRENOBLE, AS2200 - Renater
AS559 - SWITCH Teleinformatics Services

ARDs can be:

transit (see B and D),

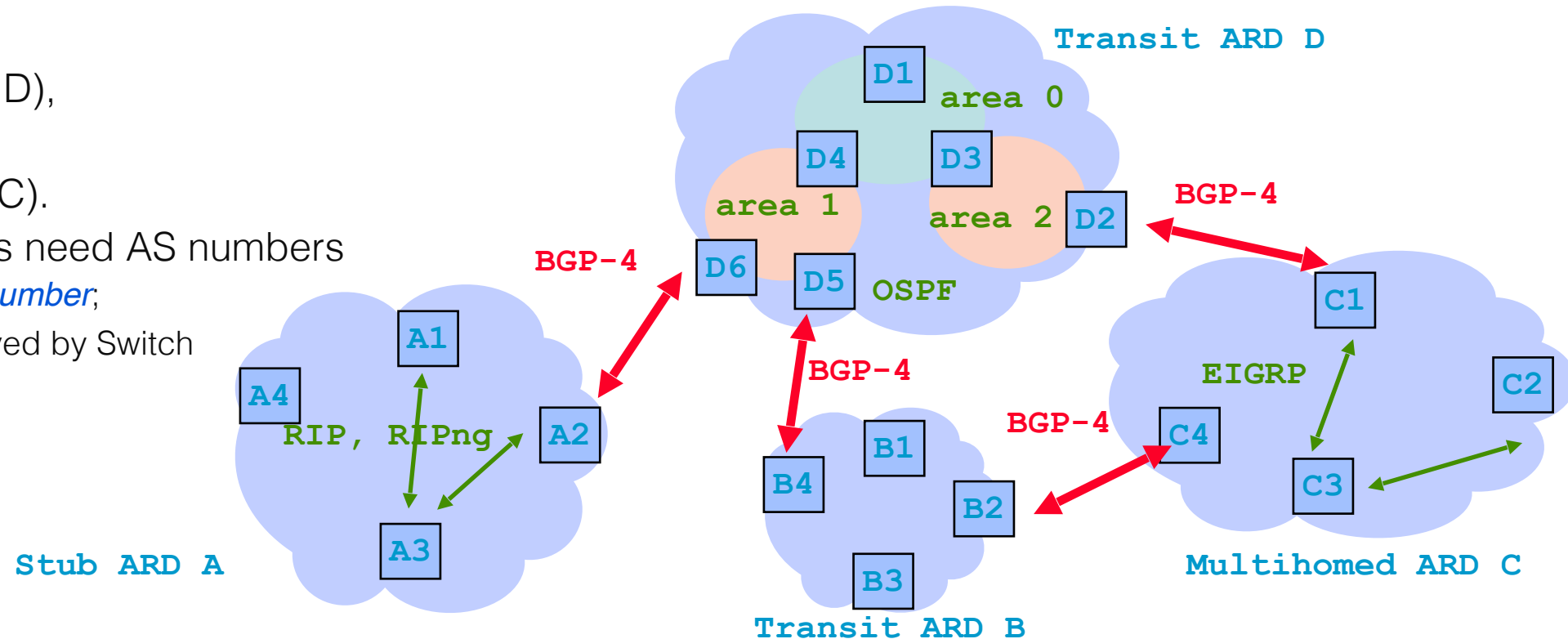
stub (see A) or

multi-homed (see C).

Only non-stub domains need AS numbers

e.g. EPFL: ARD *w/o number*;

all external traffic served by Switch



Part A: BGP at high level

1. Inter-Domain Routing

Context

The Internet is *too large* + *heterogeneous* (i.e. it is split into various domains) to be run by one routing protocol.

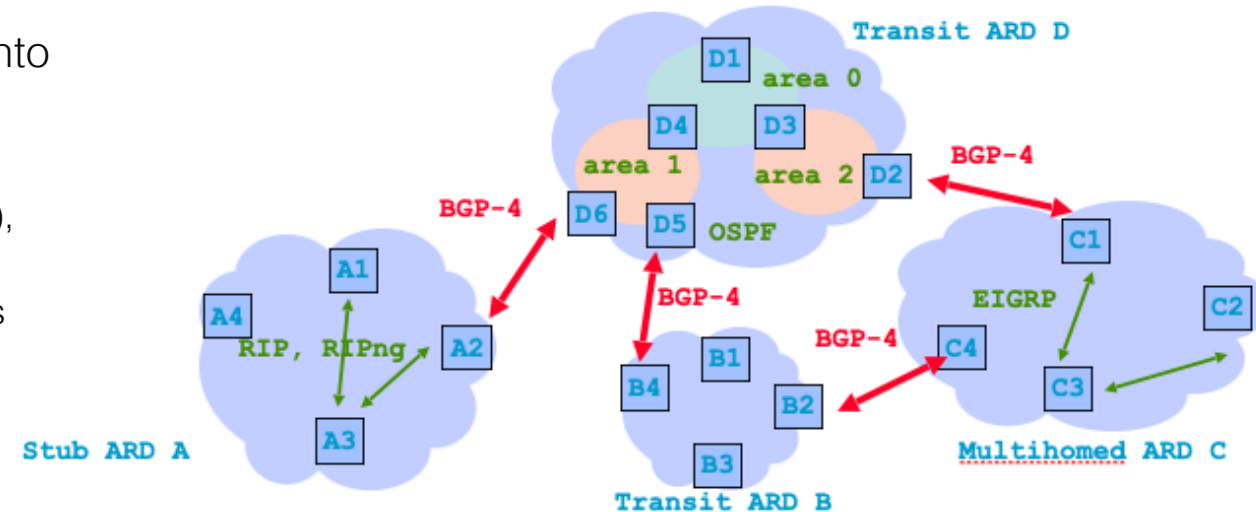
We use hierarchical routing instead:

- *within* domains, we use an **IGP** (= Internal Gateway Protocol), e.g. RIP, OSPF (standard), IGRP (Cisco)
 - with OSPF: large domains are further split into Areas
- *between* domains, we use **BGP** (= Border Gateway Protocol)

What is the goal of BGP?

- Compute paths from a border router in one domain to any network prefix in the world
- Handle both IPv4 and IPv6 addresses in a single process

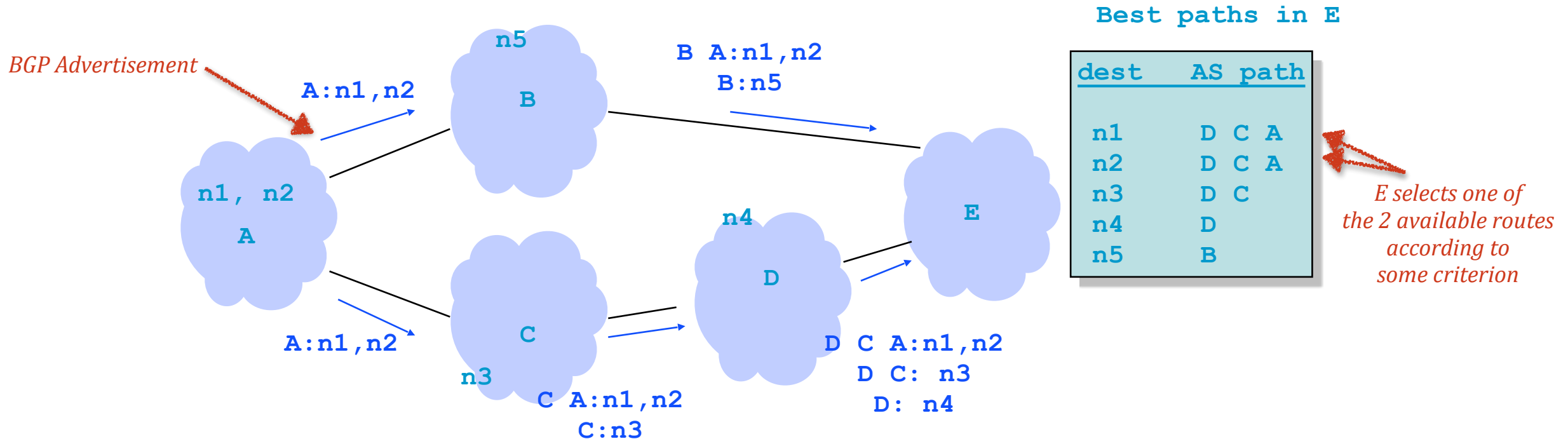
How does it achieve it? via *path-vector routing* and *policies*



Path Vector Routing (high-level example)

Goal: To compute best AS-level routes/paths.

How? ASes *advertize* to their neighbor ASes their *best routes* to destinations, by *prepending* its AS number to the routes they export. Each AS uses its *own criteria* for deciding which path is the best.



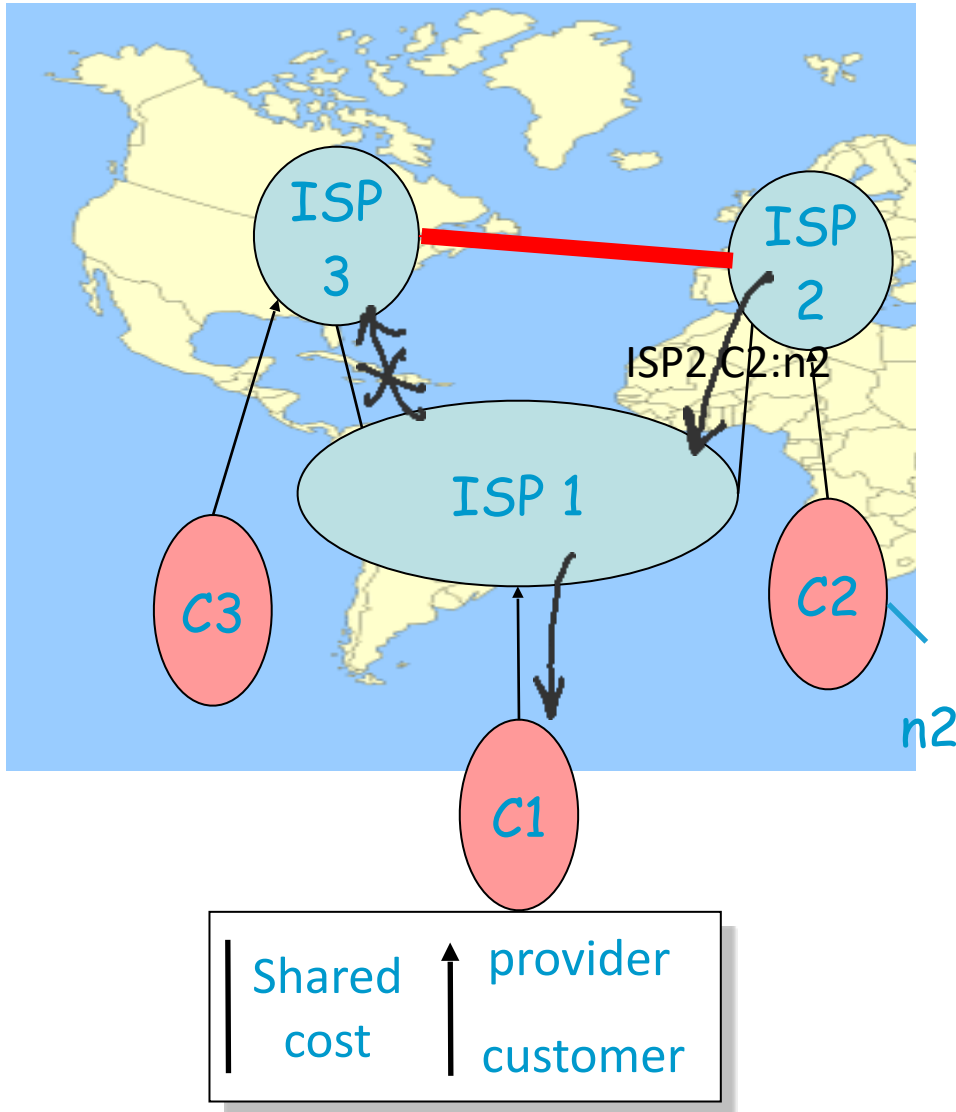
Policies...

...implement domains' business agreements
(e.g. customer-provider relationships, shared-cost peering)

via:

import (what to accept) and export rules (what to advertize to whom),
and a decision process (what is the best route to each destination)

Policies (high-level example)



Suppose:

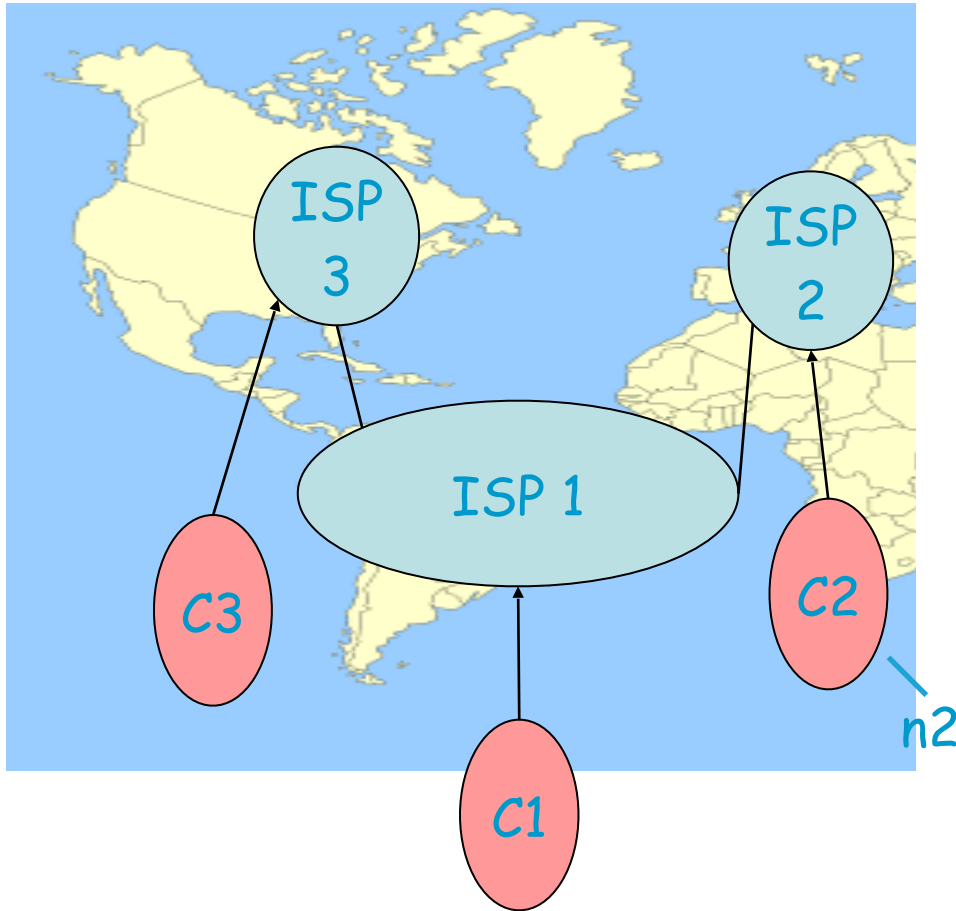
- All ISPs are shared-cost peers; C_i is customer of ISP_i .
- ISP3-ISP2 is a transatlantic link, cost-shared between ISP2 & ISP3, but it is expensive;
- ISP3-ISP1 is a local, inexpensive link;
- Problem: It is advantageous for ISP3 to send traffic to n_2 via ISP1; but...
ISP1 may not agree to carry traffic from C_3 to C_2 .
How can ISP1 apply such a *policy*:
 - “transit service” to C_1 and
 - “non-transit” service to ISP2 & ISP3 ?

A common policy rule is:

*“Routes learnt from peers or providers are **not advertized** to peers or providers.”*

Applying this to our example:

- ISP1 advertizes the route: {ISP2 C2:n2} to C1
- but not to ISP3
 - because doing so would allow ISP3 to find a route to C2 that transits via ISP1



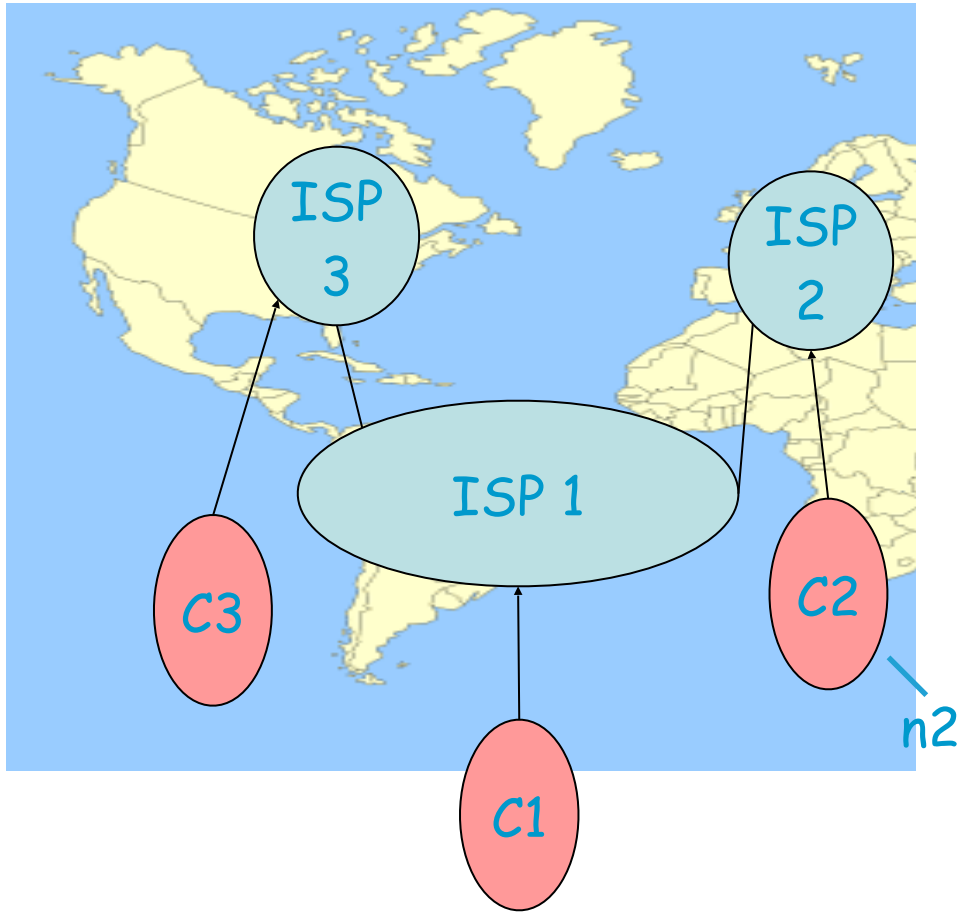
ISP1-ISP2 and ISP1-ISP3 are peers;
ISP2-ISP3 are *not* peers nor customers/providers.
All apply the rule “Routes coming from peers or
providers are not propagated to peers or providers”.
What is a valid path from C2 to C3 ?

- A. C2-ISP2-ISP1-ISP3-C3
- B. None
- C. I don't know



Go to web.speakup.info
or
download speakup app
Join room
46045

Solution



Answer B

ISP1 learns the route ISP1-ISP2-C2-n2 but refuses to announce it to ISP3 (who is a peer)

this network is **partitioned** !

Solution: internet backbone providers (eg. AT&T, OpenTransit, Orange etc, called **tier-1**):

must all exchange traffic with each other
and

all ISPs need to be connected to a tier-1

Part B.

1. How does BGP work?

- BGP routers talk to each other over TCP connections
- Each BGP router [BGP-4, RFC 4271]:
 - receives and stores *candidate routes* from its BGP neighbor peers, after applying *import* policy rules
 - applies the *decision process* to *select at most one route* per destination prefix and keeps all other accepted routes as *backup*
 - exports the *selected* routes to BGP neighbors, after applying *export* policy rules and possibly *aggregation*
- Routes are advertised via UPDATE messages that contain *only modifications*: *new paths or withdrawals*
- Other BGP messages are:
OPEN (=sync after boot-up), NOTIFICATION (= reset), KEEPALIVE (= notify BGP peers that router is running)

2 types of BGP (e-BGP and i-BGP)

A router that runs BGP is called a *BGP speaker*

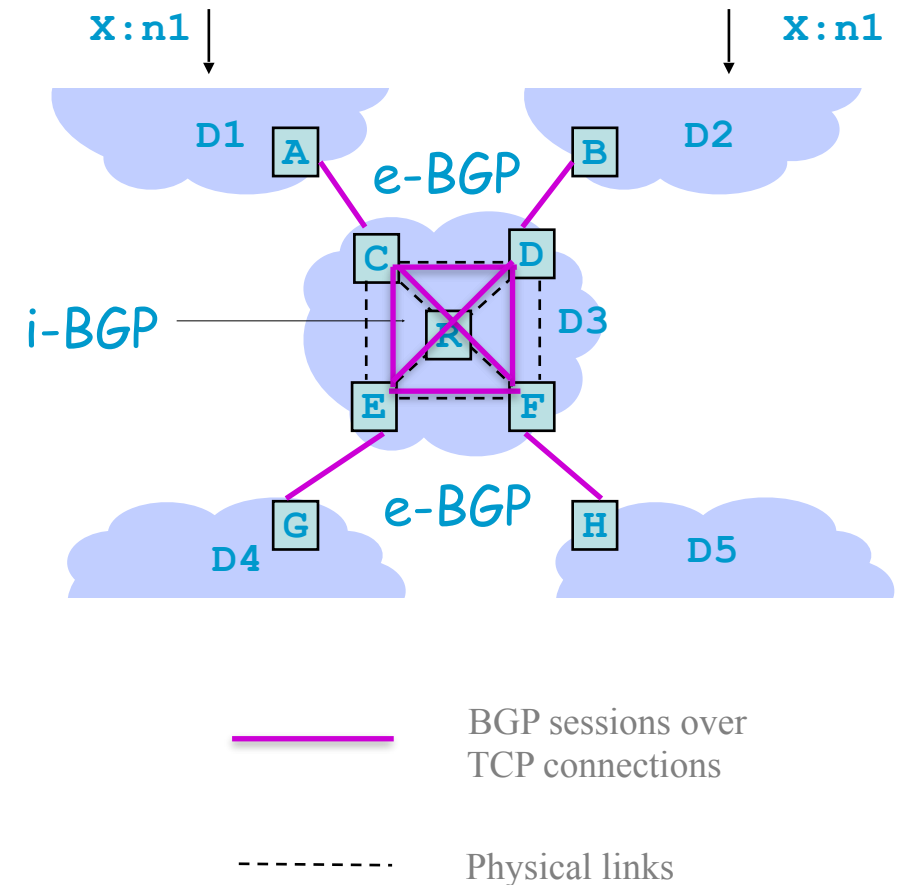
- at the **border** between 2 ARDs, there are 2 speakers, one in each ARD
- inside an ARD, there are usually several BGP speakers

BGP speakers speak:

- *externally (e-BGP)*
to advertize routes to neighbor domains [as in a previous slide]
- *internally (i-BGP)*
to exchange what they have learnt from e-BGP

In i-BGP, BGP *peers*

- communicate via a mesh network, a.k.a. “**BGP mesh**”
- advertize routes as in e-BGP but *within the domain*; so they do **not**:
 - repeat the routes learnt from i-BGP —> to avoid redundant traffic
 - prepend own AS number over i-BGP
 - modify the “NEXT-HOP” attribute of a route [see also later]
- know about all inter-domain link subnets via IGP



Say what is always true

- A. 1
 - B. 2
 - C. 1 and 2
 - D. None
 - E. I don't know
1. Two BGP peers must be connected by a TCP connection.
 2. Two BGP peers must be “on link” (on the same subnet)



Go to web.speakup.info
or
download speakup app
Join room
46045

Solution

Answer A

BGP peers communicate (typically) with TCP.

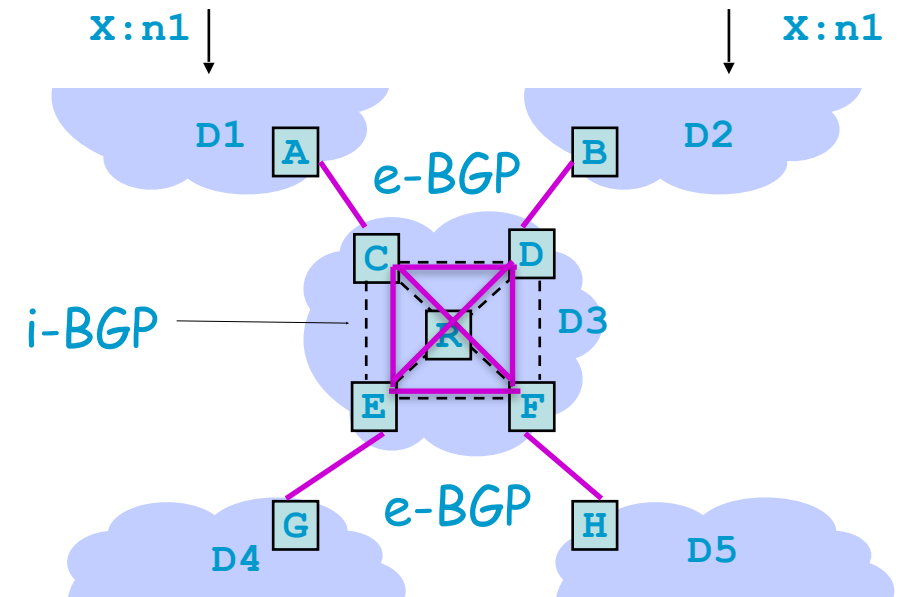
External peers are typically “on link”.

