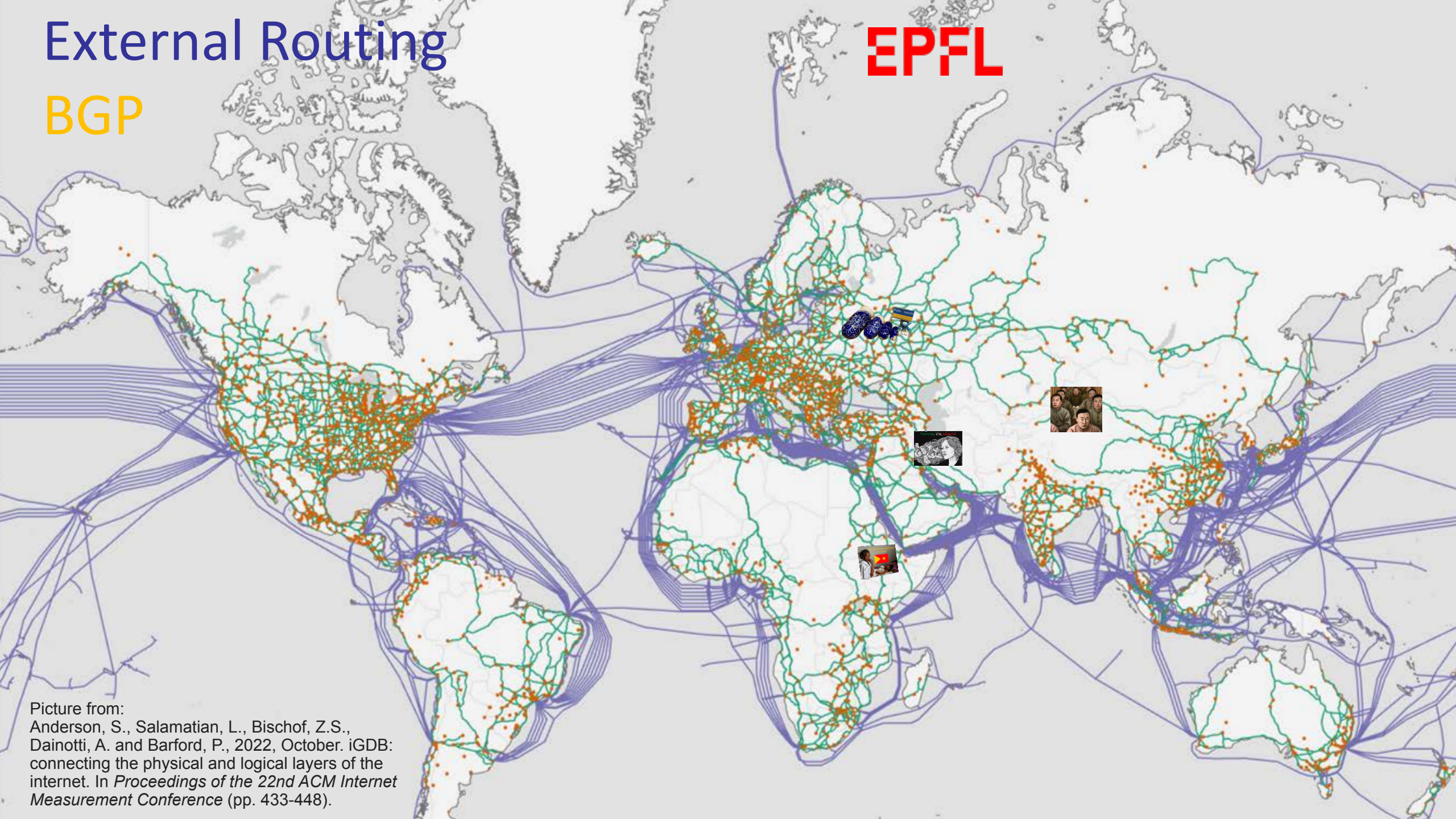
External Routing

BGP



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

Recall: BGP is an example of path vector routing

Path Vector

- Every router knows only:
  - its neighbors +
  - explicit *paths* to all destinations
- Optimization criterion is *not only cost*
- Typically used for *inter-domain routing*, where it is hard to assign costs
  - e.g. BGP computes domain-level paths
  - and each router selects best path to each destination according to *multiple criteria*
  - and populates forwarding table

Part A: BGP at high level

1. Inter-Domain Routing

Context

The Internet is *too large* + *heterogeneous* (i.e. it is split into multiple different domains) to be run by one routing protocol.
We use hierarchical routing instead:

- *between* domains, we use BGP (= Border Gateway Protocol)
- *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