Internal peers need not be “on link”.

Which BGP updates may be sent ?

- A. 1
- B. 2
- C. 3
- D. 1 and 2
- E. 1 and 3
- F. 2 and 3
- G. All
- H. None
- I. I don't know

1. $C \rightarrow A : D3 - D2 - X : n1$
2. $D \rightarrow E : D2 - X : n1$
3. $C \rightarrow E : D2 - X : n1$



———— BGP sessions over TCP connections
----- Physical links



Go to web.speakup.info
or
download speakup app
Join room
46045

Solution

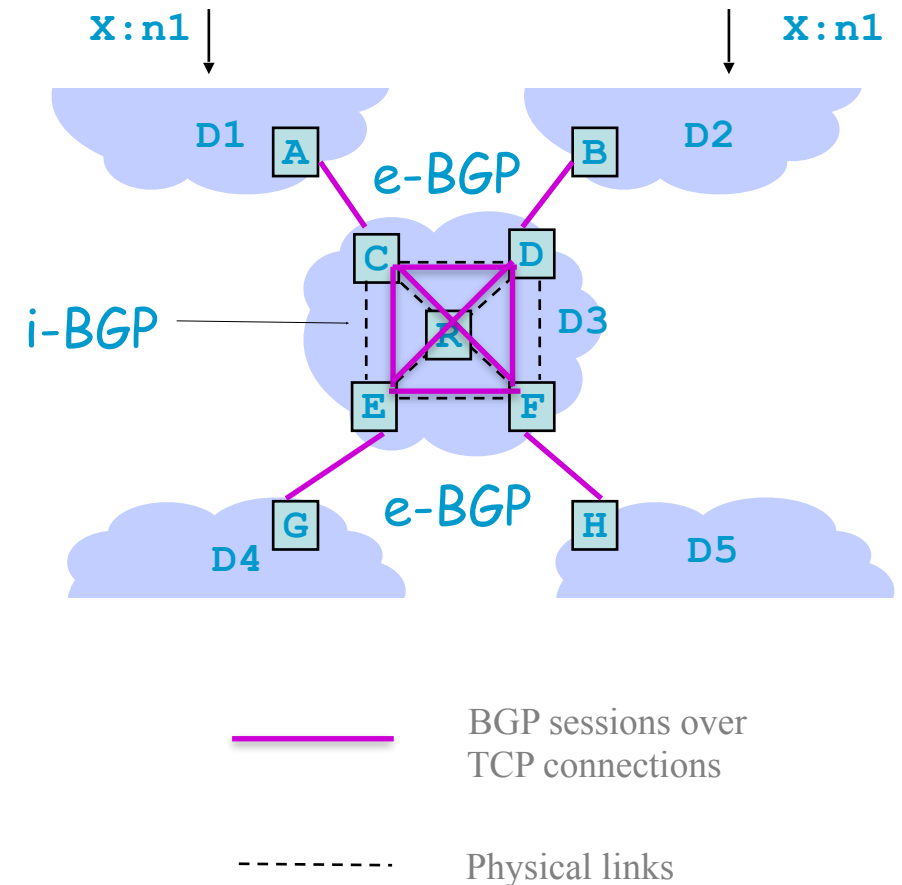
1. $C \rightarrow A : D3 - D2 - X : n1$
2. $D \rightarrow E : D2 - X : n1$
3. $C \rightarrow E : D2 - X : n1$

Answer D.

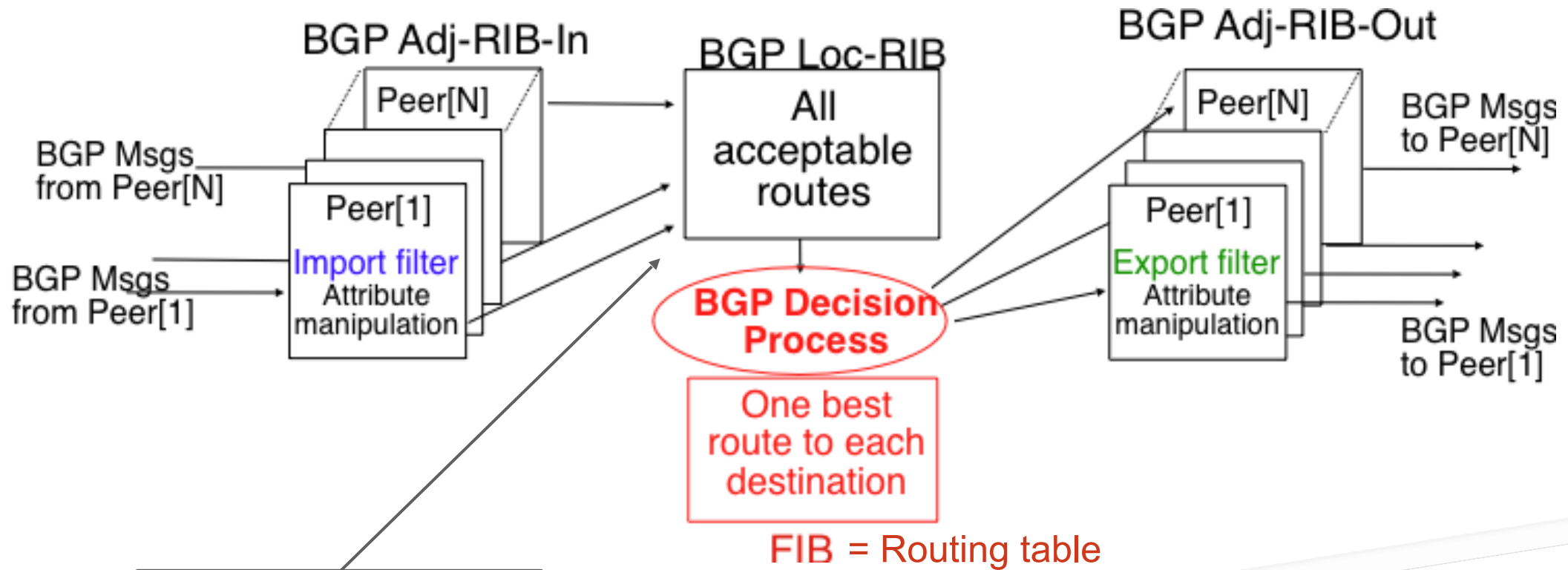
The route $C \rightarrow E : D2 - X : n1$ was learnt by C from D , i.e. via internal BGP (i-BGP).

Therefore it should *not be re-advertized* over i-BGP. There is no need since all other routers inside the domain have learnt this route from D .

Only routes 1 and 2 should be repeated.



Operation of a BGP Router

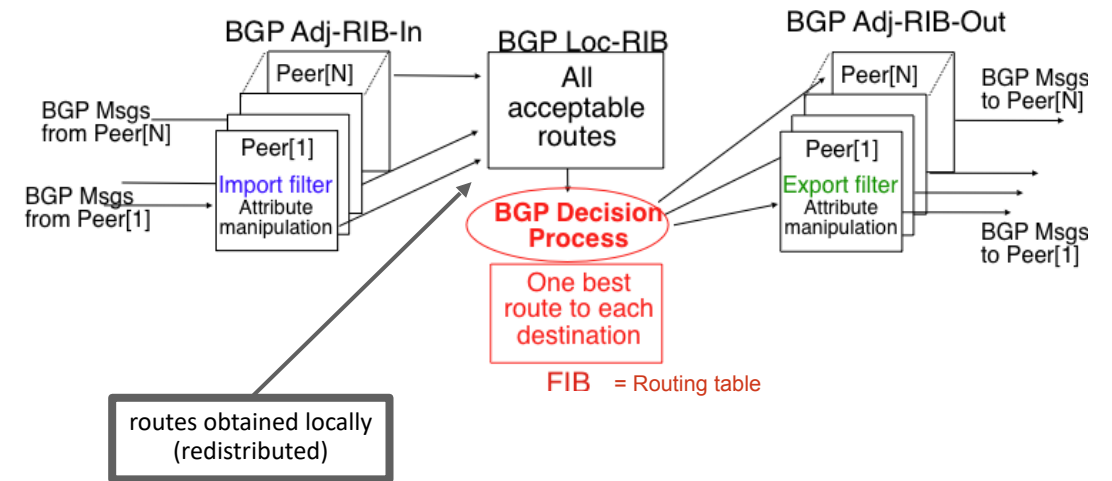


- Each BGP router [BGP-4, RFC 4271]:
 - receives and stores **candidate routes** from its BGP neighbor peers, after applying **import** policy rules
 - applies the **decision process** to **select at most one route** per dest prefix and keeps all other accepted routes as **backup**
 - exports the **selected** routes to BGP neighbors, after applying **export** policy rules and possibly **aggregation**

Routes, RIBs, Routing Table

A **route** has several attributes:

- **destination** (subnet) prefix
- **path** to the destination
(AS-PATH or an authenticated BGPsec_Path)
- **NEXT-HOP** (modified by e-BGP, left unchanged by i-BGP)
- **ORIGIN**: route learnt from IGP, BGP, static
- Other attributes:
 LOCAL-PREF,
 ATOMIC-AGGREGATE (= route cannot be dis-aggregated),
 MED, etc. [see later]



Routes + their attributes are stored in the **Routing Information Bases (RIBs)**:

Adj-RIB-in, Loc-RIB, Adj-RIB-out.

Like any IP host or router, a BGP router also has a **Routing Table** = IP forwarding table

Used for packet forwarding, in real time

The Decision Process

The decision process chooses *at most one route* to each different destination *prefix* as best

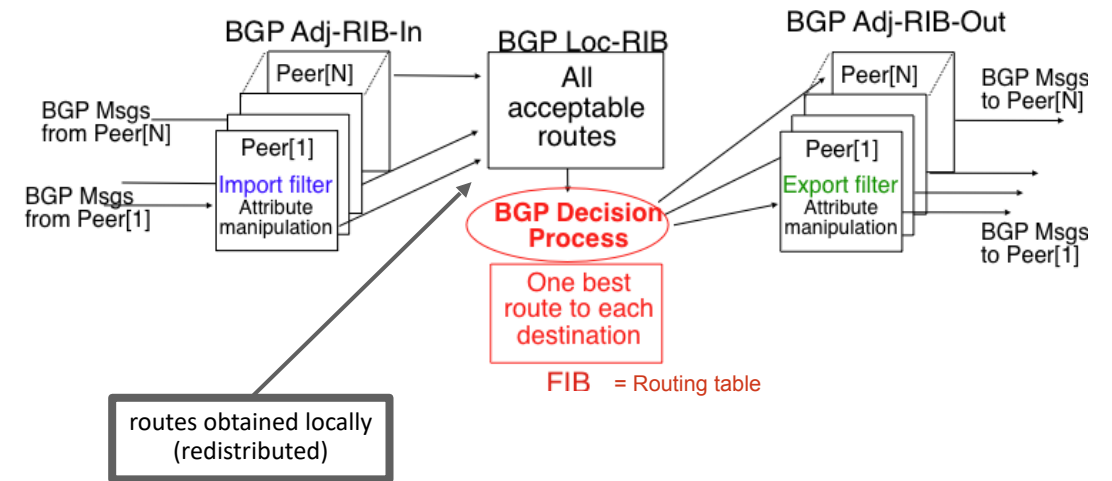
e.g.: only one route to 2.2/16 can be chosen,

but there can be different routes to 2.2.2/24 and 2.2/16

How?

- A route can be selected only if its next-hop is **reachable**
- For each dest prefix, all acceptable routes are compared w.r.t. their **attributes** using a **sequence of criteria** (until only one route remains); a common sequence is:
 0. Highest weight (Cisco proprietary)
 1. Highest LOCAL-PREF
 2. Shortest AS-PATH
 3. Lowest MED, if taken seriously by this network
 4. e-BGP > i-BGP (= if route is learnt from e-BGP, it has priority)
 5. Shortest path to NEXT-HOP, according to IGP
 6. Lowest BGP identifier (router-id of the BGP peer from whom route is received)

(The Cisco and FRR implementation of BGP, used in lab 6, have additional cases, not shown here)

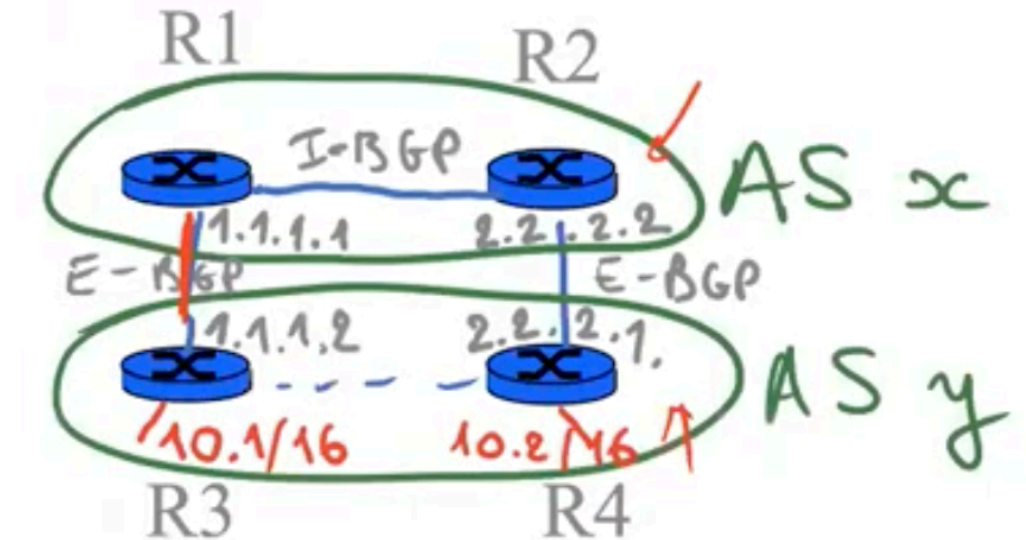


The result of the decision process is stored in forwarding table and in Adj-RIB-out (*one route per destination for each BGP peer*).

The router sends updates when Adj-RIB-out **changes** (addition or deletion) after applying **export rules**.

Fundamental Example

- 4 BGP routers communicate directly (solid lines) or indirectly (dash lines) via e-BGP or i-BGP,
- 2 ASes, x and y, each one running its own IGP, too.
- Assume R3 and R4 are configured to advertise both prefixes of y.



➡ We focus on R1 and show its BGP information:

Remarks:

- we show next only a subset of the route attributes (such as : destination, path, NEXT-HOP)
- the exact internal topology of y is not shown

Step 1

R3 —> R1

10.1/16 AS = y

10.2/16 AS = y

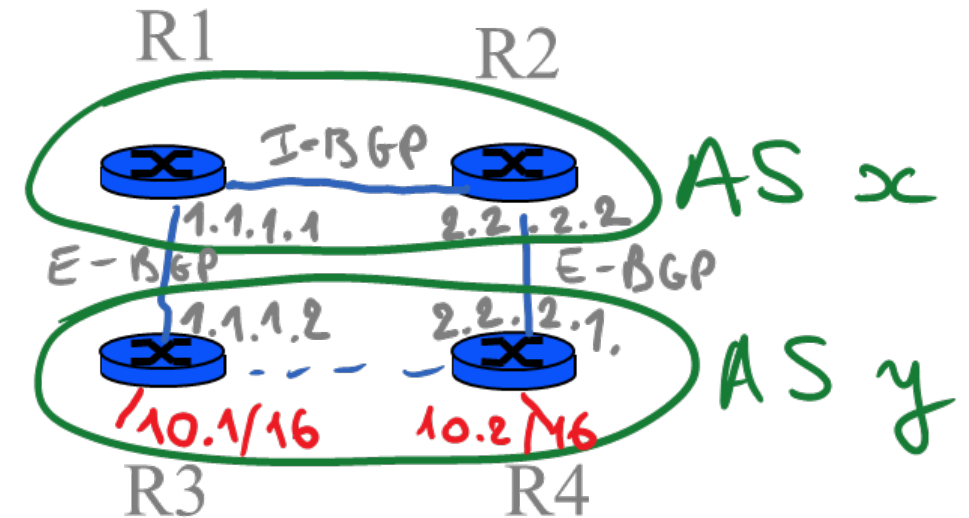
Adj-RIB-in

| | | |
|---------|---------------------------------|------|
| From R3 | 10.1/16 AS = y NEXT-HOP=1.1.1.2 | Best |
| From R3 | 10.2/16 AS = y NEXT-HOP=1.1.1.2 | Best |

Adj-RIB-out

| | |
|-------|---------------------------------|
| To R2 | 10.1/16 AS = y NEXT-HOP=1.1.1.2 |
| To R2 | 10.2/16 AS = y NEXT-HOP=1.1.1.2 |

- [import filters:] R1 accepts the updates and stores them in Adj-RIB-In
- [Decision Process:] R1 designates these routes as best routes
- [export filters:] R1 puts updates into Adj-RIB-Out, which will cause them to be sent to other BGP neighbors/peers



Step 2

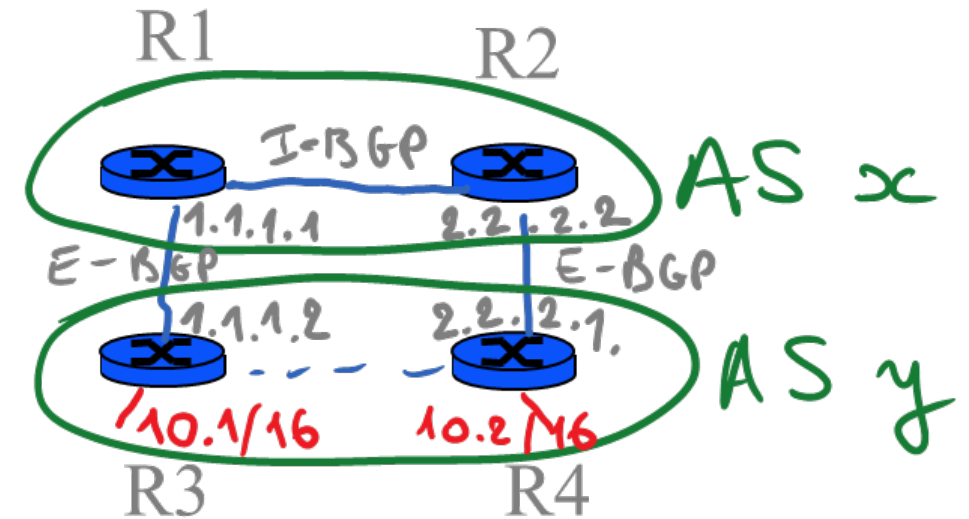
R2 → R1

10.1/16 AS = y NEXT-HOP = 2.2.2.1

10.2/16 AS = y NEXT-HOP = 2.2.2.1

Adj-RIB-in

| | | |
|---------|-----------------------------------|------|
| From R3 | 10.1/16 AS = y NEXT-HOP = 1.1.1.2 | Best |
| From R2 | 10.1/16 AS = y NEXT-HOP = 2.2.2.1 | |
| From R3 | 10.2/16 AS = y NEXT-HOP = 1.1.1.2 | Best |
| From R2 | 10.2/16 AS = y NEXT-HOP = 2.2.2.1 | |



Which of the two new routes (in red) are promoted by the decision process to “best routes” assuming WEIGHT, LOCAL_PREF and MED are empty?

- A. The first one only
- B. The second one only
- C. Both
- D. None
- E. I don't know

0. Highest weight (Cisco proprietary)
1. Highest LOCAL_PREF
2. Shortest AS-PATH
3. Lowest MED, if taken seriously by this network
4. e-BGP > i-BGP (= if route is learnt from e-BGP, it has priority)
5. Shortest path to NEXT-HOP, according to IGP
6. Lowest BGP identifier (router-id of the BGP peer from whom rc
(The Cisco and FRR implementation of BGP, used in lab 6, have additional

Step 2

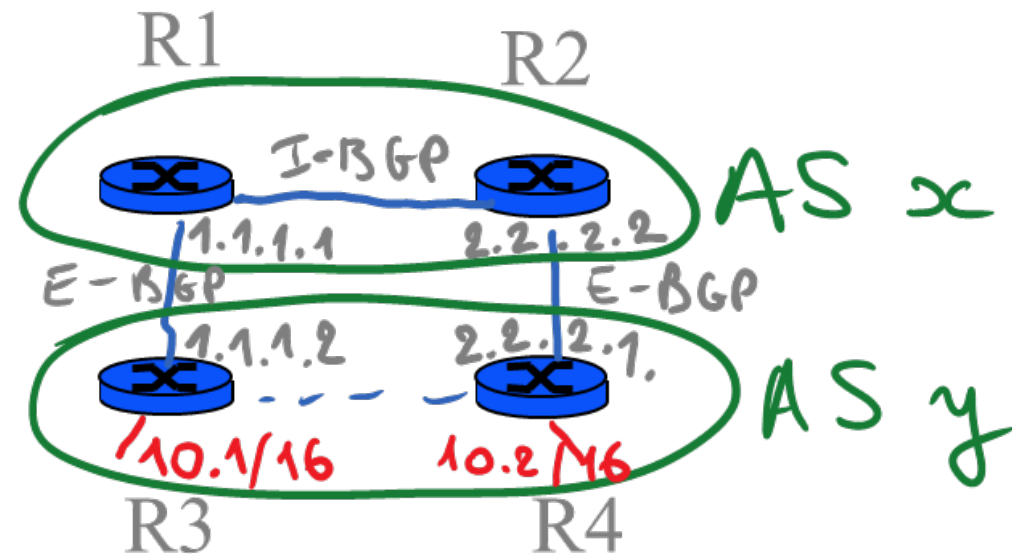
R2 → R1

10.1/16 AS=y NEXT-HOP=2.2.2.1

10.2/16 AS=y NEXT-HOP=2.2.2.1

Adj-RIB-in

| | | |
|---------|-------------------------------|------|
| From R3 | 10.1/16 AS=y NEXT-HOP=1.1.1.2 | Best |
| From R2 | 10.1/16 AS=y NEXT-HOP=2.2.2.1 | |
| From R3 | 10.2/16 AS=y NEXT-HOP=1.1.1.2 | Best |
| From R2 | 10.2/16 AS=y NEXT-HOP=2.2.2.1 | |



Answer D

R1 applies again its decision process. Now it has several possible routes to each prefix.

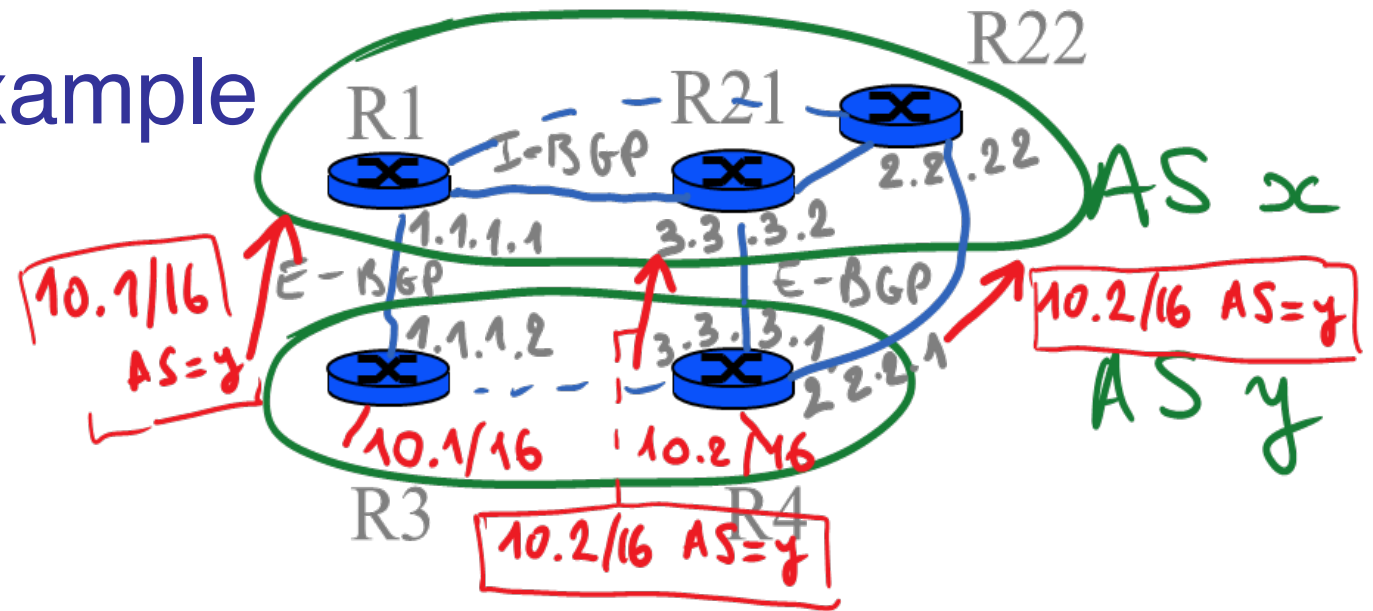
The first applicable rule in the decision process (slide “The Decision Process”) says that if a route is learnt from e-BGP it has precedence over a route learnt from i-BGP (**e-BGP > i-BGP**).

Since all routes in **Adj-RIB-In** from R2 are learnt from i-BGP, and all routes in **Adj-RIB-In** from R3 are learnt from e-BGP, the winners are the latter, so there is no change.

Since there is no change in **Loc-RIB** there is **no change** in **Adj-RIB-Out** and therefore no message is sent by R1.

Another Fundamental Example

- 3 BGP routers in AS x.
- An IGP (e.g. OSPF) also runs on R1, R21 and R22.
- Assume:
 - all link costs are equal to 1.
 - R3 and R4 advertise **only** their directly attached prefixes, as shown in the figure.



➡ We focus on R1 and show its BGP information:

Note:

The 3 BGP in AS x routers must have TCP connections with each other (same in AS y, but not shown on figure).

Step 1

R3 → R1
10.1/16 AS = y

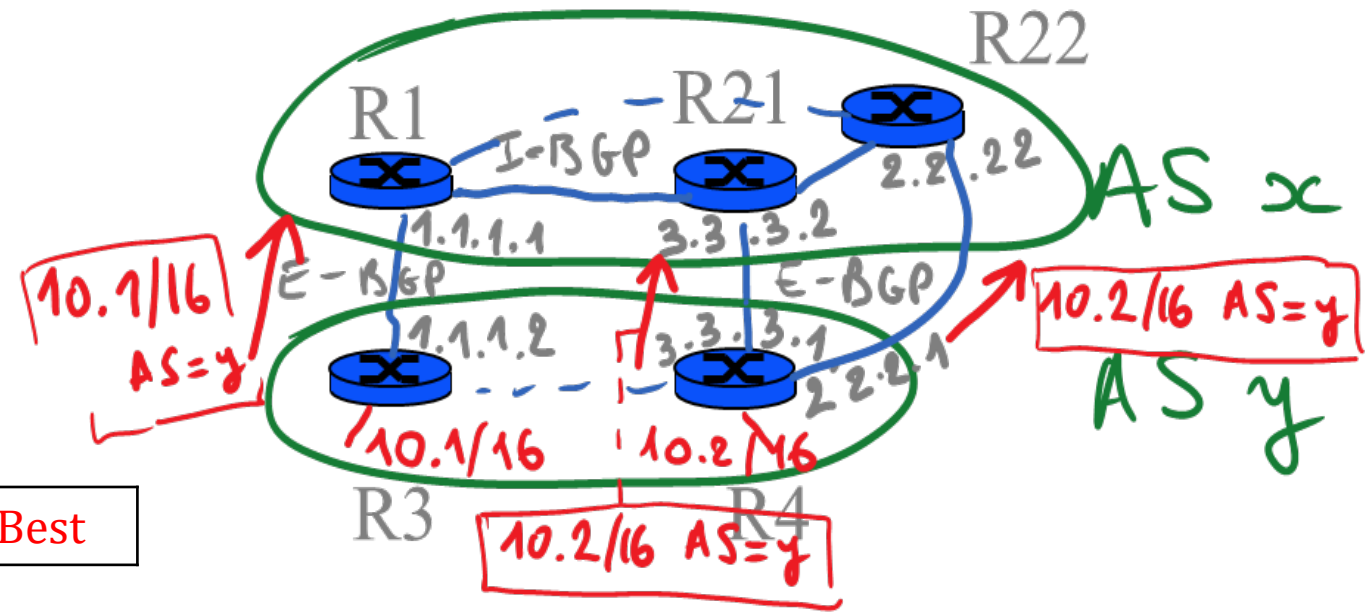
Adj-RIB-in

| | | |
|---------|---------------------------------|------|
| From R3 | 10.1/16 AS = y NEXT-HOP=1.1.1.2 | Best |
|---------|---------------------------------|------|

Adj-RIB-out

| | |
|--------|---------------------------------|
| To R21 | 10.1/16 AS = y NEXT-HOP=1.1.1.2 |
| To R22 | 10.1/16 AS = y NEXT-HOP=1.1.1.2 |

- R1 accepts the updates and stores it in Adj-RIB-In
- R1 designates this route as best route
- R1 puts route into Adj-RIB-Out, which will cause them to be sent to BGP neighbors R21 and R22



Step 2

R22 —> R1

10.2/16 AS =y NEXT-HOP=2.2.2.1

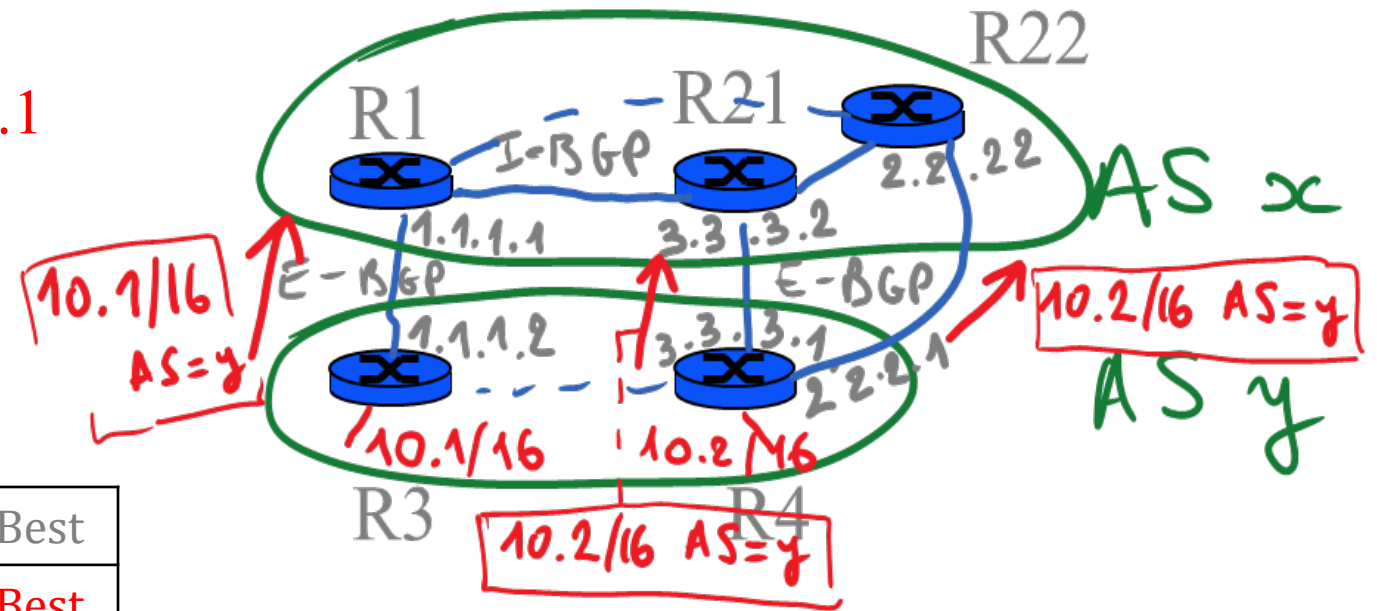
Adj-RIB-in

| | | |
|----------|--------------------------------|------|
| From R3 | 10.1/16 AS =y NEXT-HOP=1.1.1.2 | Best |
| From R22 | 10.2/16 AS =y NEXT-HOP=2.2.2.1 | Best |

Adj-RIB-out

| | |
|--------|--------------------------------|
| To R21 | 10.1/16 AS =y NEXT-HOP=1.1.1.2 |
| To R22 | 10.1/16 AS =y NEXT-HOP=1.1.1.2 |

- R1 accepts the updates and stores it in Adj-RIB-In
- R1 designates this route as best route
- R1 does **not** put route into Adj-RIB-Out to R21 because i-BGP is not repeated over i-BGP
- R1 does **not** put route into Adj-RIB-Out to R3 this would create an **AS-path loop**

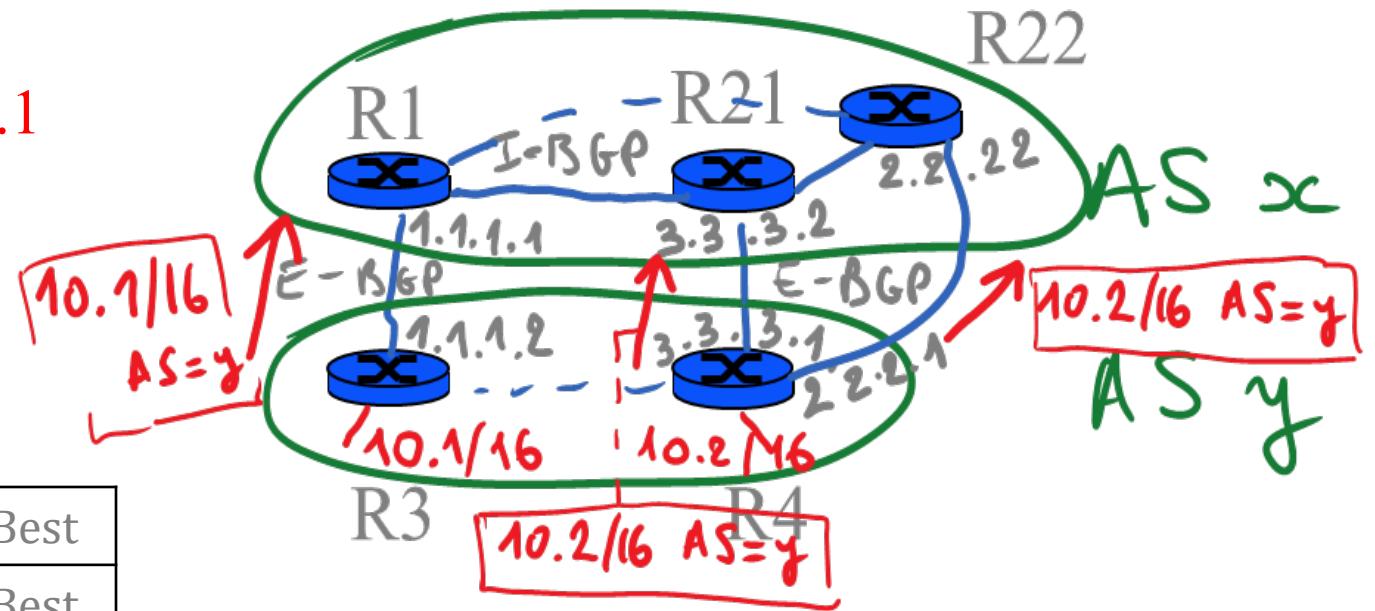


Step 3

R21 → R1
10.2/16 AS =y NEXT-HOP=3.3.3.1

Adj-RIB-in

| | | |
|----------|--------------------------------|------|
| From R3 | 10.1/16 AS =y NEXT-HOP=1.1.1.2 | Best |
| From R22 | 10.2/16 AS =y NEXT-HOP=2.2.2.1 | Best |
| From R21 | 10.2/16 AS =y NEXT-HOP=3.3.3.1 | |



Will the decision process promote the new route to “best route” assuming that WEIGHT, LOCAL_PREF, MED are empty?

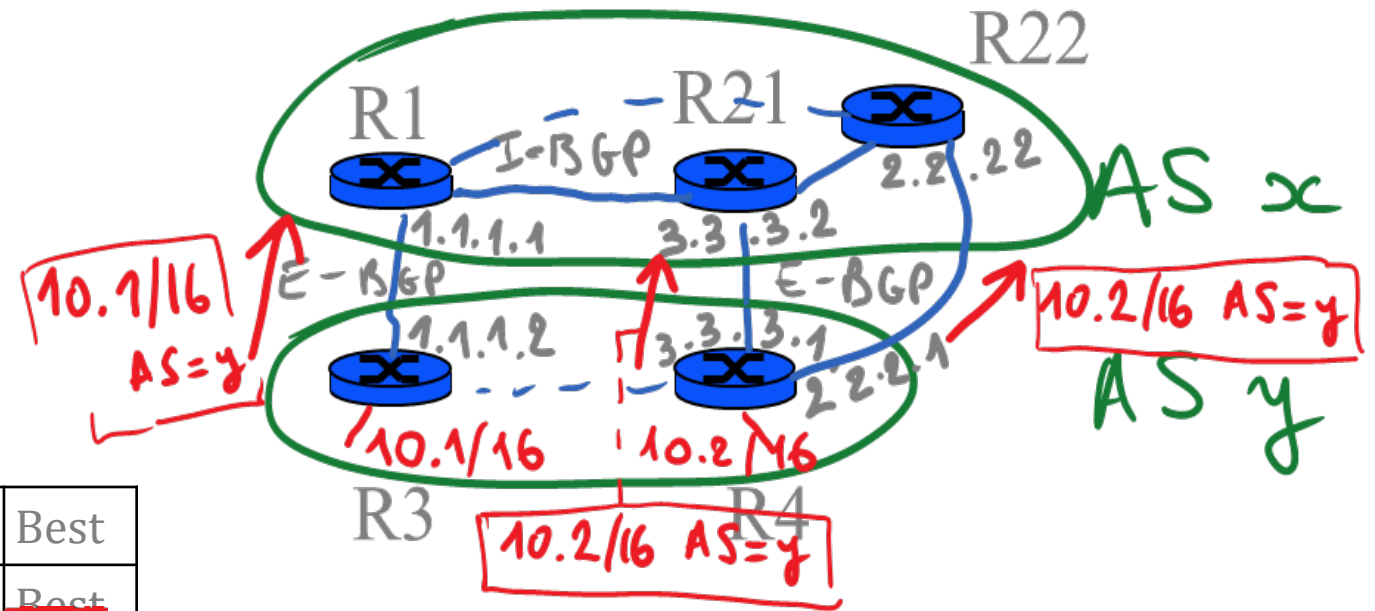
- A. Yes
- B. No, the route is worse
- C. No, it will keep both routes
- D. I don't know

0. Highest weight (Cisco proprietary)
 1. Highest LOCAL_PREF
 2. Shortest AS-PATH
 3. Lowest MED, if taken seriously by this network
 4. e-BGP > i-BGP (= if route is learnt from e-BGP, it has priority)
 5. Shortest path to NEXT-HOP, according to IGP
 6. Lowest BGP identifier (router-id of the BGP peer from whom)
- (The Cisco and FRR implementation of BGP, used in lab 6, have additional criteria)

Solution

Adj-RIB-in

| | | |
|----------|--------------------------------|-----------------|
| From R3 | 10.1/16 AS =y NEXT-HOP=1.1.1.2 | Best |
| From R22 | 10.2/16 AS =y NEXT-HOP=2.2.2.1 | Best |
| From R21 | 10.2/16 AS =y NEXT-HOP=3.3.3.1 | Best |



Answer A

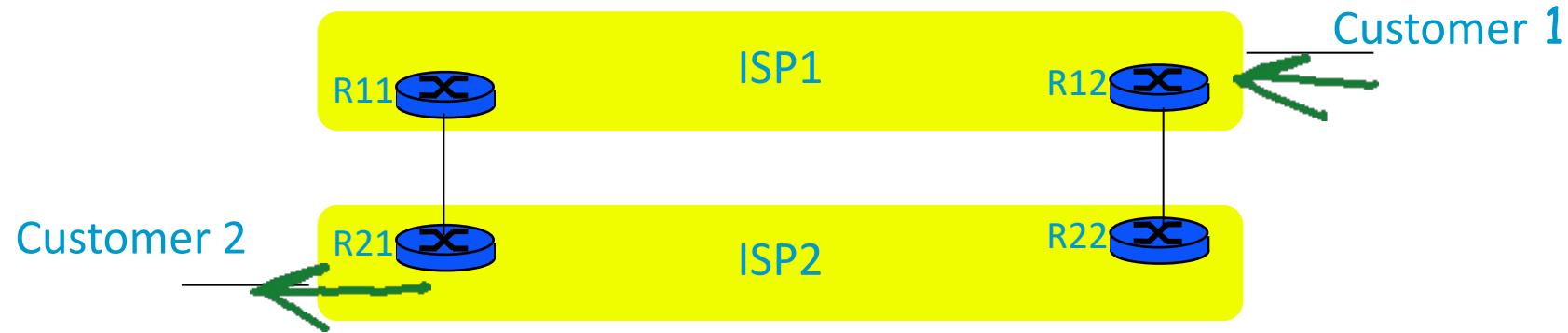
The decision process now has to choose between two routes with same destination prefix 10.2/16. Both were learnt from i-BGP, so we apply criterion 5 in slide “The Decision Process”.

The **distance**, computed by the IGP, to 2.2.2.1 is ≥ 3 and the distance to 3.3.3.1 is 2.

Thus the route that has NEXT-HOP=3.3.3.1 is preferred by the decision process, i.e. the new route is designated as “best”.

The new route is **not** put into Adj-RIB-Out for the same reasons as at step 2.

ISP1 and ISP2 are shared cost peers. Which path will be used by packets Customer 1 → Customer 2 ?

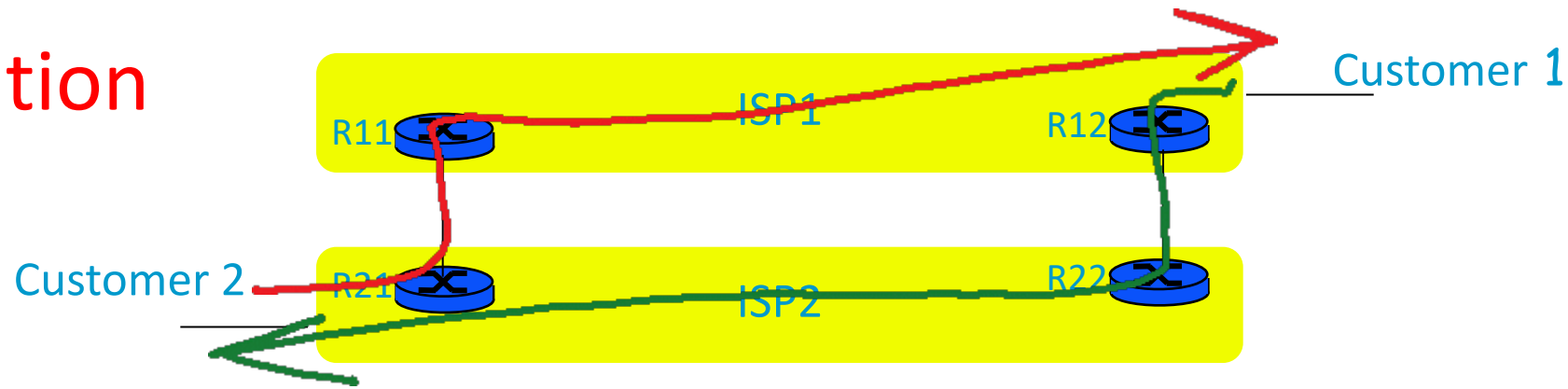


- A. R12-R11-R21
- B. R12-R22-R21
- C. It depends on the configuration of BGP at ISP1 and ISP2
- D. Both in parallel
- E. I don't know



Go to web.speakup.info
or
download speakup app
Join room
46045

Solution



Answer C: It **depends** on the configuration.

If configuration is as in “Fundamental Example”, Customer 1 → Customer 2 uses R12-R22-R21 («**Hot potato routing**»), but if the configuration is as in “Another Fundamental Example”, the other route is used (“**Cold potato routing**”)

If both ISPs do hot potato routing, Customer 2 → Customer 1 uses R21-R11-R12: routing in the global internet may be **asymmetric** !

How are routes originated (= sourced) ?

Several methods for sourcing a route:

Static configuration:

= *manually* tell this BGP router which prefixes to originate (“network” command in FRR)

Redistribute connected:

= tell this BGP router to originate *all directly attached prefixes*
(all routers in network may run i-BGP, no need for IGP in this case)

Redistribute from IGP:

= tell this BGP router to originate *all prefixes learnt by IGP*, e.g.: redistribute OSPF into BGP

- If IGP=OSPF, in principle, *only internal prefixes* should be redistributed
- Such BGP routes have attribute **ORIGIN=IGP**.
- When originated, the BGP NEXT-HOP of such a route is its **IGP next-hop**.

2. Aggregation (of routes)

Routes usually *overlap*

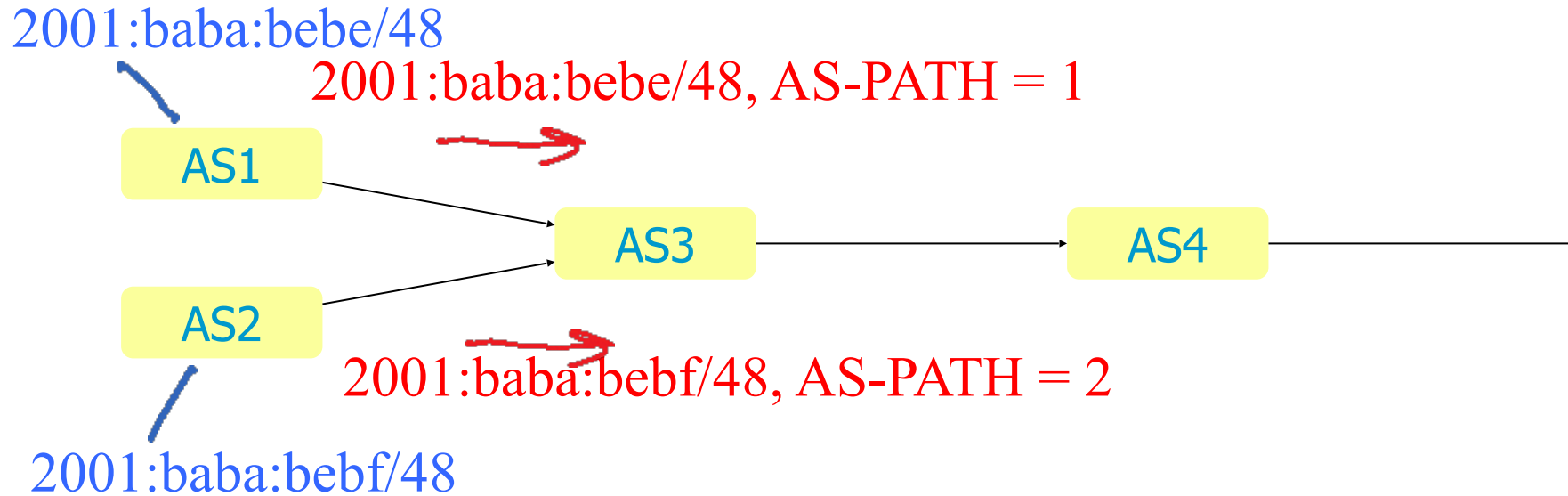
- expected to be very frequent with IPv6 (recall the way we delegate prefixes), less with IPv4

So, *prefix aggregation* can reduce the number of routes

- in IP forwarding tables
- in BGP advertizements

otherwise several hundreds of thousands of entries or advertizements (e.g. consider transit ISPs without a default route)

Can AS3 aggregate these routes into a single one?