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*

Domains — terminology

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

AS = Autonomous System = ARD with a number ("AS number")

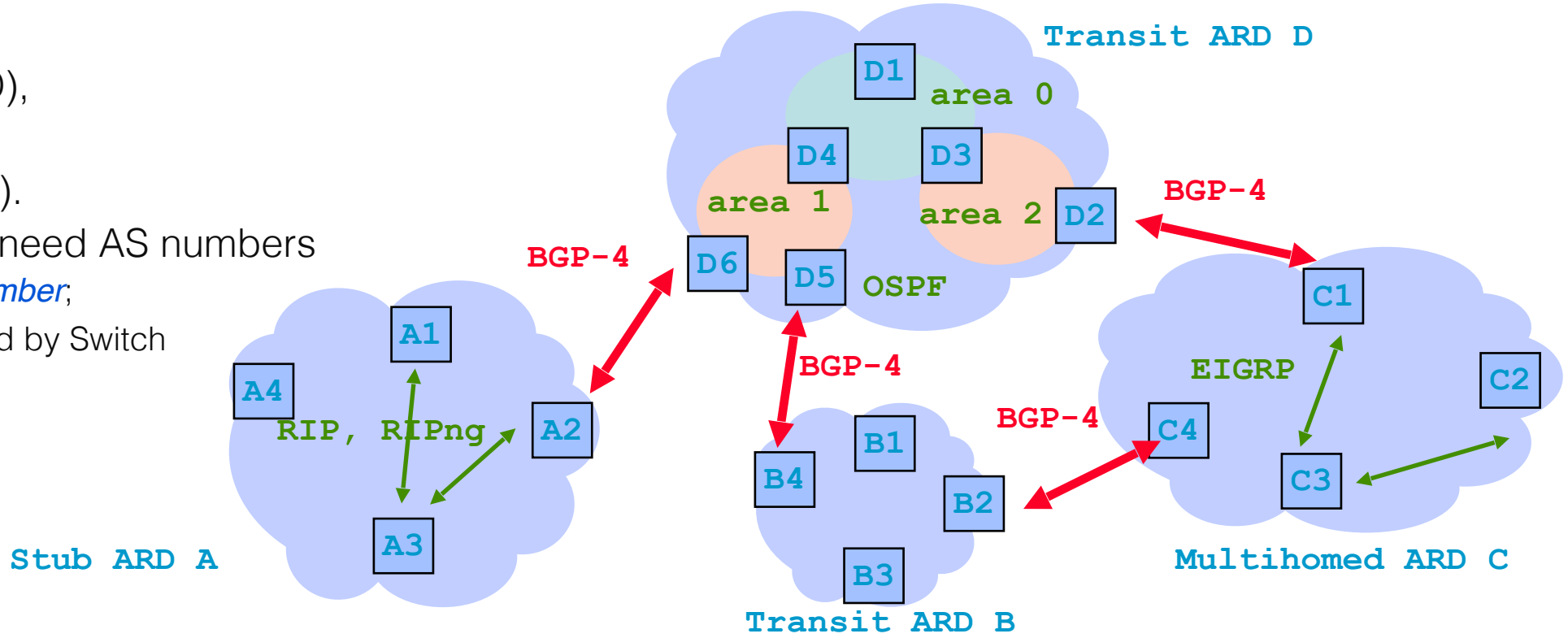
- AS number is 32 bits, written in 2-field dotted decimal notation: e.g. 23.3456, and leading zeros may be omitted: e.g. 0.559 means 559
- Private AS numbers are: 0.64512 – 0.65535
- Real examples: AS1942 - CICG-GRENOBLE, AS2200 - Renater
  AS559 - SWITCH Teleinformatics Services

ARDs can be:

transit (see B and D),
stub (see A) or
multi-homed (see C).