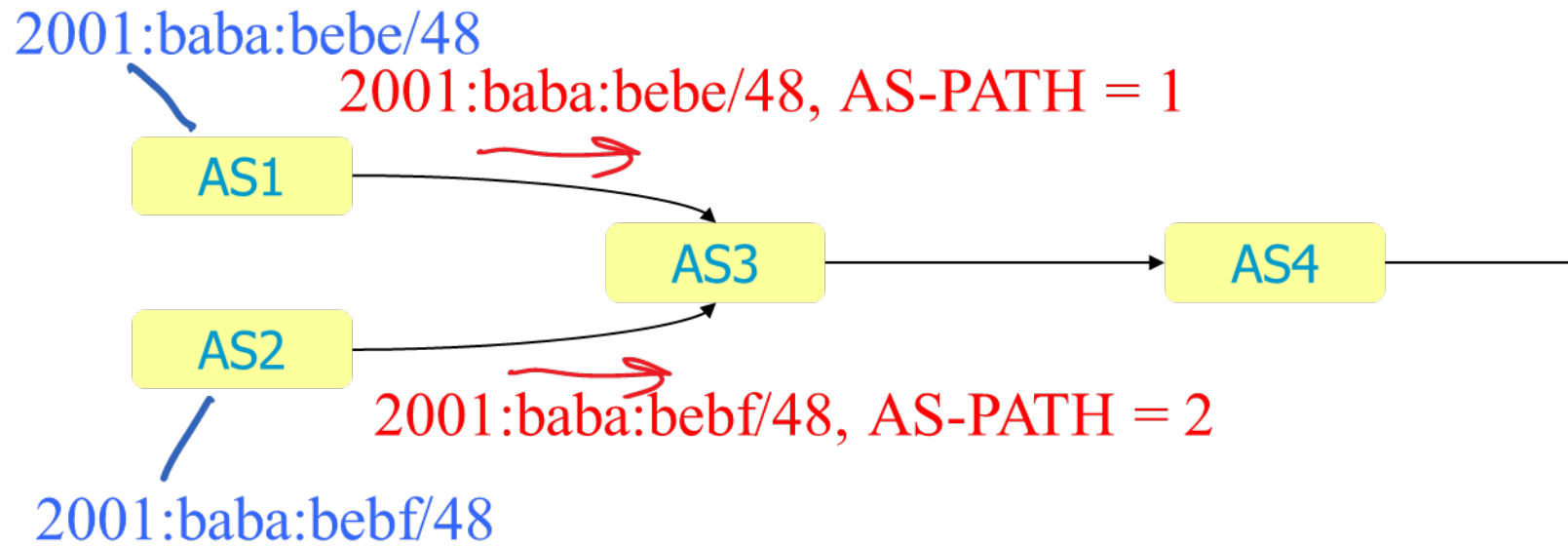


- A. Yes and the aggregated prefix is 2001:baba:bebe/47
- B. Yes and the aggregated prefix is 2001:baba:bebf/48
- C. Yes but the aggregated prefix is none of the above
- D. No
- E. I don't know



Go to web.speakup.info
or
download speakup app
Join room
46045

Solution



Answer A.

The two prefixes are contiguous and can be aggregated as 2001:baba:bebe/47

Actual advertizements:

AS3 sends to AS4 the UPDATE

2001:baba:bebe/47 AS-PATH = 3 {1 2}

AS4 sends the UPDATE

2001:baba:bebe/47 AS-PATH = 4 3 {1 2}

{ } means aggregation

2001:baba:bebe/48

1110

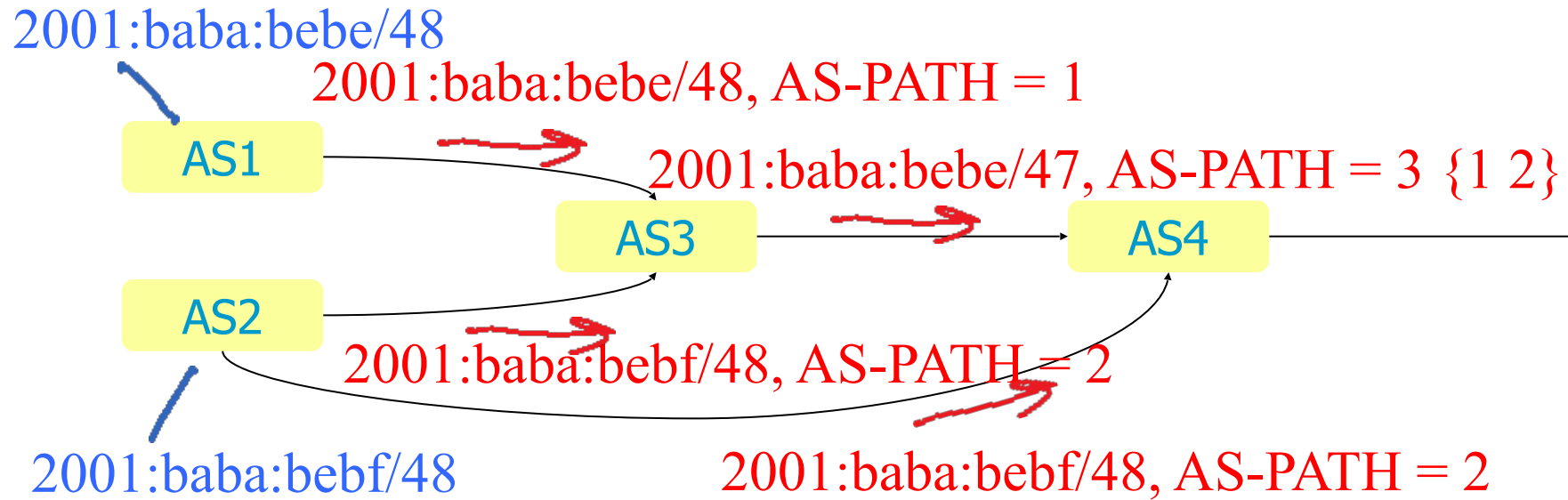
2001:baba:bebf/48

1111

2001:baba:bebe/47

1110

Which routes may the decision process in AS4 designate as best ?

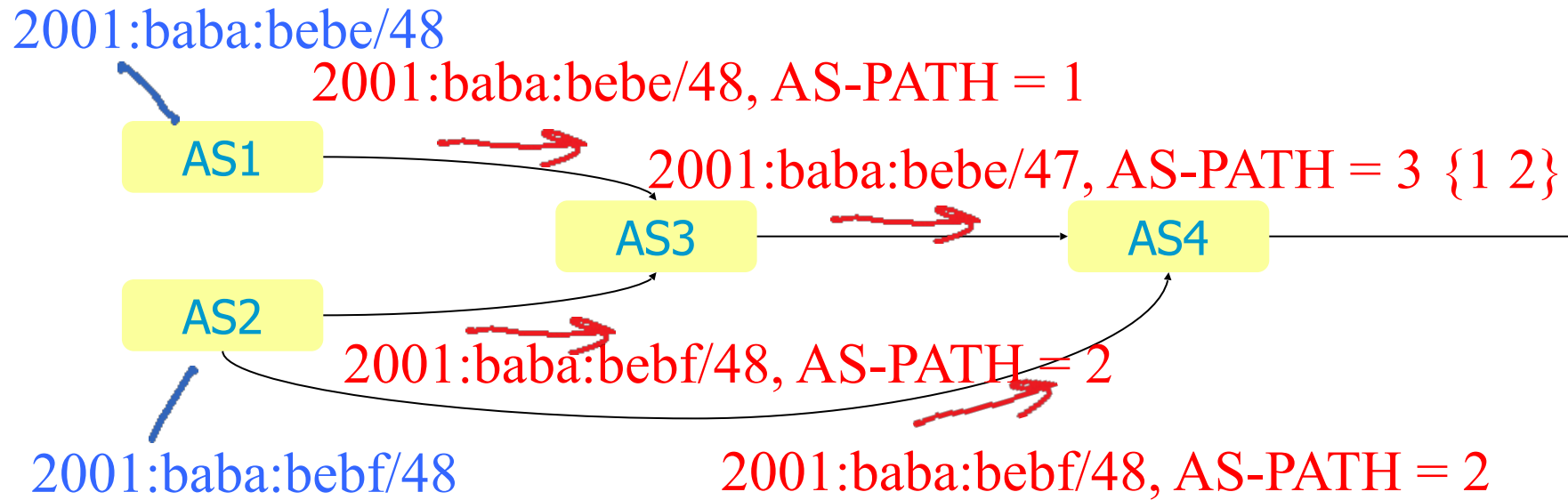


- A. The top route
- B. The bottom route
- C. Both
- D. I don't know



Go to web.speakup.info
or
download speakup app
Join room
46045

Solution



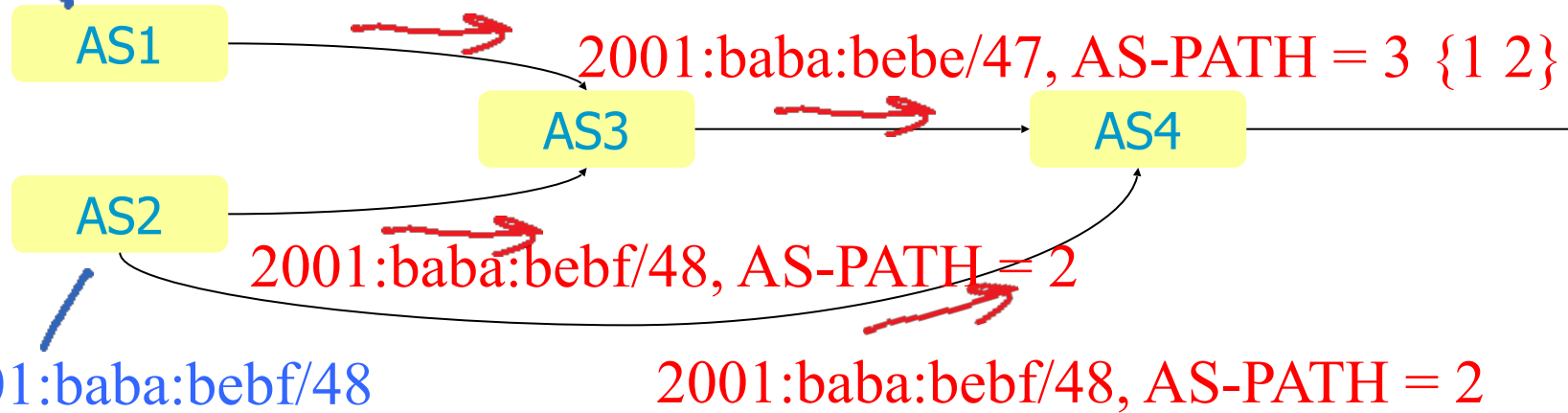
Answer C.

The decision process in AS4 may select **both** routes because they are to **different destinations**. Overlapping routes are considered different.

Assume the decision process in AS4 designates both routes as best.
Which path does a packet from AS4 to 2001:baba:bebf/48 follow ?

2001:baba:bebe/48

2001:baba:bebe/48, AS-PATH = 1

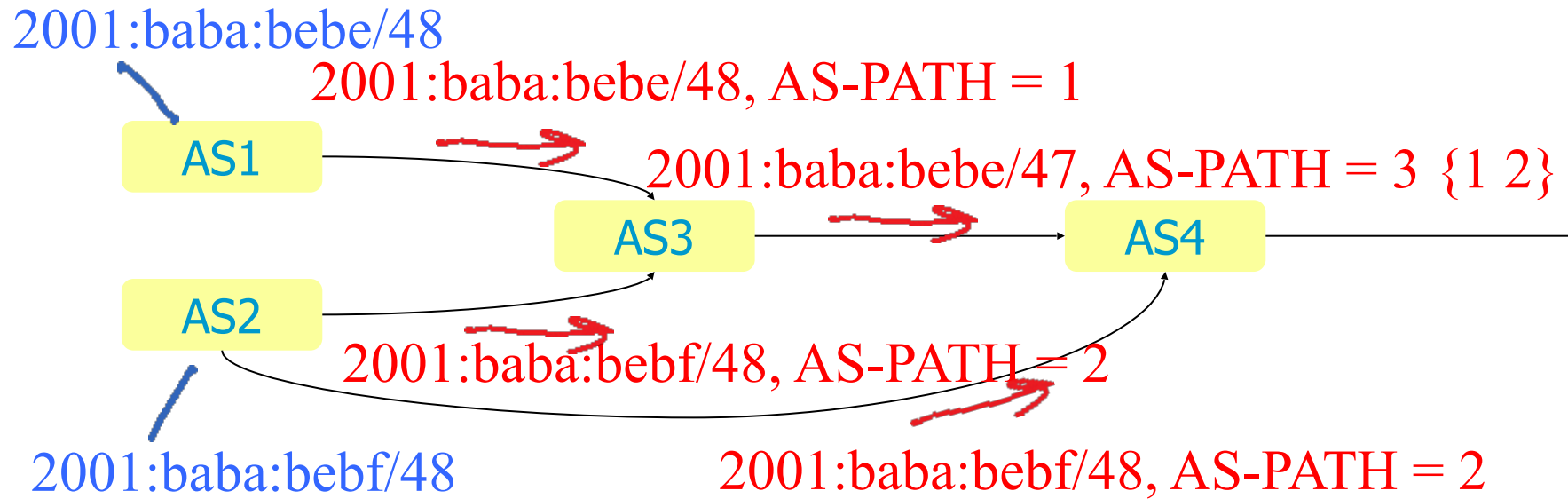


- A. AS4-AS3-AS2
- B. AS4-AS2
- C. I don't know



Go to web.speakup.info
or
download speakup app
Join room
46045

Solution



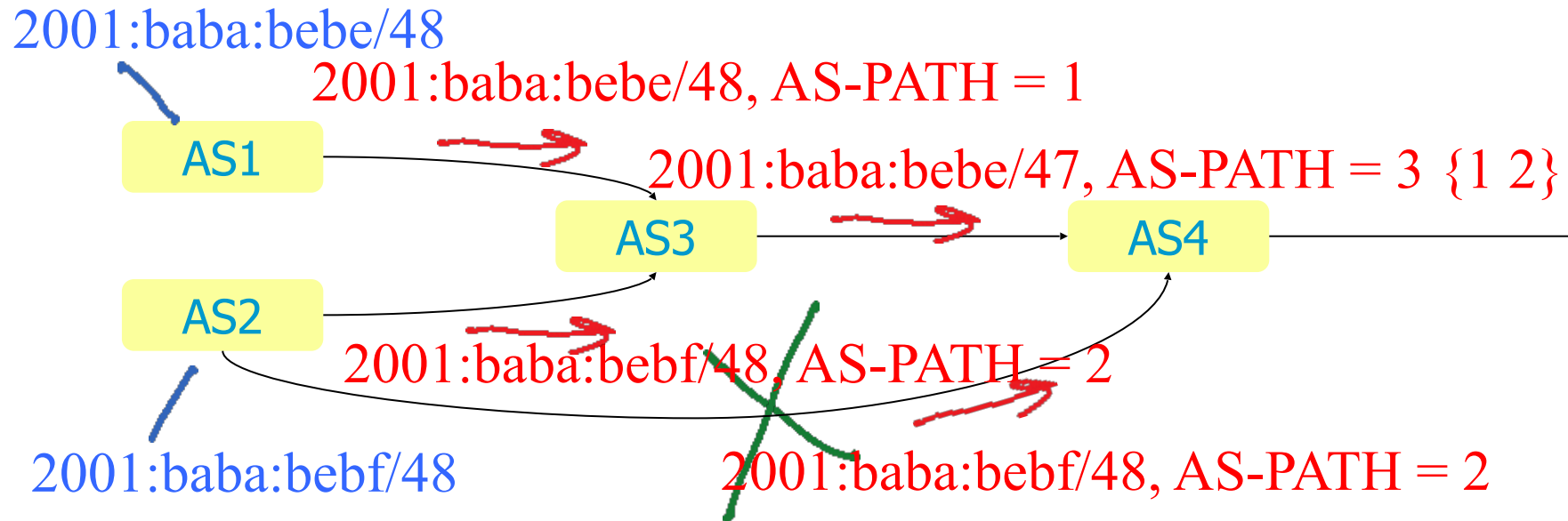
Answer B.

Recall: a BGP router is still an IP forwarding device.

So, it uses **longest prefix match** \Rightarrow packet goes AS4-AS2.

Another example: a packet to 2001:baba:bebe/48 will go AS4-AS3-AS1.

Assume the link AS2-AS4 breaks ...



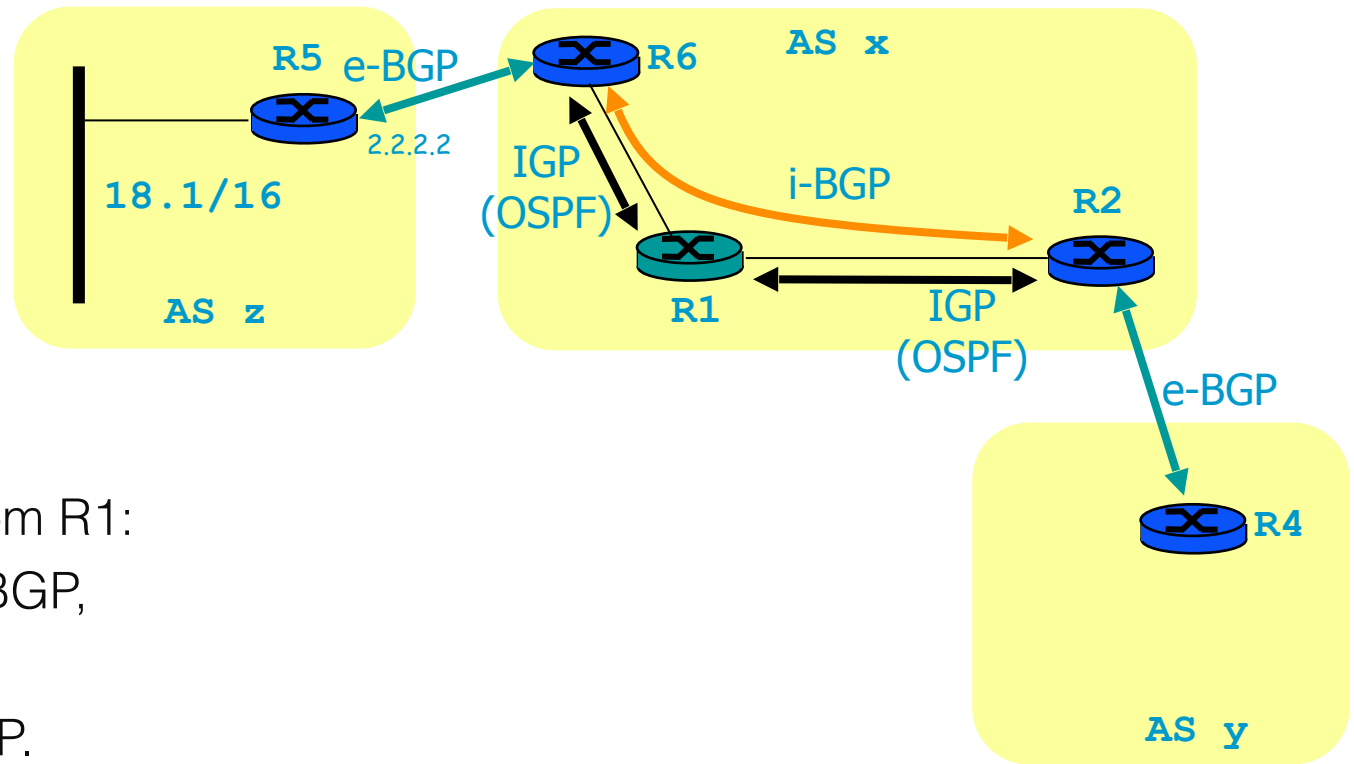
- At AS4: KEEPALIVE mechanism detects that the border router at AS2 is unreachable
- Related Adj-RIB-In, Adj-RIB-loc routes are declared invalid
- Decision process recomputes best route to 2001:baba:bebf/48 —> there is none
- The forwarding table entry 2001:baba:bebf/48 is removed
- **but...** a packet to 2001:baba:bebf/48 matches the route 2001:baba:bebe/47 and can go via AS3

3. How routes learnt by BGP are written into Forwarding Tables?

There are two possible ways:

1. *Redistribution of BGP into IGP*: routes learnt by BGP are *passed to IGP* (e.g.: OSPF)
 - Typically *only* routes learnt by *e-BGP* are redistributed
(unless BGP redistribute-internal is used)
 - IGP propagates the routes to all routers in domain
 - Works with OSPF, might not work with other IGPs (table too large for IGP)

Example (re-distribution)



Assume all routers run BGP apart from R1:

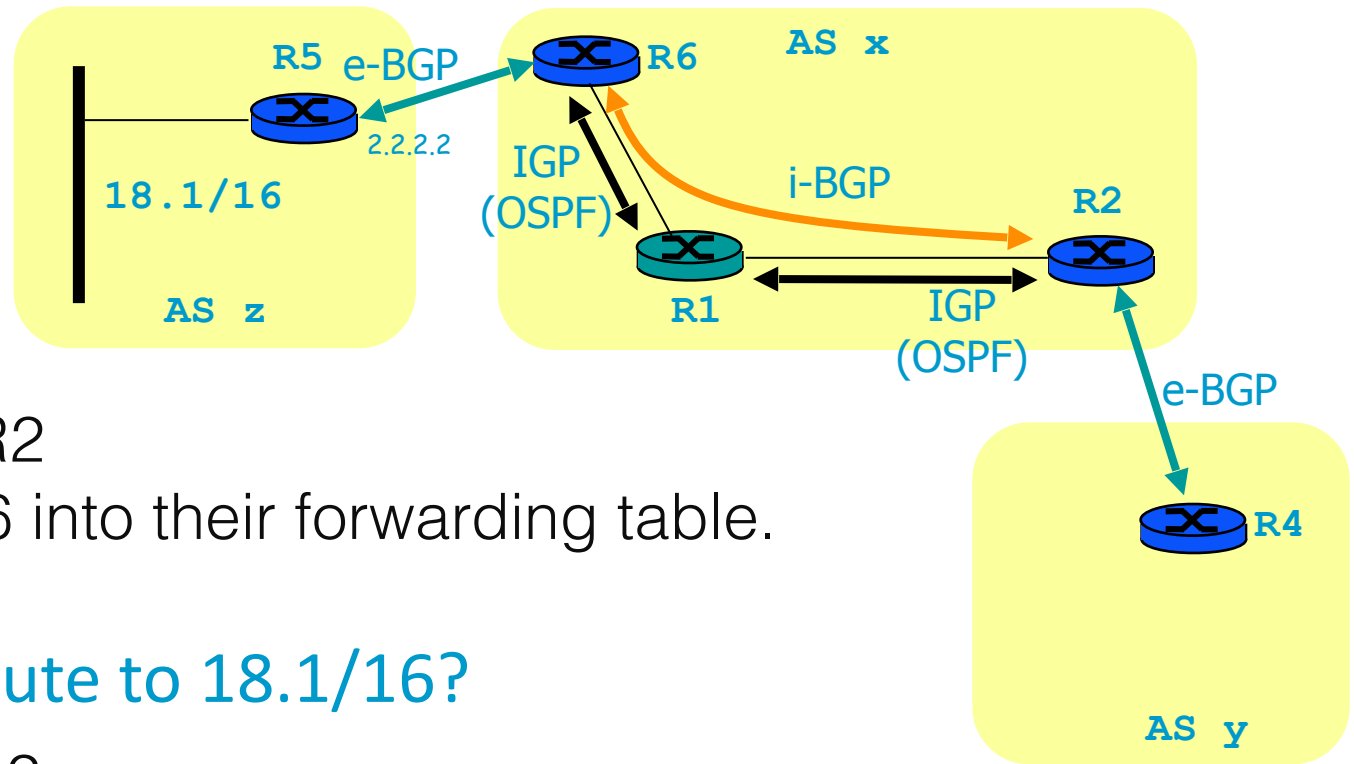
- R5 advertises 18.1/16 to R6 via e-BGP,
- R6 advertises it to R2 via i-BGP,
- R2 advertises route to R4 via e-BGP.
- R6 **redistributes** 18.1/16 (learnt from e-BGP) into IGP
 - For the IGP, it is as if 18.1/16 were *directly connected* to R6.
The IGP cost, if required, is usually set to a value *higher* than all IGP distances.
 - IGP propagates 18.1/16 *internally* (in OSPF: there is a separate LSA for this—type 5).
 - R1, R2, R6 update forwarding tables. R1, R2 now have a route to 18.1/6.
 - Packet to 18.1/16 from AS y finds forwarding table entries in R2, R1 and R6

How routes learnt by BGP are written into Forwarding Tables?

There are two possible ways:

1. **Redistribution of BGP into IGP**: routes learnt by BGP are *passed to IGP* (e.g.: OSPF)
 - Typically only routes learnt by e-BGP are redistributed (unless BGP redistribute-internal is used)
 - IGP propagates the routes to all routers in domain
 - Works with OSPF, might not work with other IGPs (table too large for IGP)
2. **Injection**: Routes learnt by BGP are *directly written/copied* into forwarding table of this BGP router
 - **Why** used? IGP avoids dealing with a large number of routing entries (consider potential convergence issues in distance-vector algorithms, such as RIP).
 - Routing information is not propagated to other intra-domain routers; so, injection helps only the particular BGP router.
 - Typically used in Cisco routers and FRR (in the lab).

Example (injection)



Assume BGP routers R6 and R2
inject/copy the route to 18.1/16 into their forwarding table.

What is the next-hop for a route to 18.1/16?

- A. At R6: 2.2.2.2, at R2: 2.2.2.2
- B. At R6: 2.2.2.2, at R2: the IP address of R1-east
- C. At R6: 2.2.2.2, at R2: the IP address of R6-south
- D. None of the above
- E. I don't know

Solution

Answer A.

When a BGP router injects a route into the forwarding table, it **copies** the BGP NEXT-HOP into the forwarding table's next-hop.

Ideally, the correct answer should be B but is in fact A.

Normally, the next-hop in a forwarding table is on-link (inside the same subnet) and is the **interface** of the next router on the path, i.e. R1-east.

However, in this case, this requires that R2 learns the path to 18.1/16, by the IGP. Since 18.1/16 is not redistributed into the IGP, there is a **problem**.

The problem is usually solved via ***recursive table lookup***. (See next slides)

Recursive Table Lookup

Why?

A BGP router *injects* a route into its forwarding table = it *copies* the BGP NEXT-HOP into the forwarding table's next-hop.

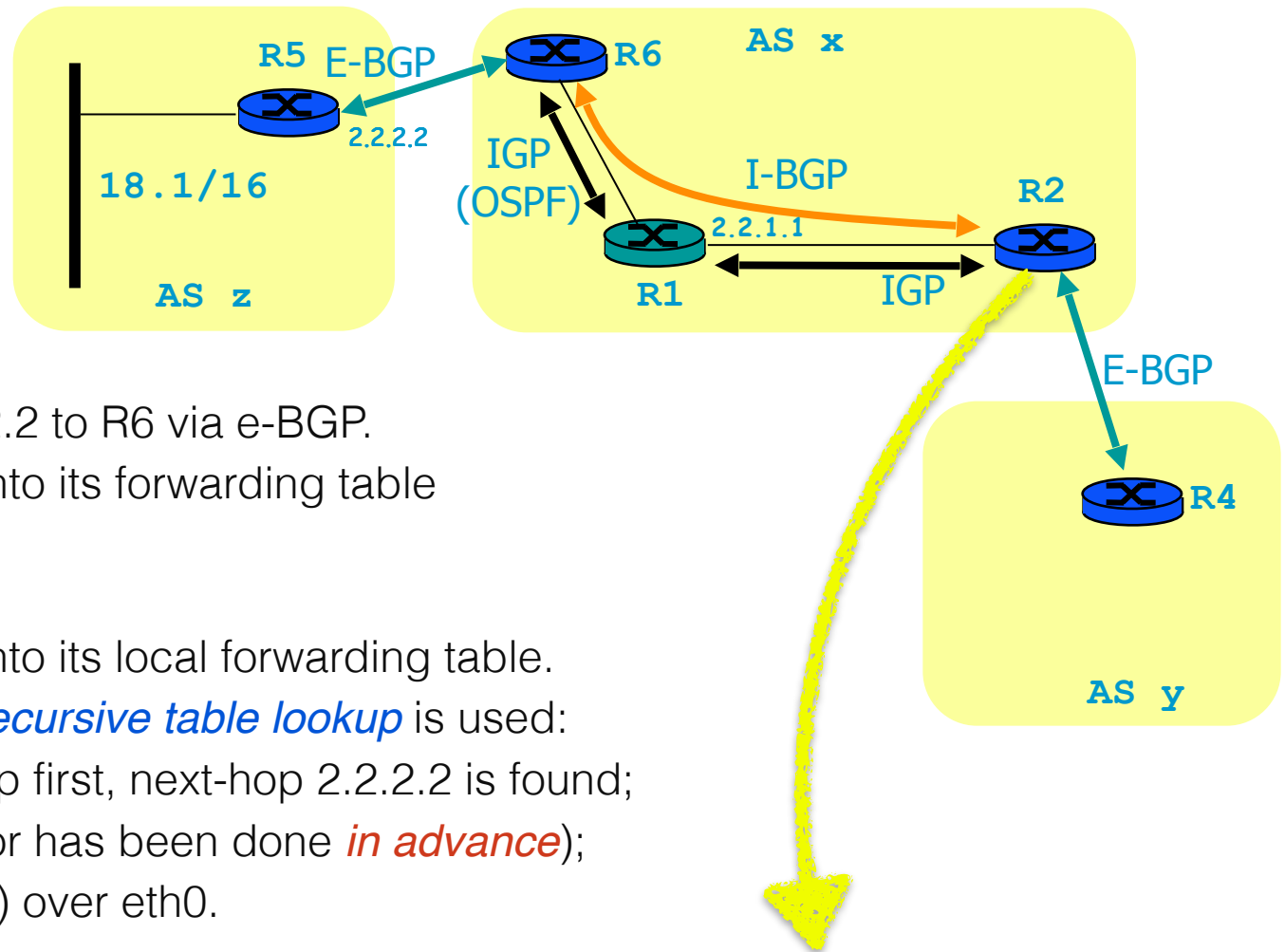
So, the forwarding table may indicate next-hops which are not “on-link” (i.e. within the subnets of this router).

How?

To resolve the non-on-link next-hop into an on-link next-hop neighbor, a *second lookup* is done into the forwarding table

in fact, the second lookup may be done *in advance*—not in real time—by *pre-processing* the routing table.

Example (injection, cont'd)

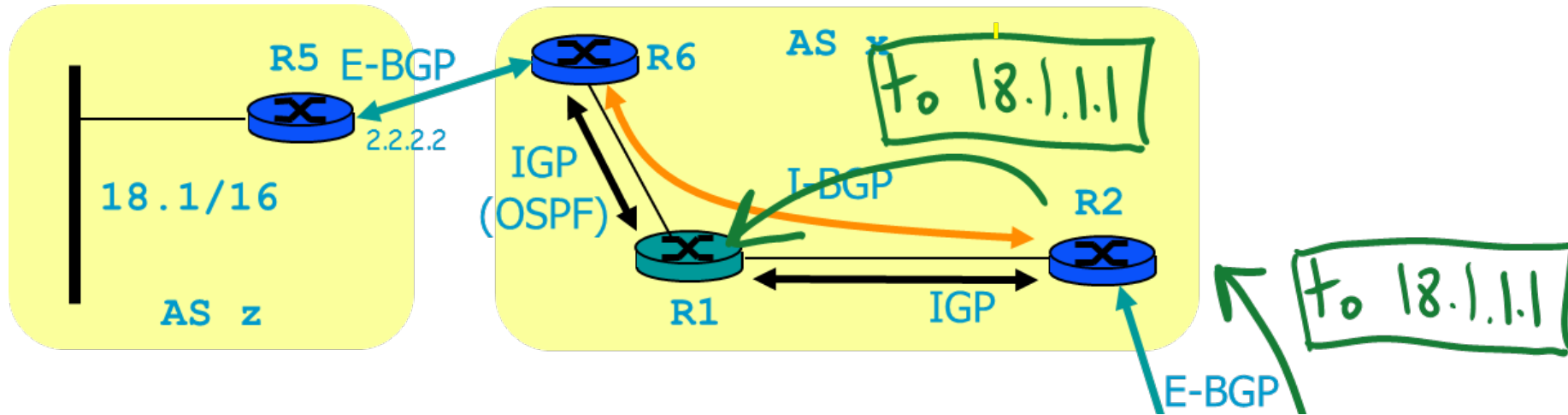


- R5 advertises 18.1/16, NEXT-HOP = 2.2.2.2 to R6 via e-BGP.
- R6 injects 18.1/16, NEXT-HOP = 2.2.2.2 into its forwarding table (does not re-distribute into OSPF).
- R2 learns route from R6 via i-BGP.
- R2 injects 18.1/16, NEXT-HOP = 2.2.2.2 into its local forwarding table.
- IP packet to 18.1.1.1 is received by R2, *recursive table lookup* is used:
 - the forwarding table at R1 is looked up first, next-hop 2.2.2.2 is found;
 - a second lookup for 2.2.2.2 is done (or has been done *in advance*);
 - packet is sent to R1 (interface 2.2.1.1) over eth0.

Forwarding Table at R2

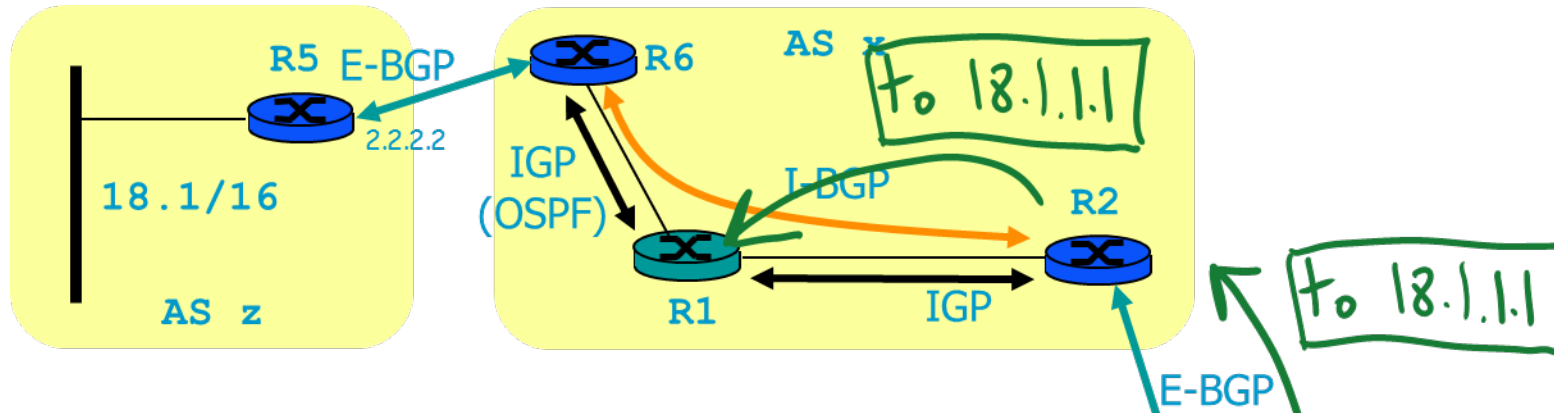
| To | next hop | interface |
|---------|----------|-----------|
| 18.1/16 | 2.2.2.2 | N/A |
| 2.2.2.2 | 2.2.1.1 | eth0 |

Injection (no redistribution into IGP): What happens to this IP packet at R1 ?



- A. It is forwarded to R6 because R1 does recursive table lookup
- B. It is forwarded to R6 because R1 runs an IGP
- C. It cannot be forwarded to R6
- D. I don't know

Solution



Answer C

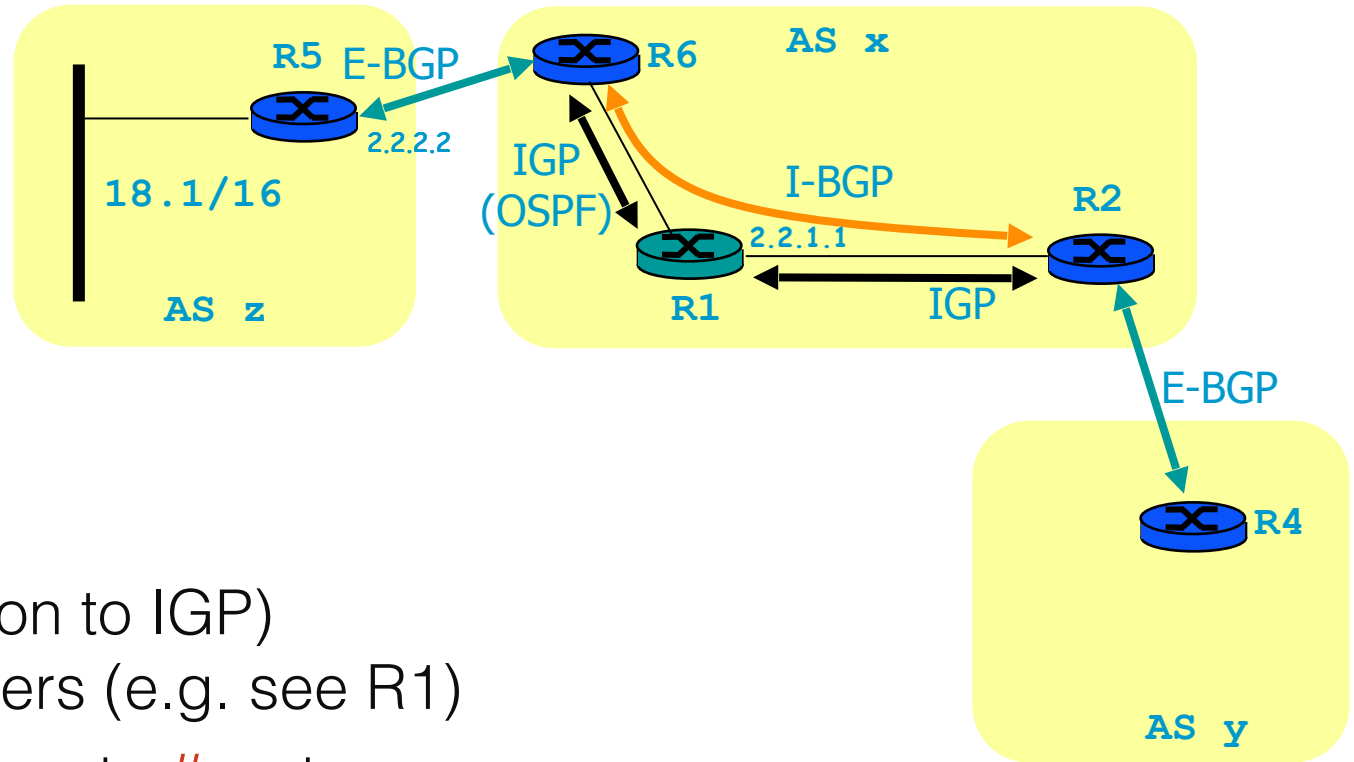
The IGP announces only internal routes since we use only injection and we do not re-distribute BGP into IGP.

R1 does not run BGP.

Thus R1 does not have any route to 18.1/16 in its forwarding table.

The packet cannot be forwarded by R1 (“**destination prefix not found**”).

In practice, simple injection implies that *all* routers need to run BGP

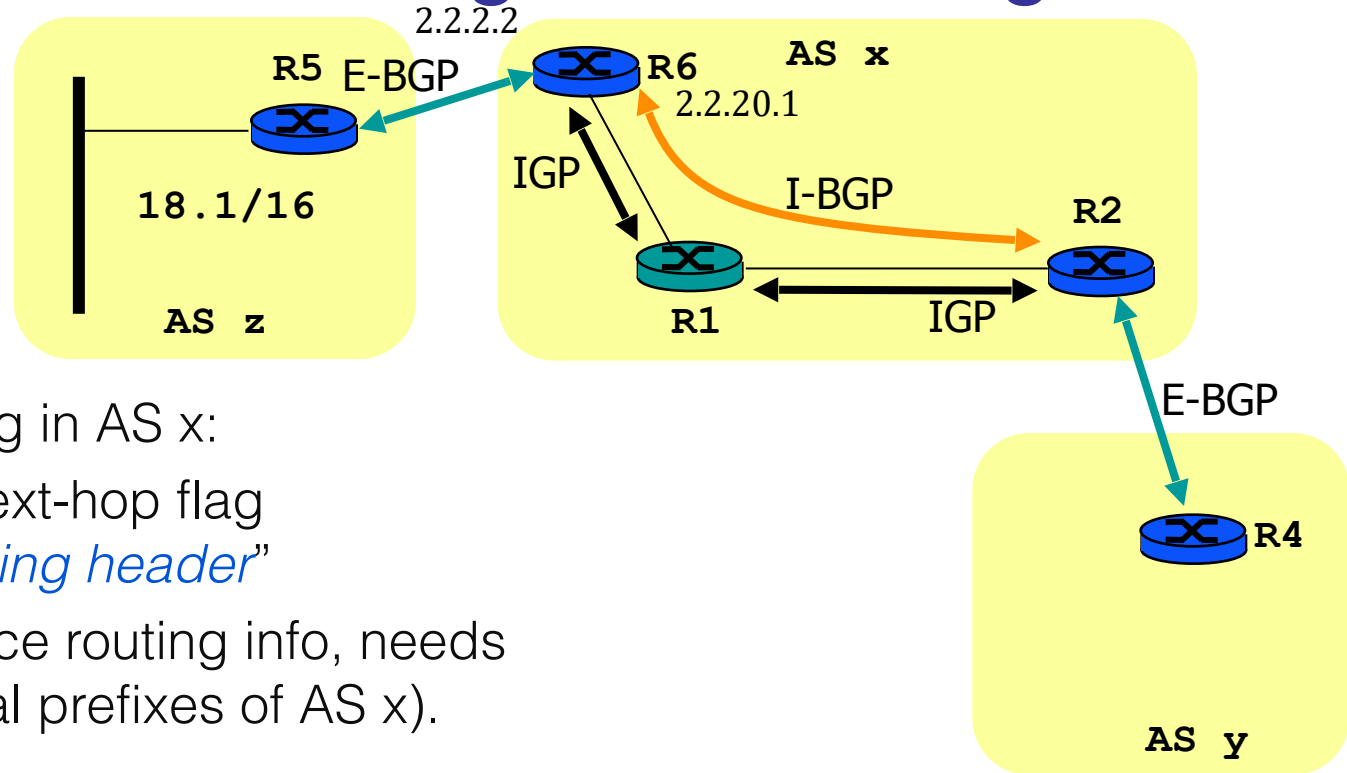


The “injection-only” BGP setup:

- *all* routers run BGP (in addition to IGP) even if not being border routers (e.g. see R1)
- recursive table lookup is done at *all* routers
- potential problem: size of i-BGP mesh —> use reflectors (see later)
- *IGP is still needed* to discover paths to next-hops; but handles only internal prefixes – very few

Alternative: BGP with source/segment routing

Alternative to redistribution
or running i-BGP in all routers:



Use source routing / segment routing in AS x:

- Routing table at R2 contains next-hop flag
“*insert next-hop as source routing header*”
- R1 forwards packet using source routing info, needs only small routing table (internal prefixes of AS x).

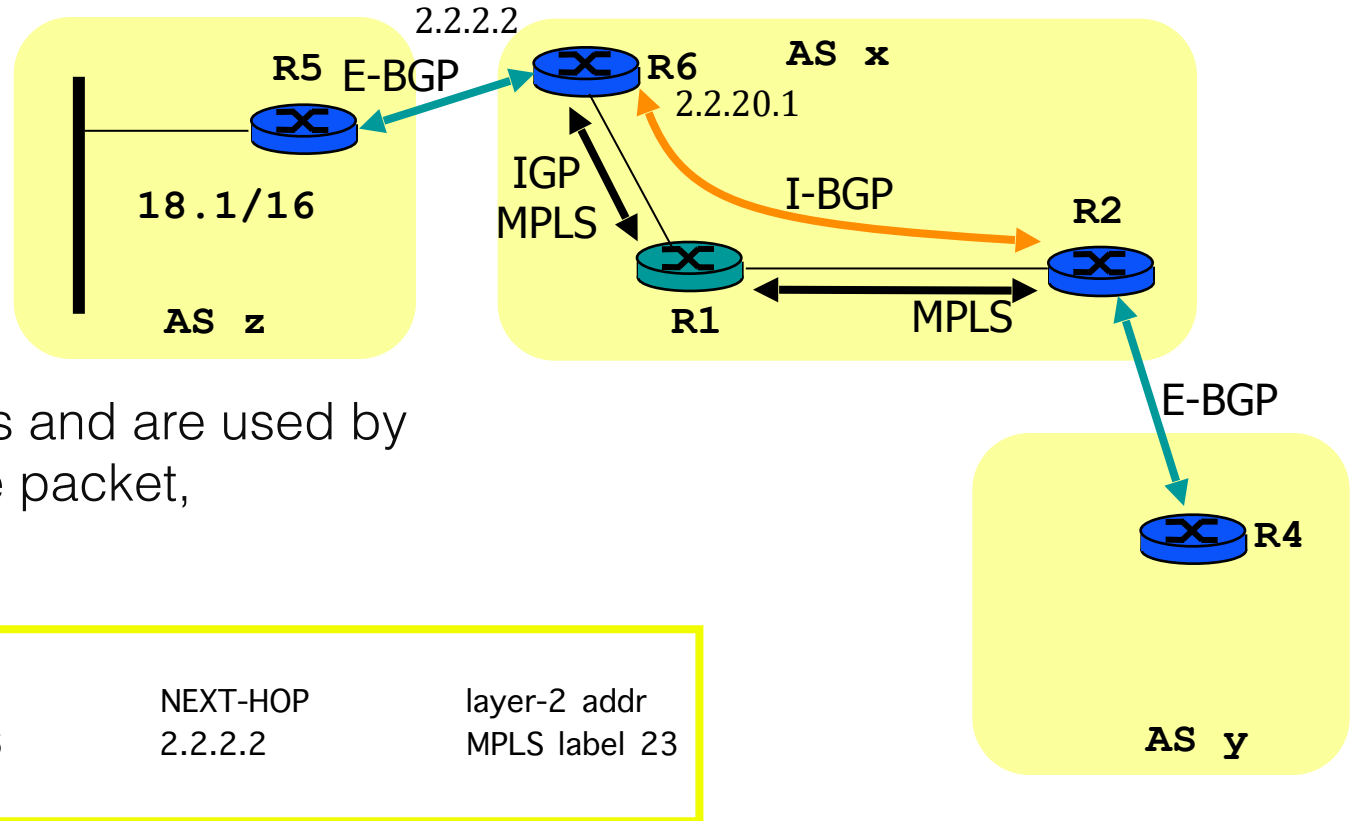
| | | | |
|-------|---------|----------|--|
| at R2 | To | NEXT-HOP | Flags |
| | 18.1/16 | 2.2.2.2 | insert next-hop as source routing header |

- **SCION** (alternative to BGP) uses a similar mechanism.

Alternative: BGP with MPLS

Alternative to redistribution
or running I-BGP in all routers:

Associate *MPLS labels* to exit points
MPLS labels are similar to VLAN tags and are used by
MPLS-capable routers to forward the packet,
without looking at the IP header.



Example:

- R1, R2 and R6 support IP and MPLS
- R2 creates a “*label switched path*” to 2.2.2.2, with label 23
- At R2: Packets to 18.1/6 are associated with this label
- *R1 runs only IGP and MPLS*—no BGP → only very small routing tables

Injection conflicts

In FRR and cisco, BGP always injects routes into forwarding table, even if these routes are redistributed into IGP. This may cause injection conflicts:

- a route may be *injected* into the forwarding table by *both IGP (e.g. OSPF) and BGP*.

To solve the conflicts, every route in the forwarding table has an attribute called *administrative distance*, which depends on which process wrote the route:

E-BGP = 20,

OSPF = 110,

RIP = 120,

I-BGP = 200

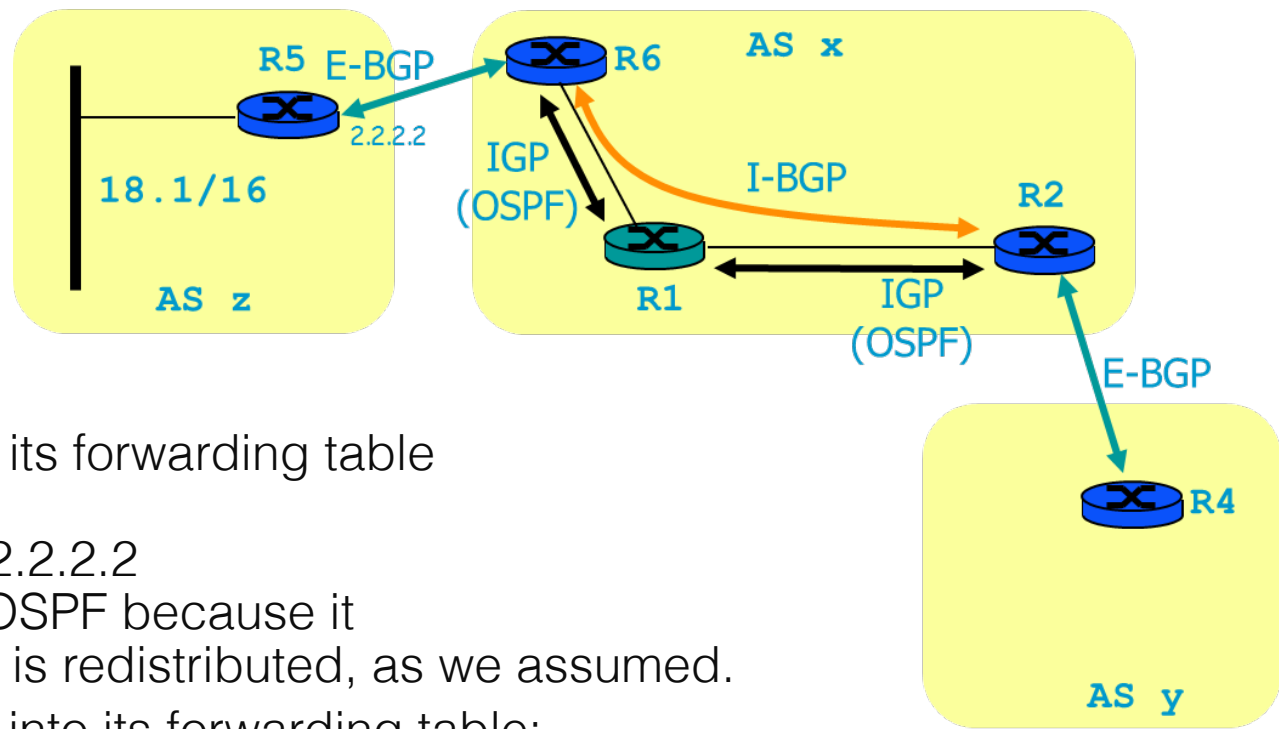
Administrative distance is compared before the usual distance.