Only non-stub domains need AS numbers

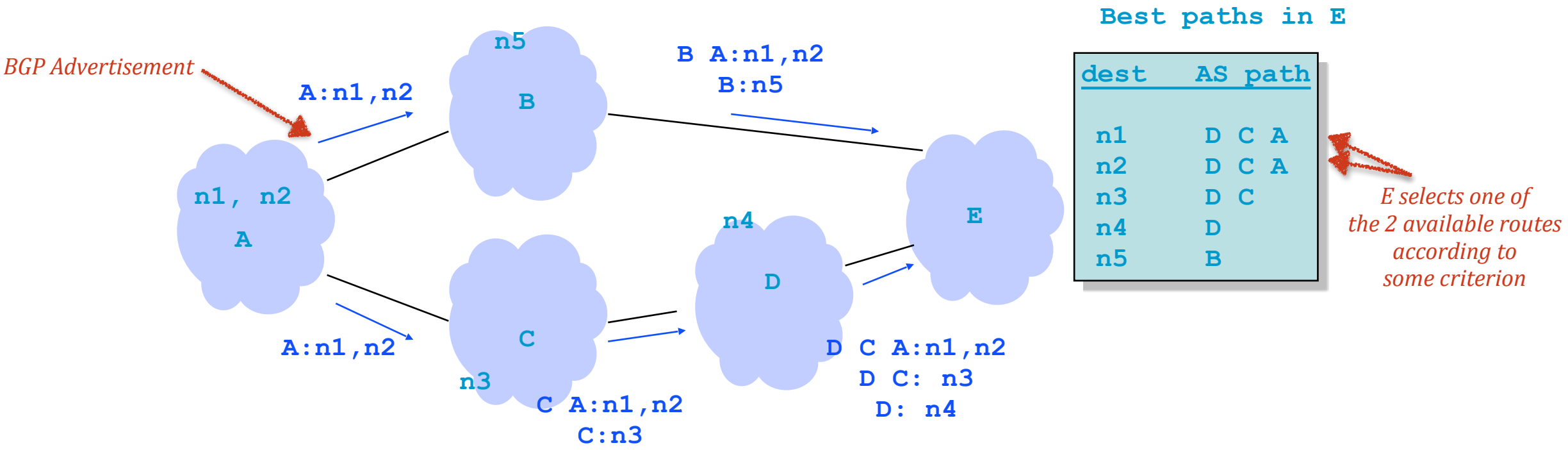
e.g. EPFL: ARD *w/o number*;
all external traffic served by Switch



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 exports. Each AS uses its *own criteria* for deciding which path is the best.



Not all routers run 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 over *TCP* connections:

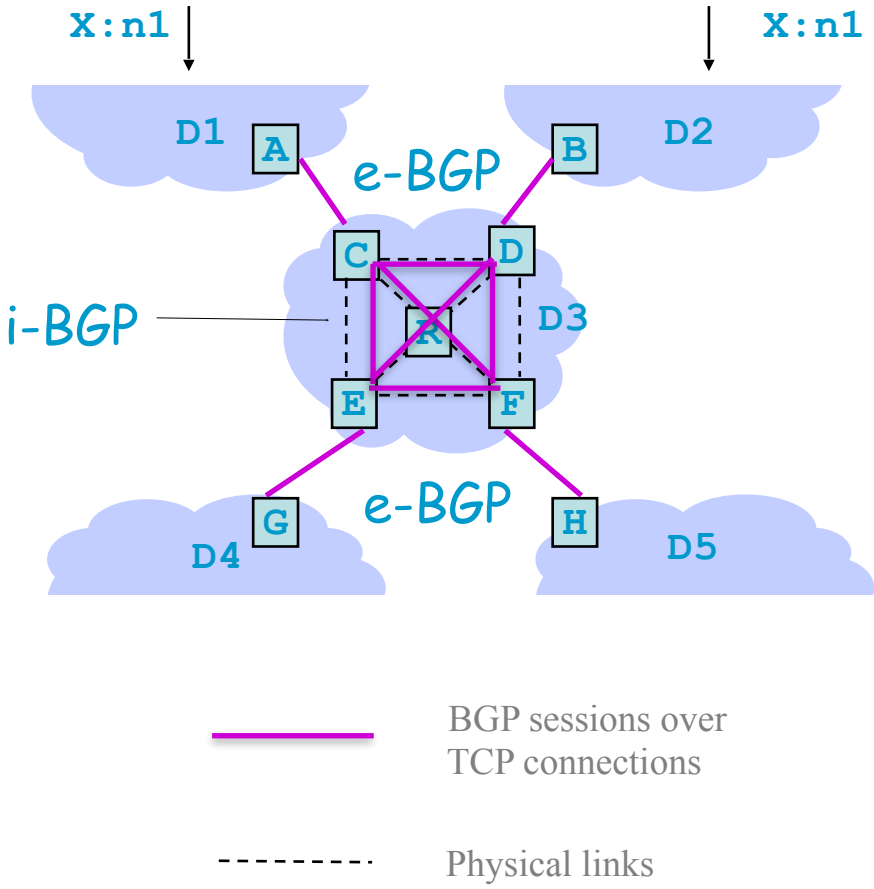
- *externally* (e-BGP)

  to advertize routes to neighbor domains [as in previous slide]
- *internally* (i-BGP)

  to exchange what they have learnt from e-BGP

In i-BGP, BGP peers

- communicate via a mesh network, called "BGP mesh"
- do the same as in e-BGP but 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.