- Only the route with the *smallest administrative distance* is selected to forward IP packets.
- The decision process selects a BGP route, only if there is no route to same destination prefix with smaller administrative distance in the forwarding table.

Admin distance is *local* and is not used by routing protocols.

Example

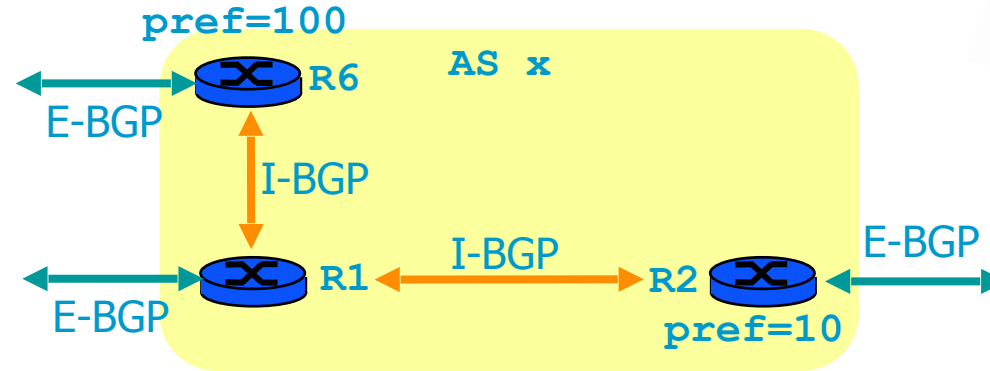
Assume R2 and R6 redistribute e-BGP into OSPF, but also inject routes directly.



- at t_1 : R2 injects 18.1/16 from BGP into its forwarding table
In **R2's forwarding table** we see:
18.1/6, Admin Dist = 200, Next-Hop=2.2.2.2
R2 does not redistribute 18.1/16 into OSPF because it was learnt with i-BGP and only e-BGP is redistributed, as we assumed.
- at $t_2 > t_1$: R6 injects 18.1/16 from BGP into its forwarding table;
In **R6's forwarding table** we see:
18.1/6, Admin Dist = 20, Next-Hop=2.2.2.2
then R6 redistributes 18.1/16 from BGP into OSPF with OSPF cost = 20 (an arbitrary value chosen as Cisco's default).
- at $t_3 > t_2$: via OSPF R2 learns the route and injects it into its forwarding table.
In **R2's forwarding table** we see an **injection conflict**:
18.1/6, Admin Dist = 110, cost = 22, Next-Hop=R1-east
18.1/6, Admin Dist = 200, Next-Hop 2.2.2.2
- The **Admin Distance** solves the conflict: R2 uses only the first route.

4. Other Route Attributes

LOCAL-PREF



0. Highest weight (Cisco proprietary)
 1. Highest LOCAL-PREF
 2. Shortest AS-PATH
 3. Lowest MED, if taken seriously by this network
 4. e-BGP > i-BGP (= if route is learnt from e-BGP, it has priority)
 5. Shortest path to NEXT-HOP, according to IGP
 6. Lowest BGP identifier (router-id of the BGP peer from whom route is learnt)
- (The Cisco and FRR implementation of BGP, used in lab 6, have additional)

- Used *inside an AS* to express preference.
- Assigned by BGP router when *receiving* route over E-BGP.
- Propagated without change over i-BGP; *not used (ignored)* over e-BGP.

Example

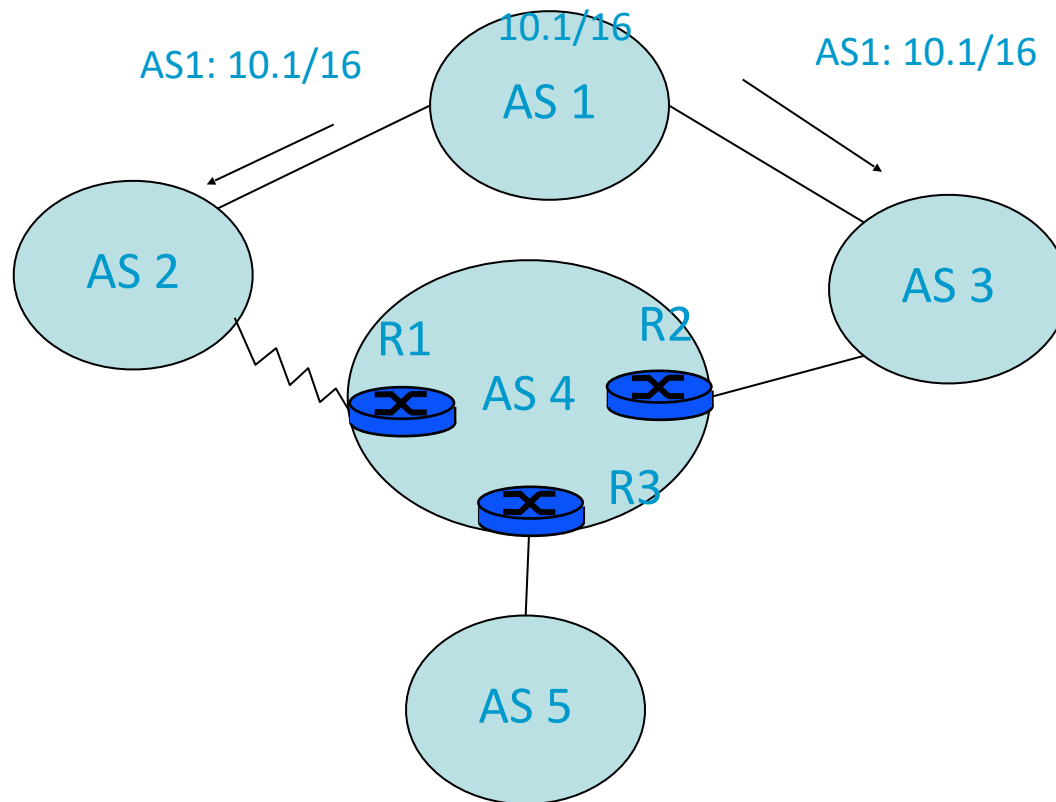
R6 associates pref=100, R2 pref=10

R1 chooses the largest preference

LOCAL-PREF Example: Link AS2-AS4 is expensive

AS 4 sets LOCAL-PREF to:

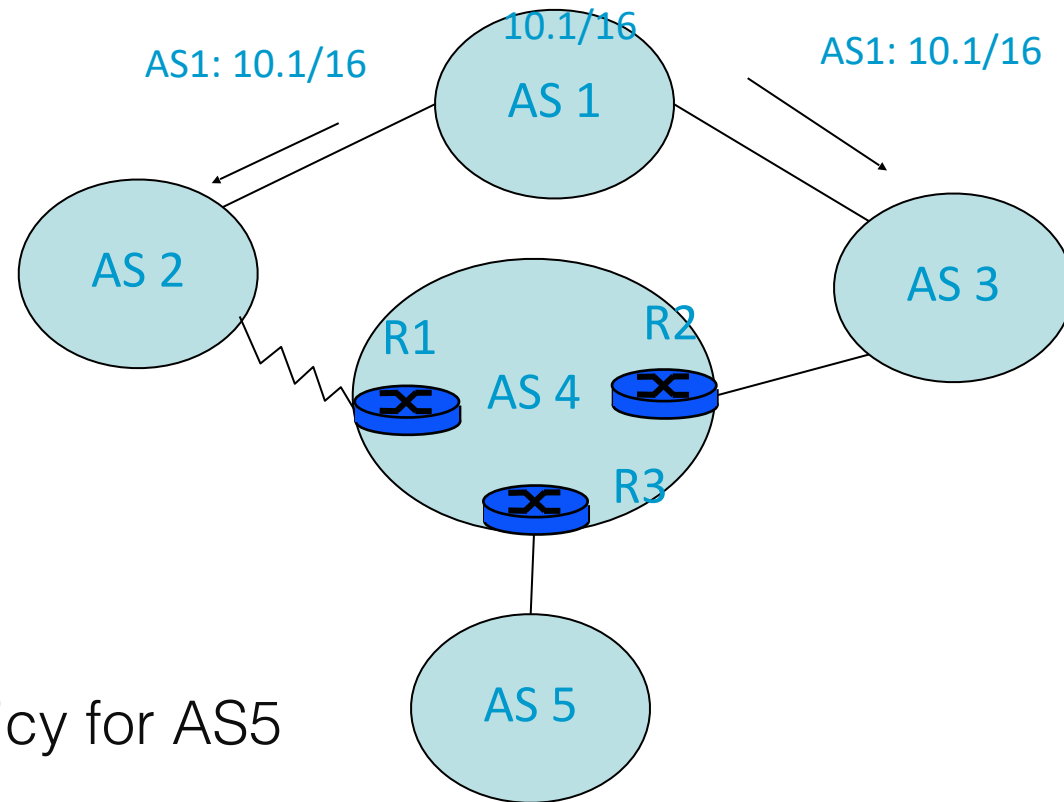
- 100 to all routes received from AS 3
- 50 to all routes received from AS 2



R1 receives the route
AS2 AS1 10.1/16
over e-BGP; sets LOCAL-PREF to 50

R2 receives the route
AS3 AS1 10.1/16
over e-BGP; sets LOCAL-PREF to 100

What does R3 announce to AS5?



- A. 10.1/16 AS-PATH=AS4 AS2 AS1
- B. 10.1/16 AS-PATH=AS4 AS3 AS1
- C. Any of the two, depending on policy for AS5
- D. Both
- E. None
- F. I don't know

Solution

Answer B

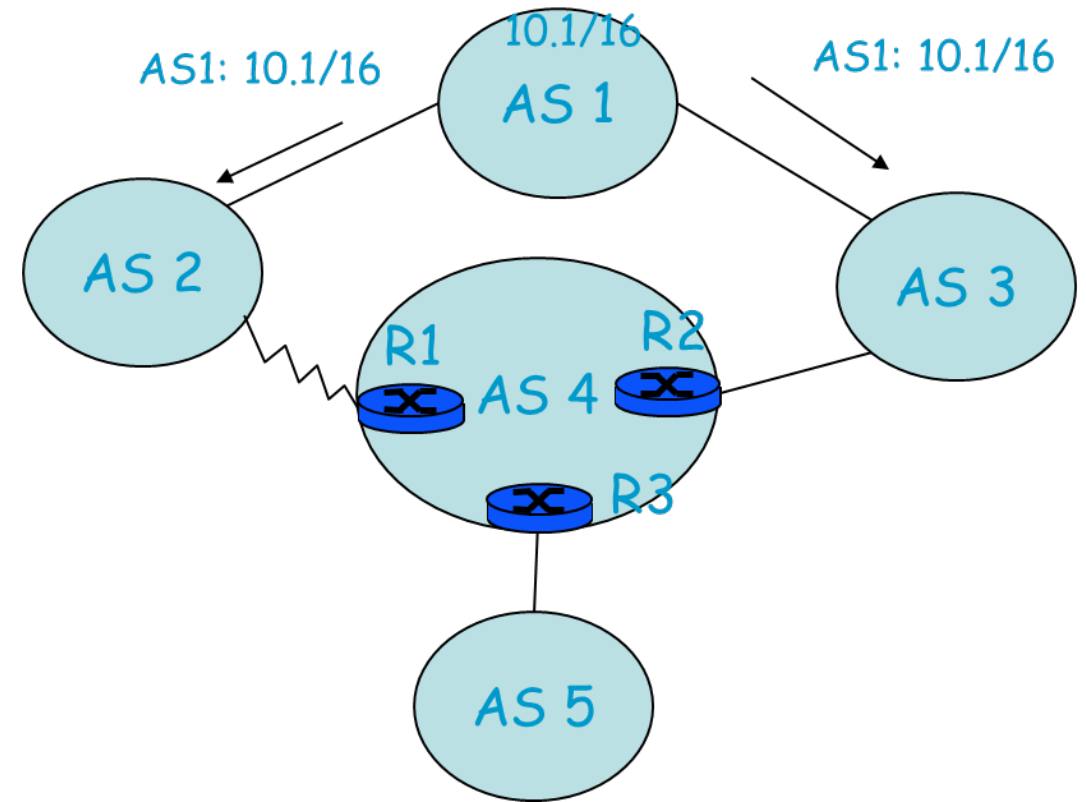
R1, R2 and R3 all select the route via AS3 as best route to 10.1/16 because of the LOCAL-PREF attribute

R3 advertises only its best route to AS5, i.e. 10.1/16 AS-PATH=AS4 AS3 AS1

R1 injects in forwarding table the next-hop corresponding to the R2-AS3 link and therefore the packet to 10.1.1.1 goes via AS3

Answer C is not possible because BGP allows only best route to be propagated

Answer E is possible if the policy in AS4 forbids propagating this route



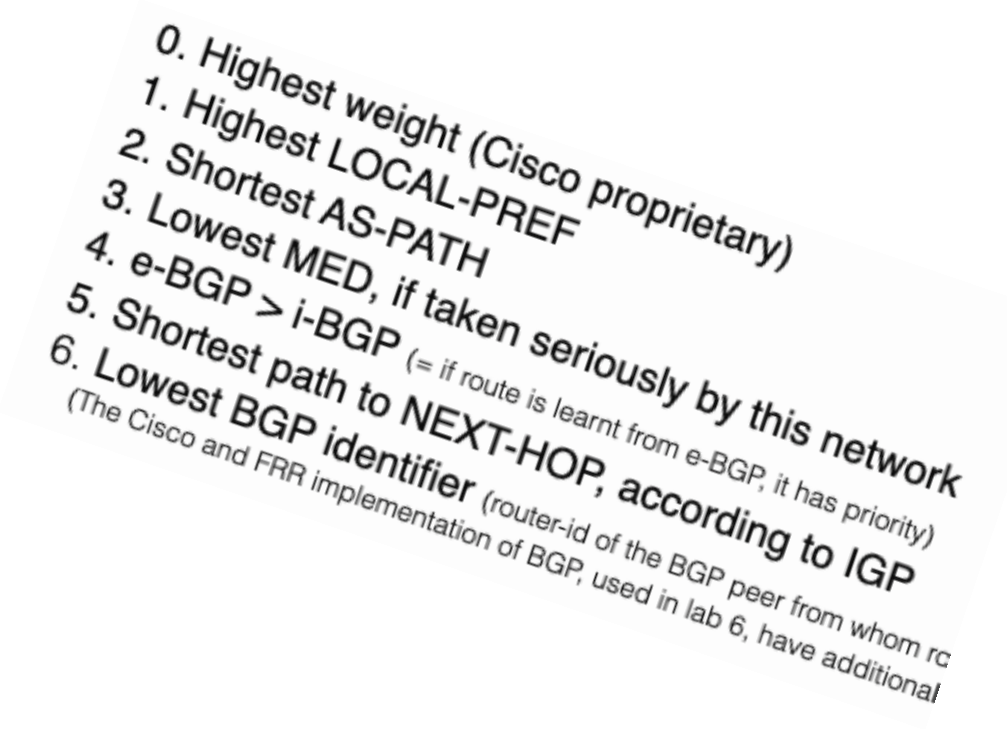
Weight

This is a route attribute given by Cisco or similar router

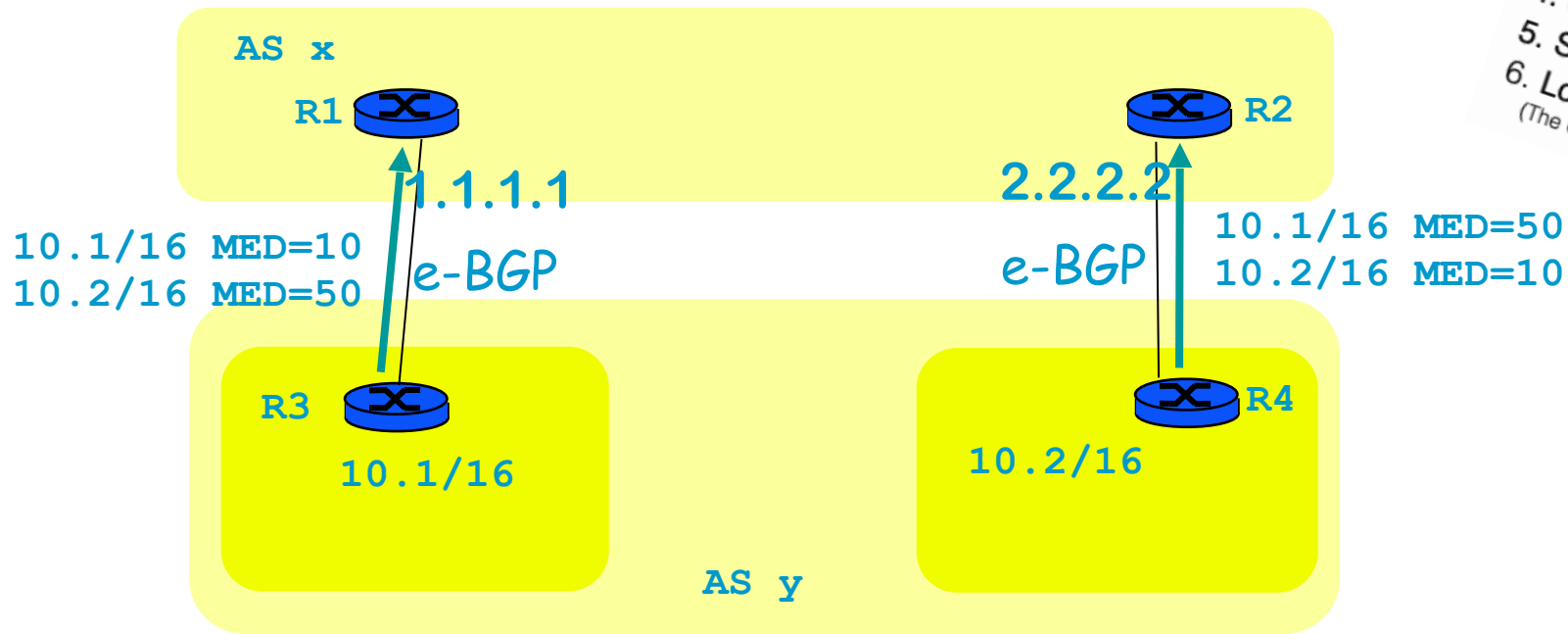
It remains **local** to this router

Never propagated to other routers, even in the same AS

Therefore there is no weight attribute in route announcements

- 
0. Highest weight (Cisco proprietary)
 1. Highest LOCAL-PREF
 2. Shortest AS-PATH
 3. Lowest MED, if taken seriously by this network
 4. e-BGP > i-BGP (= if route is learnt from e-BGP, it has priority)
 5. Shortest path to NEXT-HOP, according to IGP
 6. Lowest BGP identifier (router-id of the BGP peer from whom rc
(The Cisco and FRR implementation of BGP, used in lab 6, have additional

MULTI-EXIT-DISC (MED)



0. Highest weight (Cisco proprietary)
 1. Highest LOCAL-PREF
 2. Shortest AS-PATH
 3. Lowest MED, if taken seriously by this network
 4. e-BGP > i-BGP (= if route is learnt from e-BGP, it has priority)
 5. Shortest path to NEXT-HOP, according to IGP
 6. Lowest BGP identifier (router-id of the BGP peer from whom route is learnt)
- (The Cisco and FRR implementation of BGP, used in lab 6, have additional)

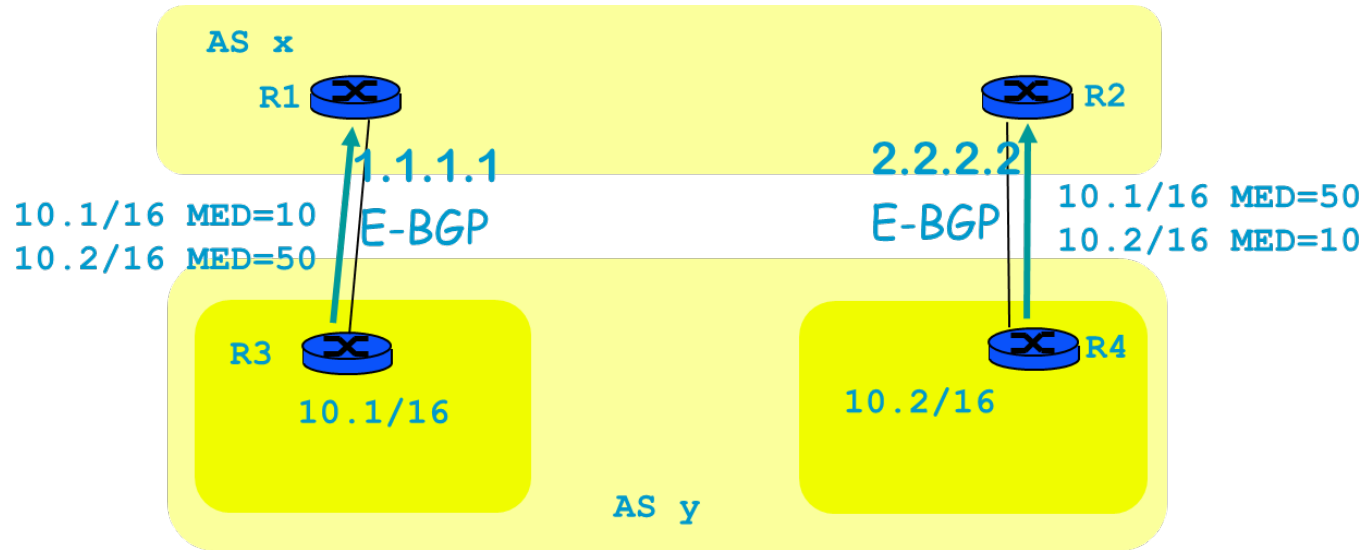
One AS connected to another over several links (*multi-homing*)

e.g.: multinational company y connected to worldwide ISP x

AS y advertises its prefixes with different MEDs (**lowest MED = preferable**)

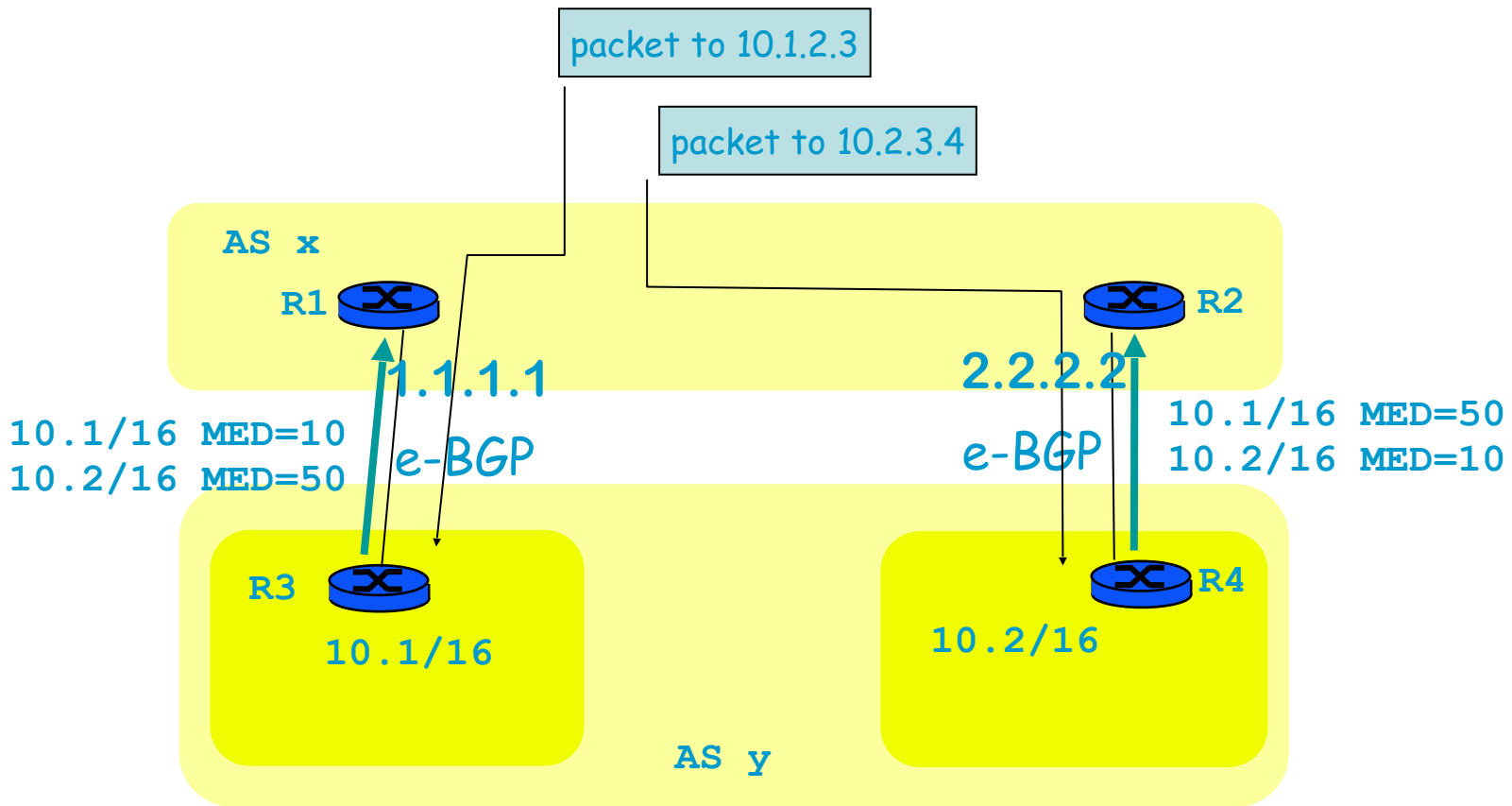
If AS x accepts to use MEDs put by AS y: traffic goes on preferred link

R1 has 2 routes to 10.2/16:
one via R3, learnt from R3 by E-BGP (MED=50),
one via R4, learnt from R2 by I-BGP (MED=10).
The decision process at R1 prefers ...