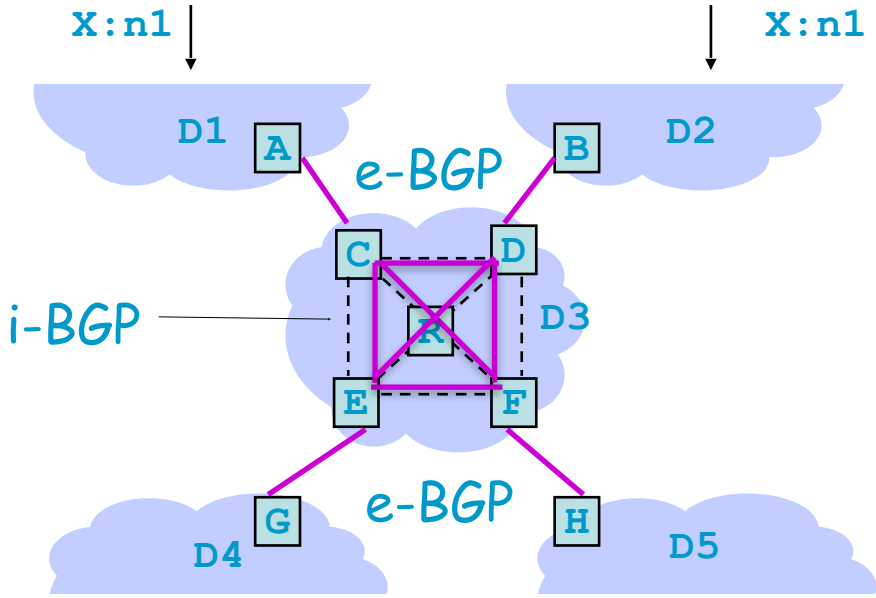
External peers are typically “on link”.

Internal peers need not be “on link”.

Which BGP updates may be sent ?

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

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



BGP sessions over TCP connections

Physical links

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

Solution

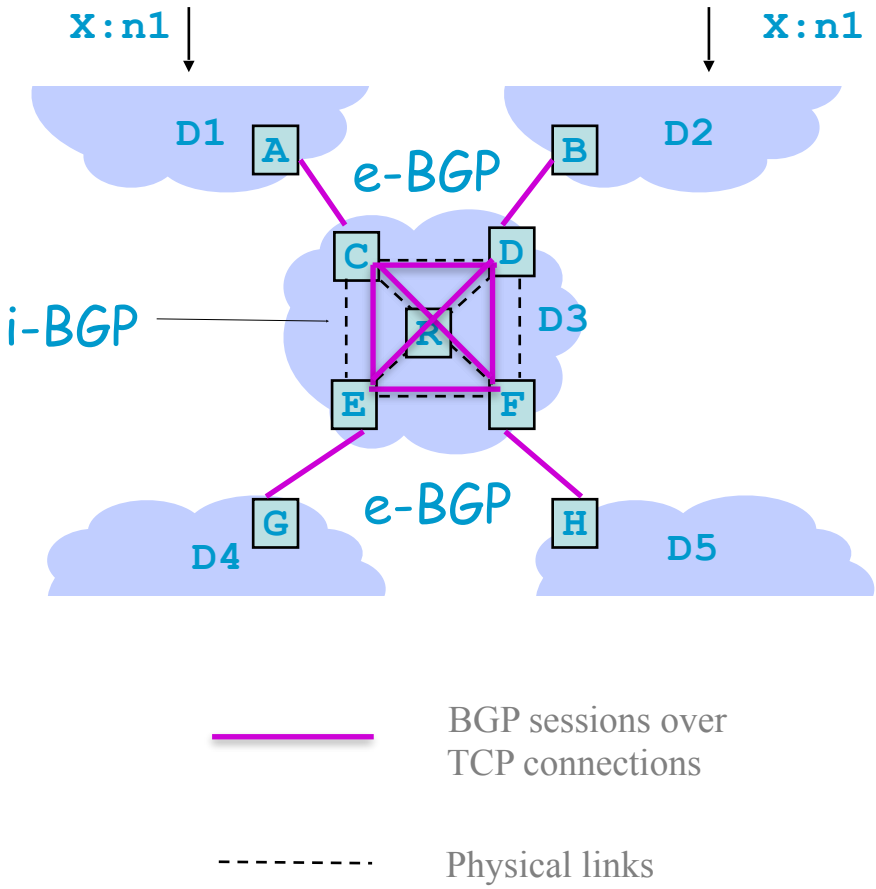
1. $C \rightarrow A : D3 - D2 - X : n1$
2. $D \rightarrow E : D2 - X : n1$
3. $C \rightarrow E : D2 - X : n1$

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.