- A. The route via R2
- B. The route via R3
- C. Both
- D. I don't know

Solution



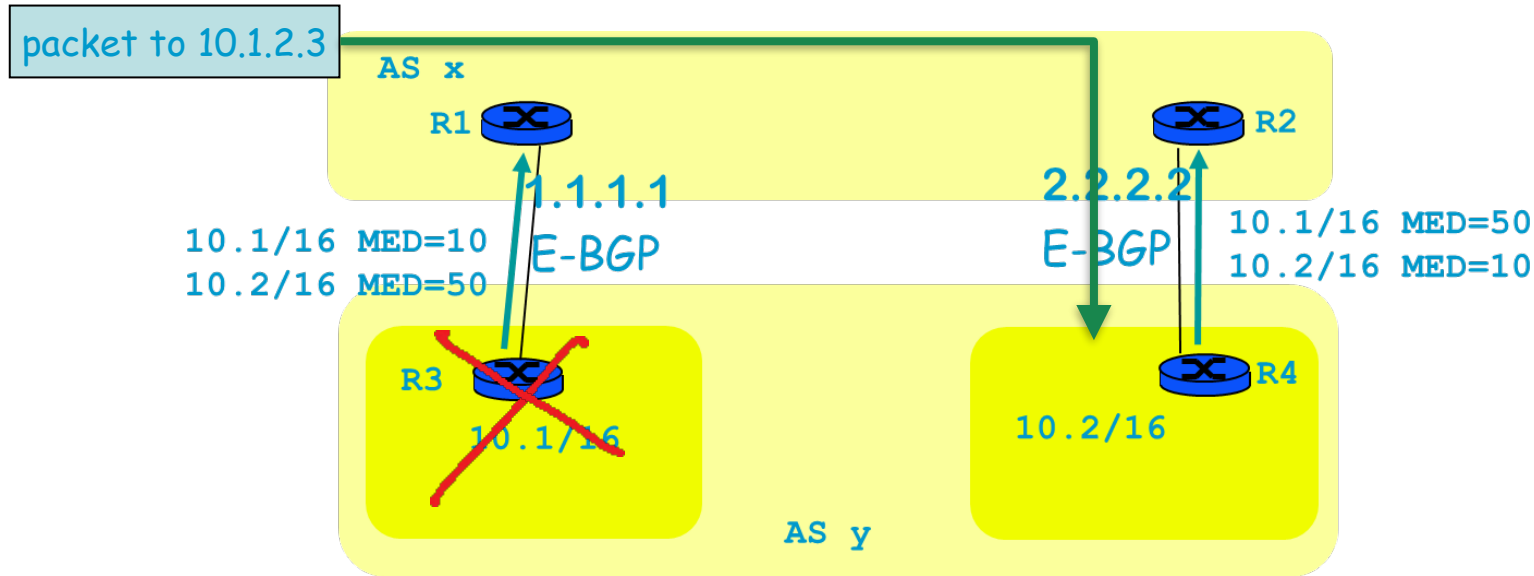
Answer A

R1 prefers the route via R2 because the decision process tests MED before e-BGP > i-BGP

Similarly, R2 has 2 routes to 10.1/16, R2 prefers the route via R1

Traffic from ASx to 10.1/16 flows via R1, traffic from ASx to 10.2/16 flows via R2

Router R3 crashes ...



R1 clears routes to ASy learnt from R3 (keep-alive mechanism) and selects as best route to 10.1/16 the route learnt from R2

R2 is informed of the route suppression by i-BGP

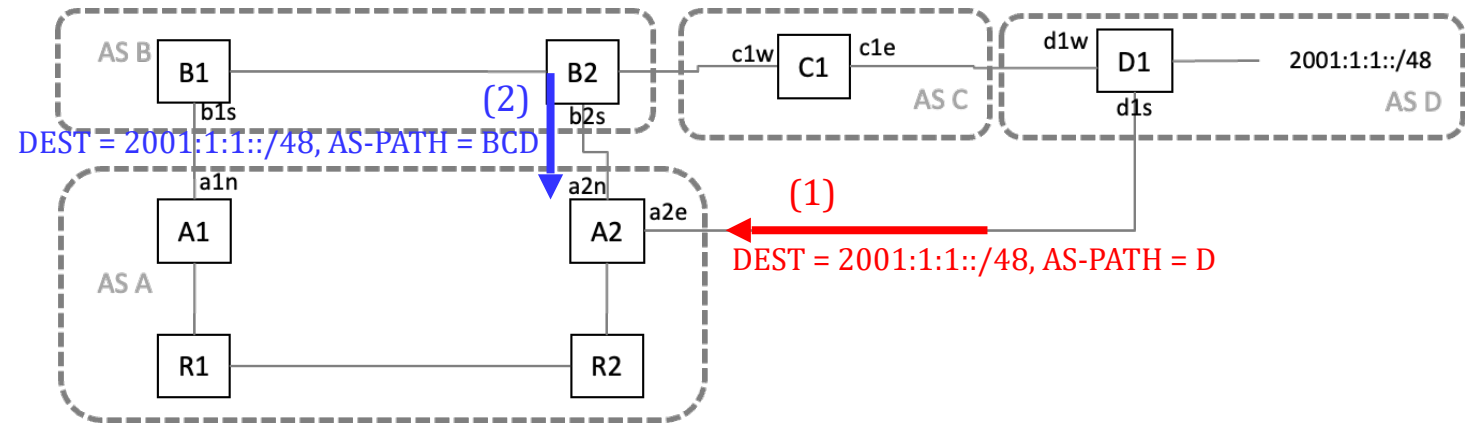
R2 has now only 1 route to 10.1/16 and 1 route to 10.2/16;

traffic to 10.1/16 now goes to R2

MED allows AS y to be dual homed and use closest link – other links are used as *backup*

Convergence of BGP

It is hoped that BGP converges and in practice it does, however there may be configurations with *no equilibrium* (*oscillations*) or with *multiple equilibria*:



Example: *A prefers B over D and sets LOCAL-PREF = 100 to updates received from B*

- If A2 receives (1) DEST = 2001:1:1::/48, AS-PATH = D from D1 before A receives any route to 2001:1:1::/48 from B then B2 receives DEST = 2001:1:1::/48, AS-PATH = A D, selects it as best route (prefers it over DEST = 2001:1:1::/48, AS-PATH = CD received from C, same AS-PATH length, smaller identifier) and sends nothing to A. A2's best route is DEST = 2001:1:1::/48, AS-PATH = D, NEXT-HOP = d1s
- If A2 receives (2) DEST = 2001:1:1::/48, AS-PATH = BCD from B2 before receiving a route to 2001:1:1::/48 from D, A2 stores it and will prefer it over any route to 2001:1:1::/48 received later from D. A2's best route is DEST = 2001:1:1::/48, AS-PATH = BCD, NEXT-HOP = b2s

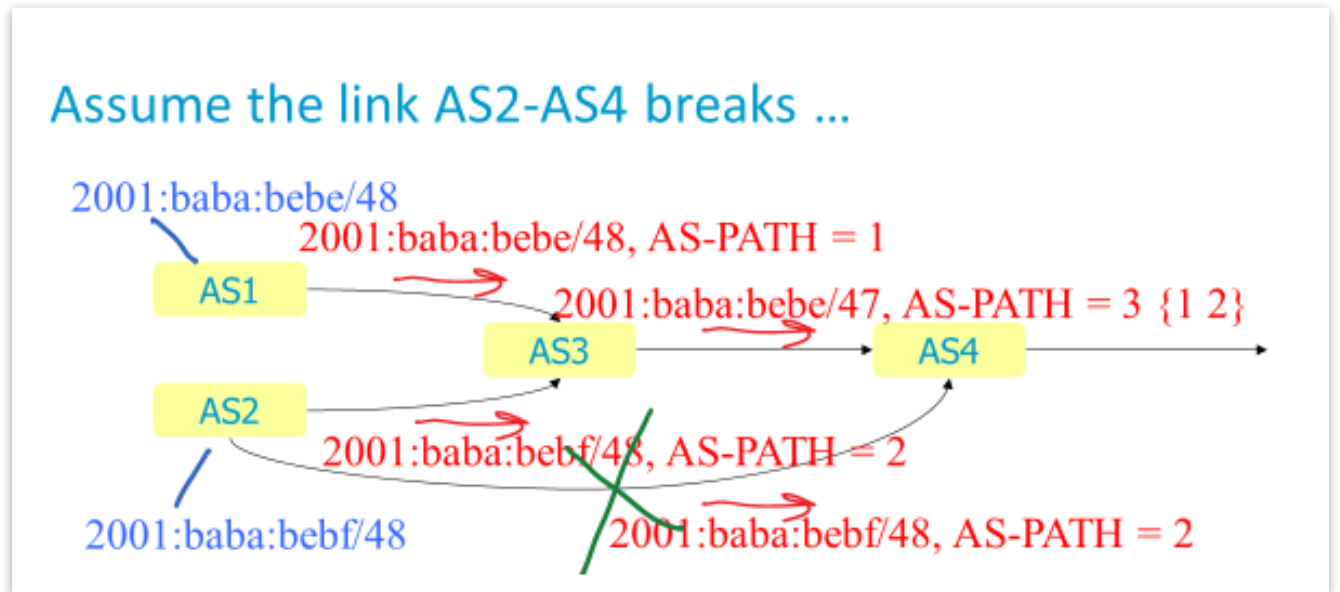
Two equilibria are possible here, depending on message timings/order.

Griffin, T.G. and Wilfong, G., 1999. An analysis of BGP convergence properties. *ACM SIGCOMM Computer Communication Review*, 29(4), pp.277-288

5. BGP: other bells and whistles

What happens when a BGP router loses its best route to some destination ?

- A. It will send an update in the next periodic KEEPALIVE message
- B. It sends a WITHDRAW update to the BGP peers to whom it had sent this route, as soon as possible
- C. It does not inform its BGP peers, they will recompute best routes and will find out
- D. I don't know



Solution

Answer B

BGP sends modifications to neighbors, including additions and withdrawals of best routes.

Route flap damping (or dampening)

Why?

Route flap: a route is successively withdrawn, updated, withdrawn, updated etc.

Caused e.g. by unstable BGP routers (crash, reboot, crash, reboot...) or by non convergence (oscillations).

- ▶ The flap propagates to the AS and to other ASes. Causes CPU **congestion** on BGP routers.

How?

Withdrawn routes are **kept** in Adj-RIN-in, with a **penalty** counter and a SUPPRESS state.

WITHDRAW \Rightarrow penalty incremented;

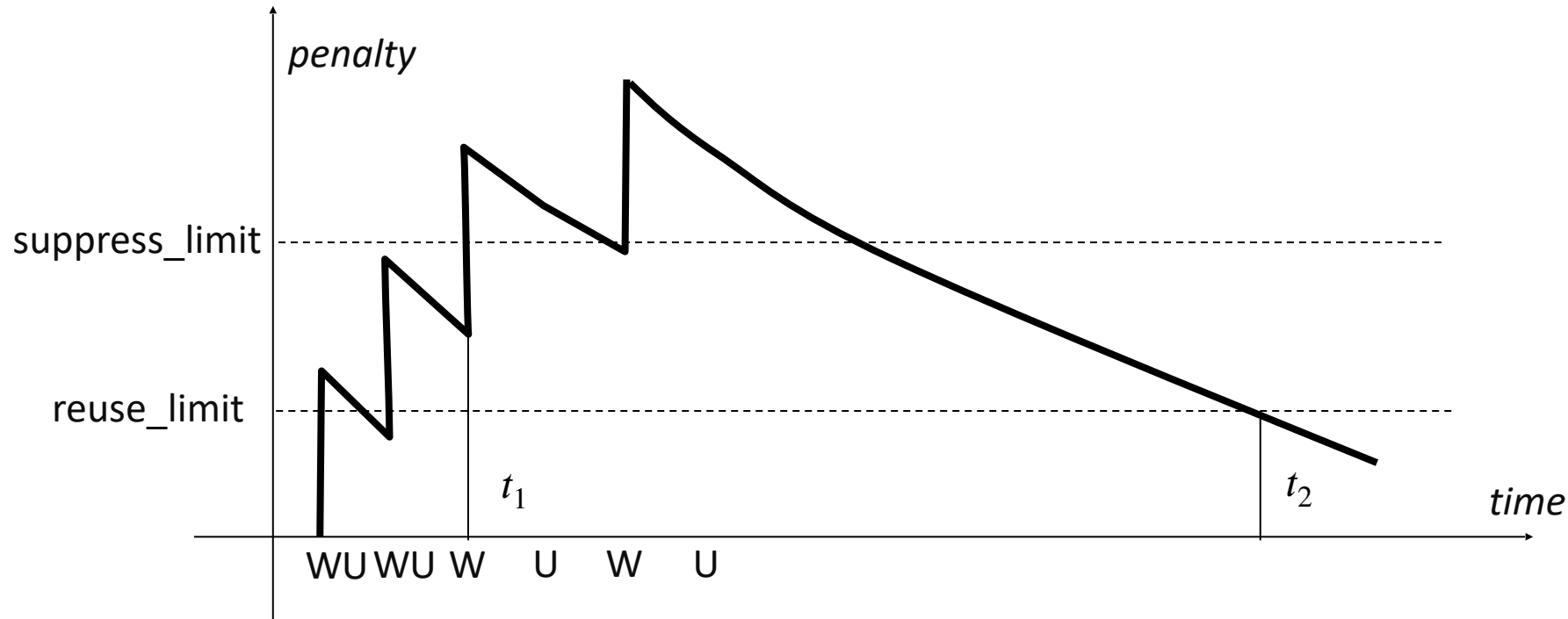
updated ADVERTISEMENT \Rightarrow if penalty > suppress_limit, then SUPPRESS = true

penalty is updated e.g. every <5 sec, with exponential decay; when

penalty < reuse_limit, then SUPPRESS = false and route is re-announced

routes that have SUPPRESS==true are ignored by the decision process

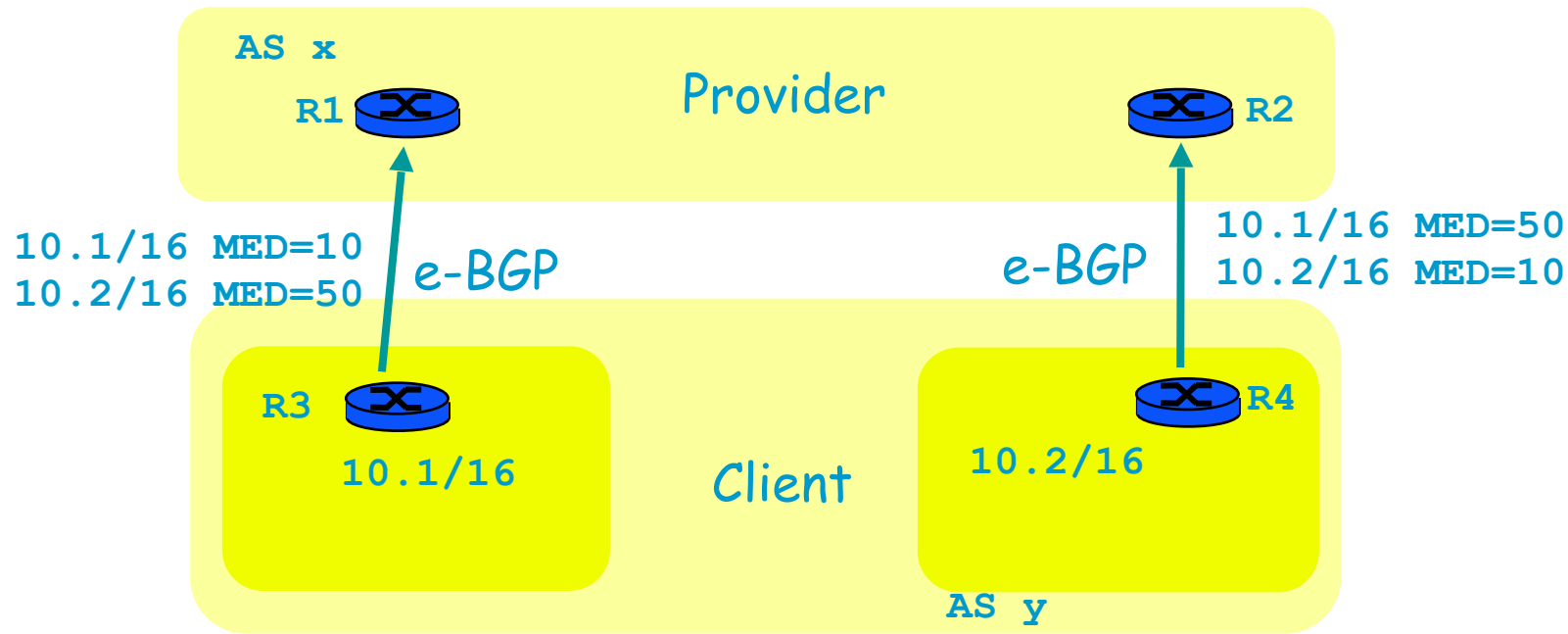
Route Flap Damping



W: reception of WITHDRAW, U: reception of updated advertisement

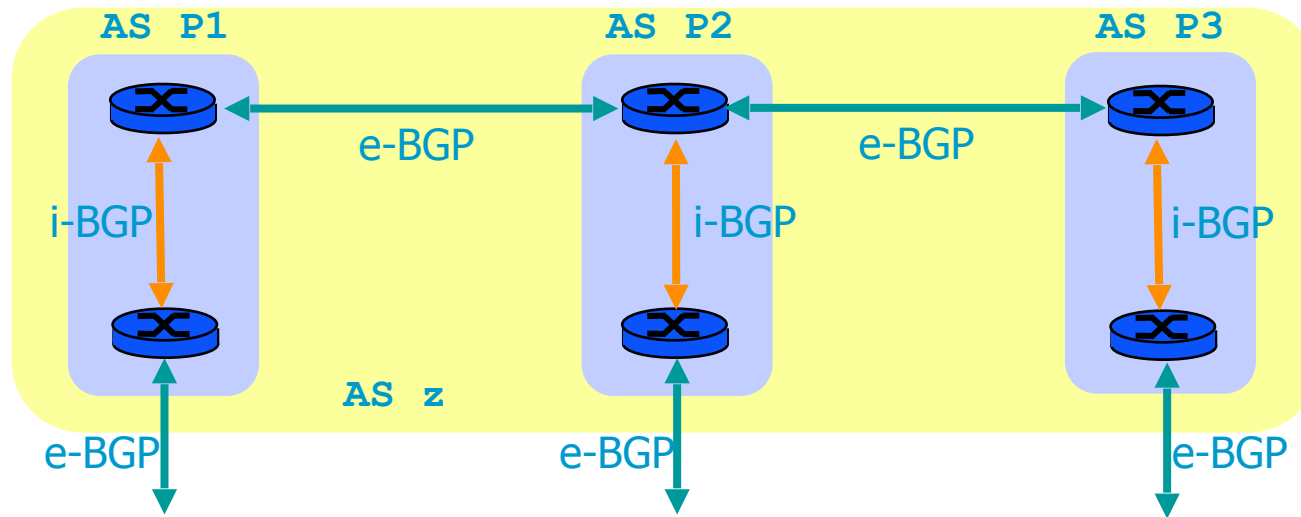
- in $[0, t_1]$ two flaps occur and propagate
- at t_1 the route has SUPPRESS = true
- in $[t_1, t_2]$ the route is ignored
- at t_2 the route has SUPPRESS = false and is used again

Private AS Number



- Client uses BGP with MED to control flows of traffic (e.g provider should use R1-R3 for all traffic to 10.1/16)
- Stub domains (e.g., EPFL) can use a *private AS number* -- not usable in the global internet, used only between Client and Provider (e.g., SWITCH)
- Provider *translates* this number to his own when exporting routes to the outside world.
- Client does not need a public AS number.

Avoiding i-BGP Mesh: Confederations

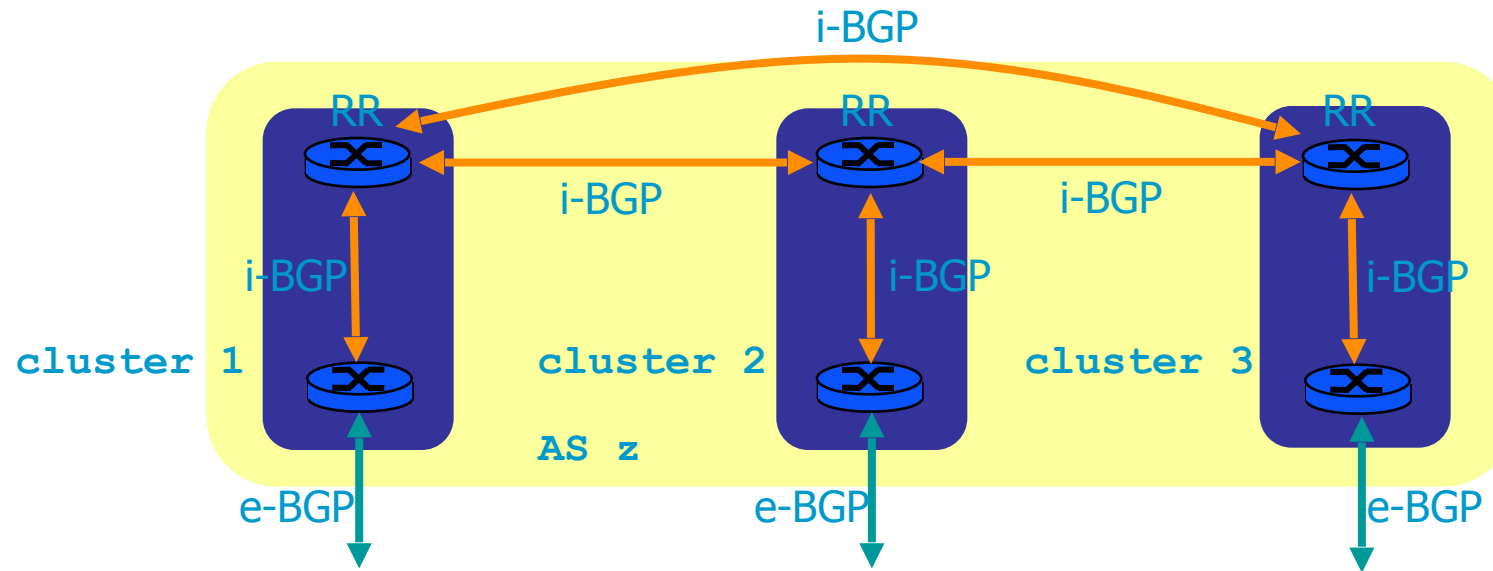


AS **decomposed** into sub-AS with private AS number
Similar to OSPF areas

i-BGP inside sub-AS (full interconnection)

e-BGP between sub-AS

Avoiding i-BGP Mesh : Route reflectors



Cluster of routers

- one i-BGP session between each client and a route reflector (RR)

Route reflector acts like a proxy:

- re-advertises a route learnt via i-BGP

This architecture results in fewer iBGP internal peerings (no mesh, but hierarchy), and avoids loops

CLUSTER_ID attribute associated with the advertisement

An Interconnection Point



[E-Mail](#) | [Credits](#)

[Expand all](#) | [Collapse all](#)

General Information

Services

Costs

[Membership fees](#)

[Connection fees](#)

Legal

[Articles of association](#)

[Peering Policy](#)

[Connection agreement](#)

Members

[Member list](#)

[Board members](#)

[Membership application](#)

Member Login

Tech Corner

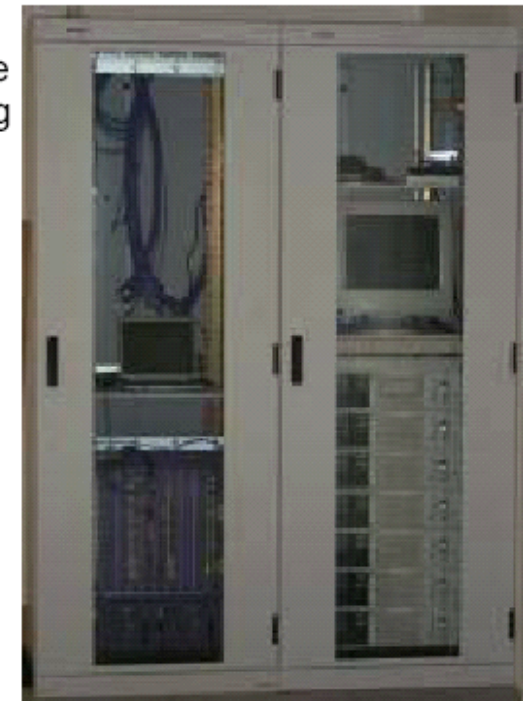
Links

Welcome to swissix

The Swissix (Swiss Internet Exchange) in Zurich, Switzerland, is now open. We are pleased to welcome ISPs and hosting companies as members and peering partners.

With continued growth of Internet traffic, we want to make sure that there is sufficient reliability built into the Swiss Internet. By exchanging traffic at multiple exchanges points, you can help ensure that consumers have fast Internet access and network operators have multiple routes for their traffic flows.

The Swiss Internet Exchange (swissix) is a neutral and independent exchange and a place for Internet Service Providers (ISPs) to interconnect and exchange IP traffic with each other at a national or international level.



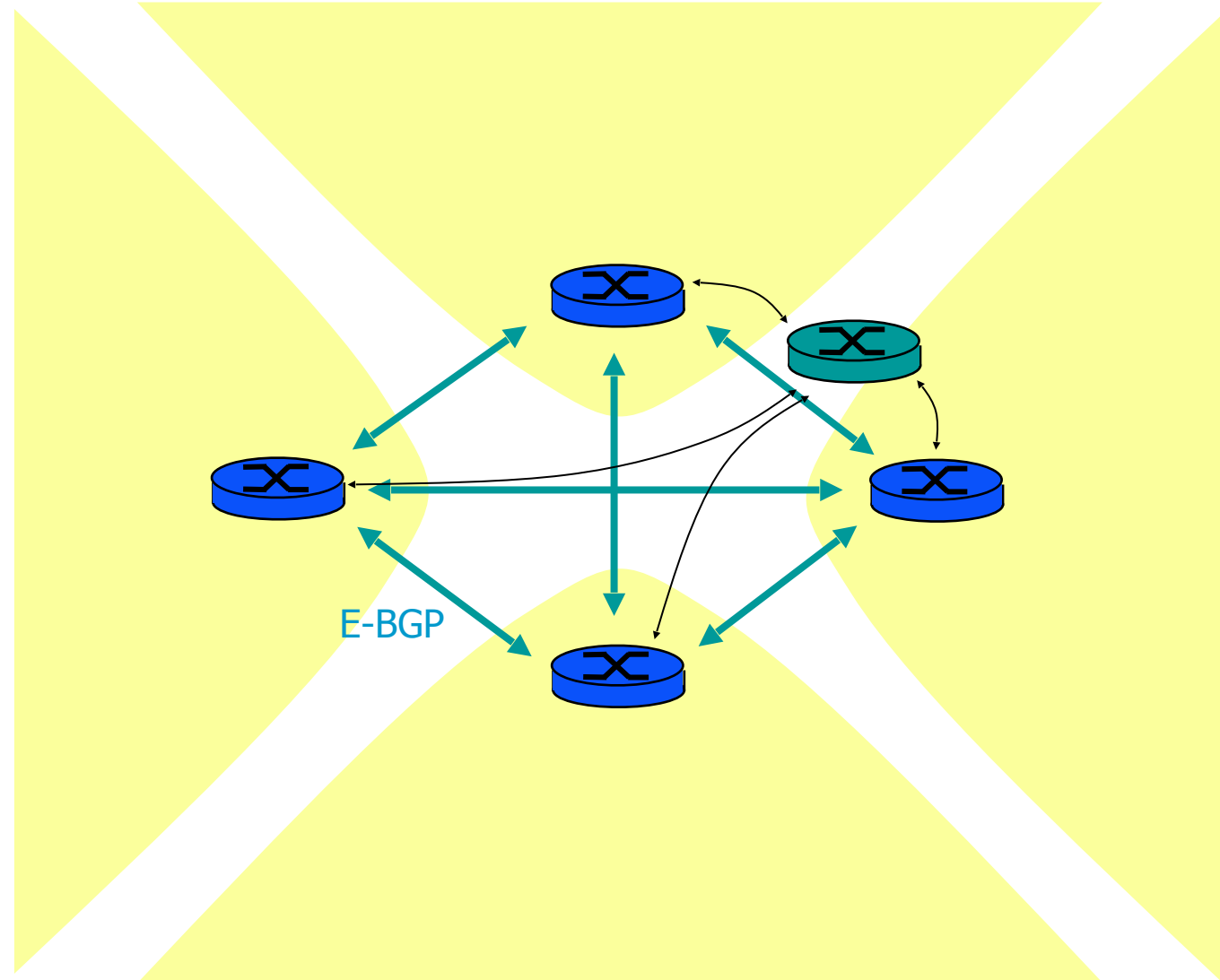
Avoiding e-BGP mesh: Route server

Problem: At an interconnection point, there might be a large e-BGP mesh

Instead of $n(n-1)/2$ peer-to-peer E-BGP connections, we use n connections to Route Server (similarly to reflectors for i-BGP)

To avoid loops ADVERTISER attribute indicates which router in the AS generated the route

Many route servers publish their advertisements



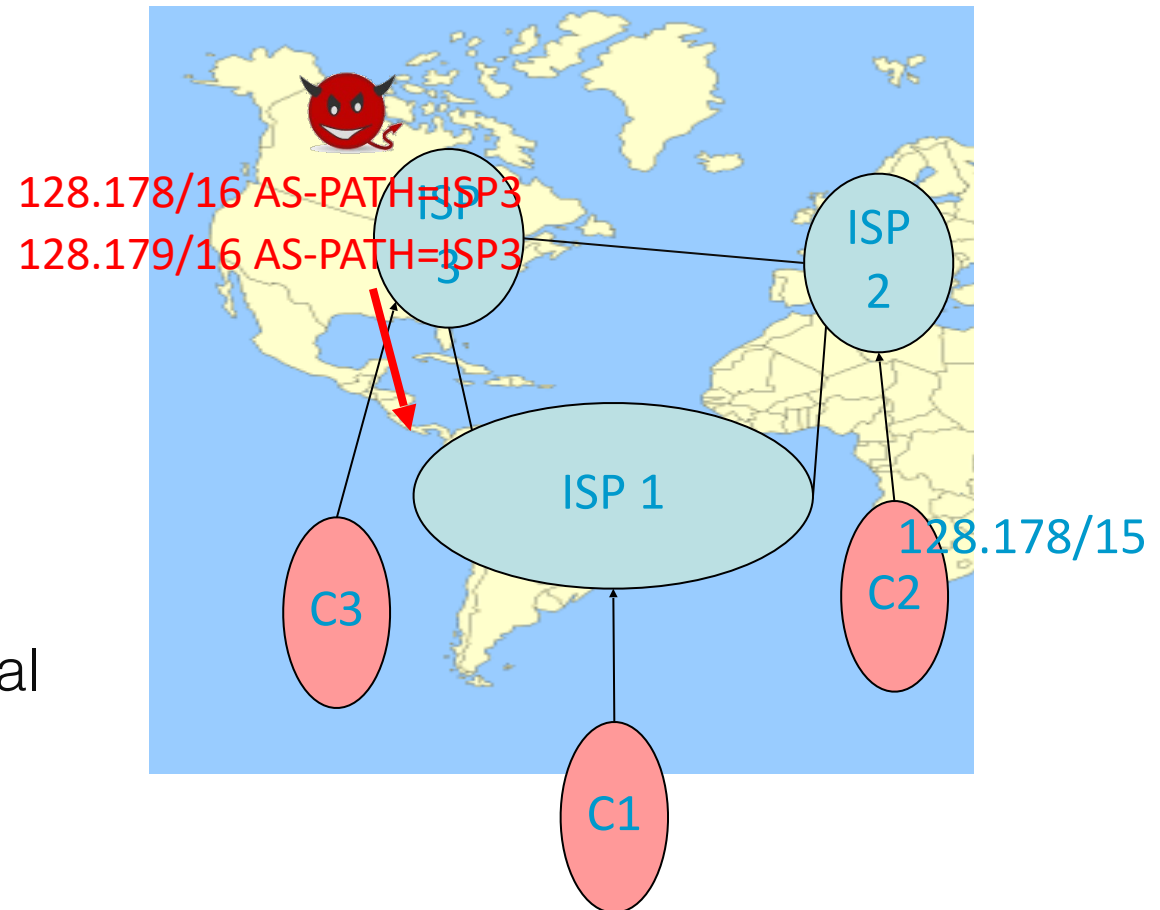
6. Security Aspects

Malicious or simply buggy BGP updates may cause damage to global internet

Example 1 (**subprefix hijack**): Assume ISP3 (malicious) announces to ISP1 a route to 128.178/16 and a route to 128.179/16 (both are EPFL prefixes)

What will happen to traffic from C1 to EPFL (i.e. C2 in the figure)?

- A. All such traffic will go to ISP3
- B. Some fraction will go to ISP3
- C. All such traffic will go to C2, as usual
- D. I don't know



Solution

Answer A or B

- If aggregation is not done by ISP1, the routes to 128.178/16 and 128.178/15 are different. By longest prefix match, all traffic to 128.178/16 (and to 128.179/16) will follow the bogus route to ISP3, who may simply discard all packets – this is called **subprefix hijack** and will cause EPFL to be unreachable from ISP1 and its customers.
- If aggregation is performed by ISP1, there are now 2 competing routes and either can be chosen, depending on the specific policy rules inside ISP1 (hot potato routing or not) leading to partial loss of traffic

BGP Security

Forged AS paths, destination prefix, next-hop etc cause traffic to go to malicious ISP -> used to deny service / spy / forge

BGP security measures:

- **Routing Registries**: PTI (IANA/ICANN, internet number authority) manages address allocations / delegated to 5 Regional Internet Registries, RIRs (for Europe: RIPE); RIPE maintains a public Routing Registry, database of address blocks + some policy information. Cooperation of Routing Registries = the Internet Routing Registry (IRR). ASes can read Routing Registries and use them to verify the routes received from BGP peers not cryptographic, best effort.

Origin Validation: ROA

Owner of an address block creates a (cryptographically signed) Route Origin Authorization (ROA) that contains AS number and IP address block; this validates origination - prevents bogus origination. More secure than IRR.

Uses the RPKI (resource public key infrastructure) rooted at IANA/ICANN and deployed in RIRs.



Example: Switch receives block 2001:620::/32 from RIPE (European authority), obtains a certificate from RIPE, and uses it to create and publish ROA for this block. Any AS can verify the ROA using the certificates of ICANN and RIPE.

try it: `whois -h whois.bgpmon.net 128.178.0.0/15` (EPFL's IPv4 block)
`whois -h whois.bgpmon.net 2001:620::/32` (Switch's IPv6 block)

Beyond ROA: Validation of Path with BGPsec

BGPsec authenticates the entire AS-PATHs contained in a BGP update

Between E-BGP peers

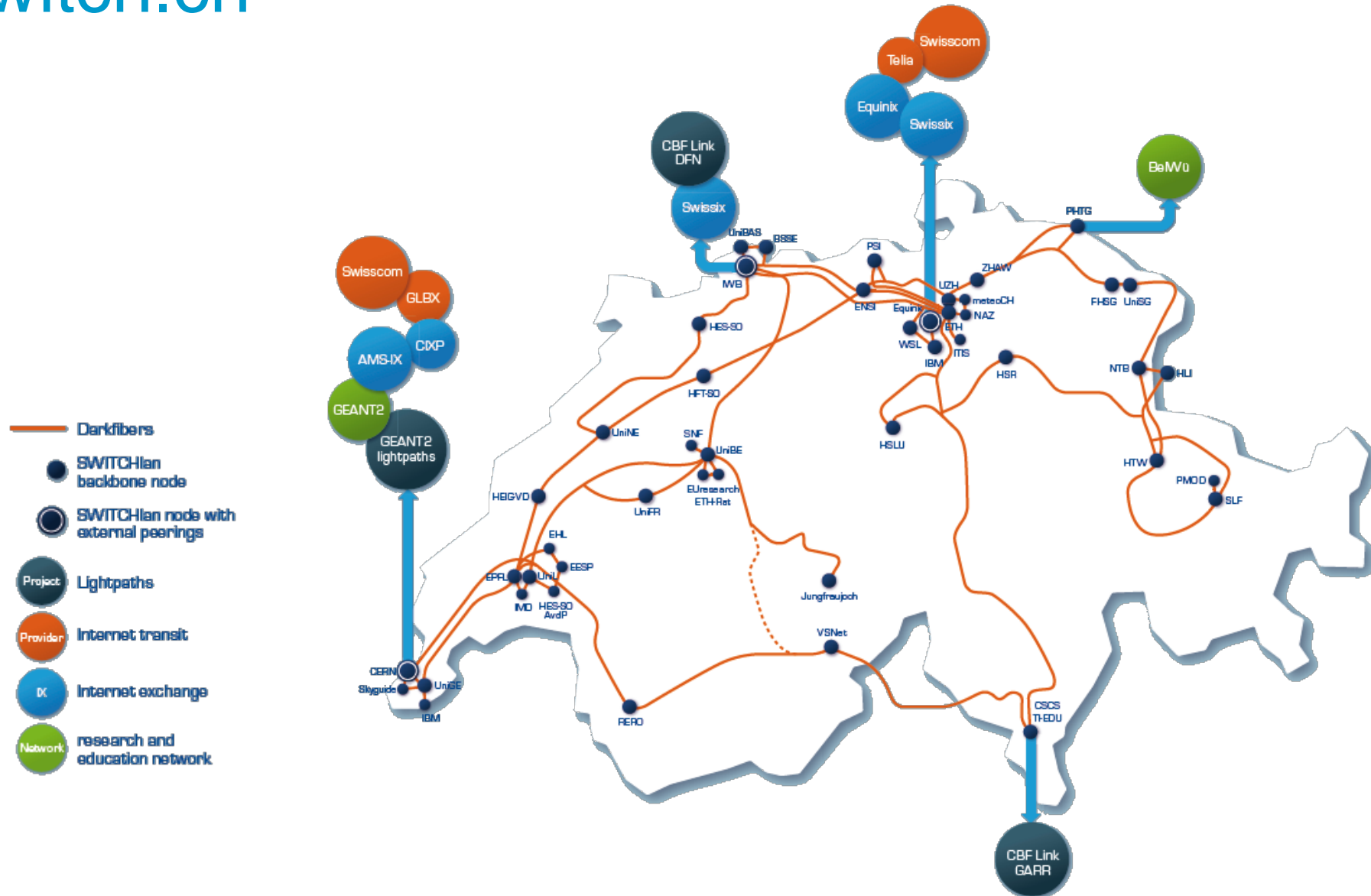
AS-PATH attribute replaced by BGPsec_Path attribute that contains the AS path + signatures of every segment of the path performed by every intermediate AS

Deployment in progress.

SCION (<https://scion-architecture.net>, ETHZ, Adrian Perrig) is an alternative to BGP (and to IP) that uses source routing and systematic encryption.

C. Illustrations: The Switch Network

www.switch.ch



BGP Toolkit Home

ks

Home

[Home](#)

[Report](#)

[Report](#)

[Report](#)

[S](#)

[Routes](#)

[Port](#)

Welcome to the Hurricane Electric BGP Toolkit.

You are visiting from **2001:620:618:197:1:80b2:9771:1**

Announced as **2001:620::/32** (SWITCH)

Announced as **2001:620::/29** (SWITCH)

Your ISP is **AS559** (SWITCH)

2001:620::/32

hks

Network Info

Whois

DNS

IRR

[Home](#)

[Report](#)



[Report](#)

[Report](#)

[as](#)

[t](#)

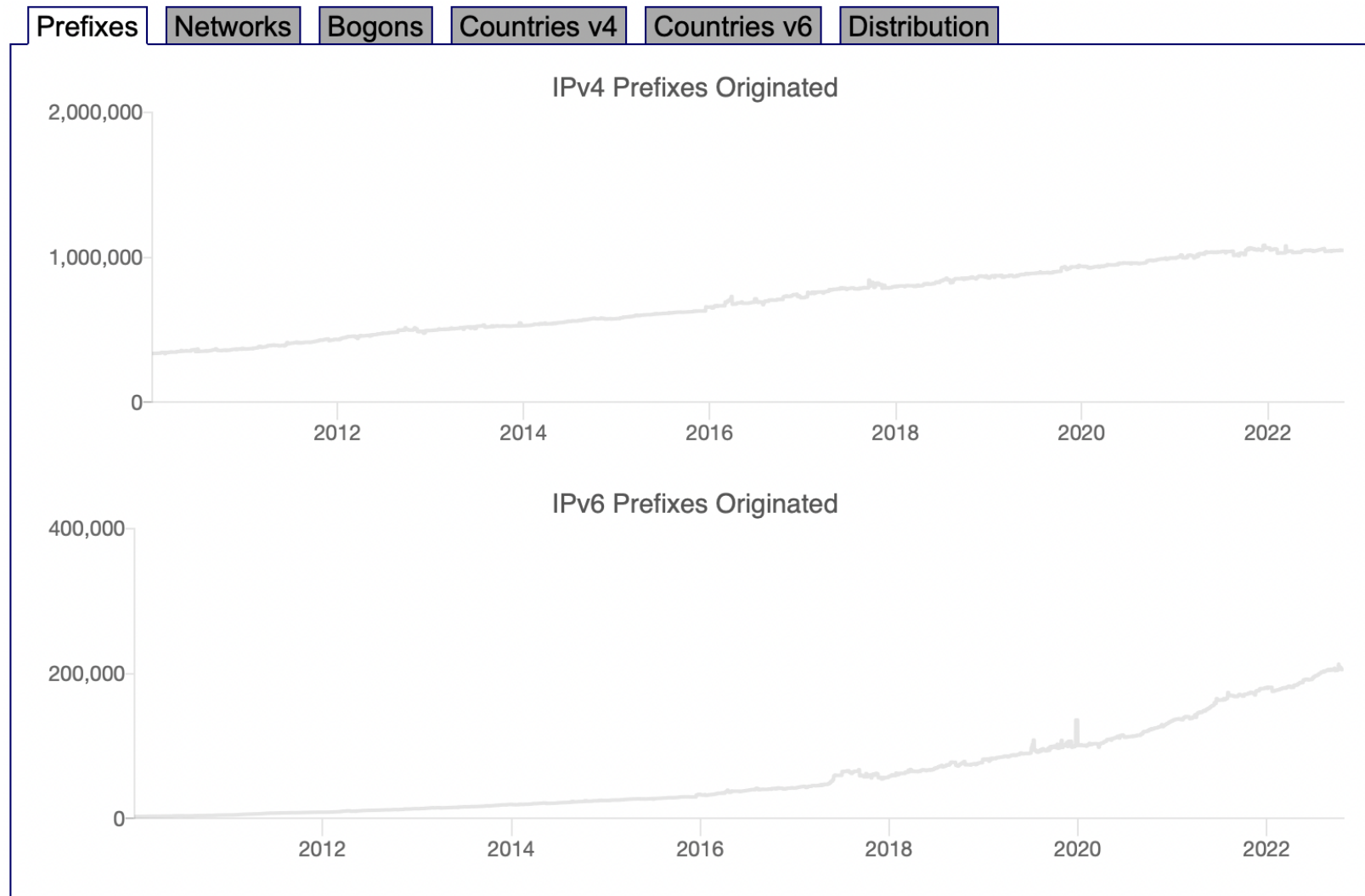
Announced By

| Origin AS | Announcement | | Description |
|-----------------------|-------------------------------|---|-------------|
| AS559 | 2001:620::/32 |   | SWITCH |



ROA signed and valid

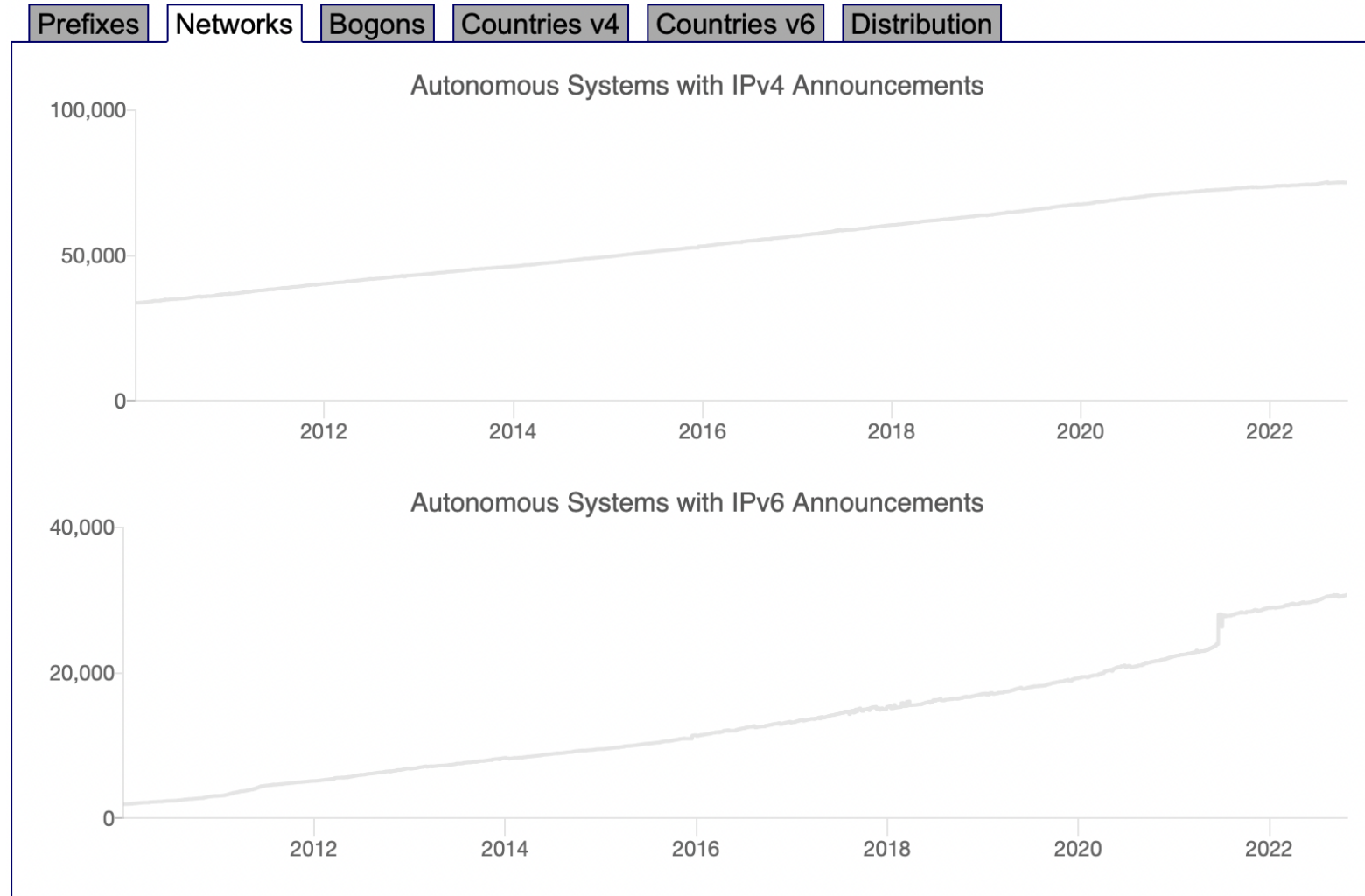
Number of announced prefixes



Updated 24 Oct 2022 17:08 PST © 2022 Hurricane Electric

seen by Hurricane Electric: bgp.he.net/report/prefixes, sampled on 2022 Oct 24

Number of ASs



Updated 24 Oct 2022 17:08 PST © 2022 Hurricane Electric

seen by Hurricane Electric: bgp.he.net/report/prefixes, sampled on 2022 Oct 24

Conclusion

BGP integrates different ASs

Interface BGP-IGP is complex and has many subtleties

Security of BGP is an active area of research and development

Beyond BGP:

SCION (<https://scion-architecture.net>, ETHZ, Adrian Perrig) is an alternative to BGP (and to IP) that uses source routing and systematic encryption. Aims to provide more security and flexibility in choice of routes.