2. Policies

Context

Self-organized interconnection of ASs (= peering) —> free market

- point to point links between networks: e.g., EPFL to Switch, Switch to Telianet
- interconnection points: all participants run a BGP router in the same LAN.
  NAP (Network Access Point), MAE (Metropolitan Area Ethernet), CIX (Commercial Internet eXchange), GIX (Global Internet eXchange), IXP, SFINX, LINX...

Various financial relationships:

- customer-provider (hierarchy): e.g. EPFL is customer of Switch—EPFL *pays* Switch
- shared-Cost peers (same level): e.g. Swisscom and Switch are peers; they collaborate to serve their customers, typically *without paying* each other
- also many other, depending on (private) business agreements

Policy rules…

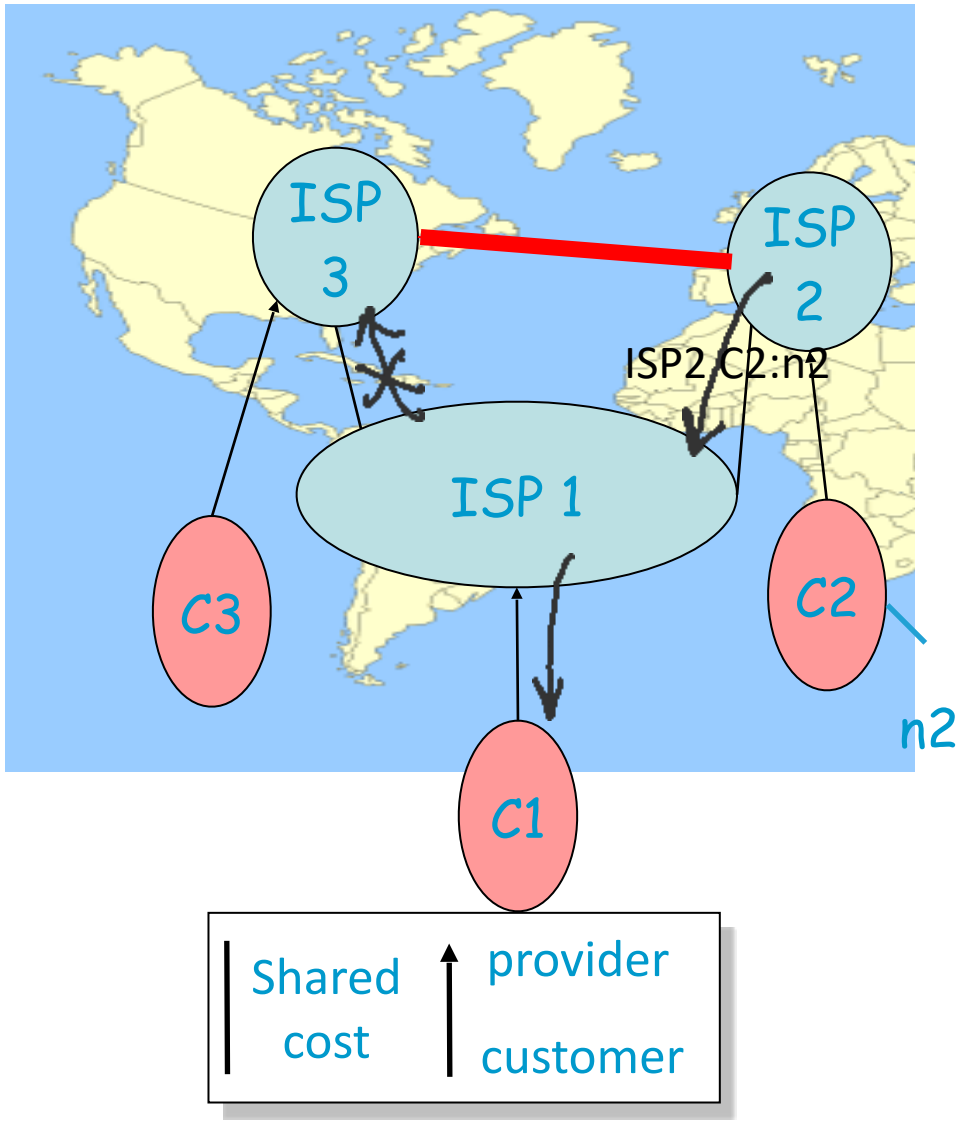
…implement such relations & business agreements

via:

import (what to accept) and export rules (what to advertize?),

and a decision process

Policies (high-level example)



Suppose:

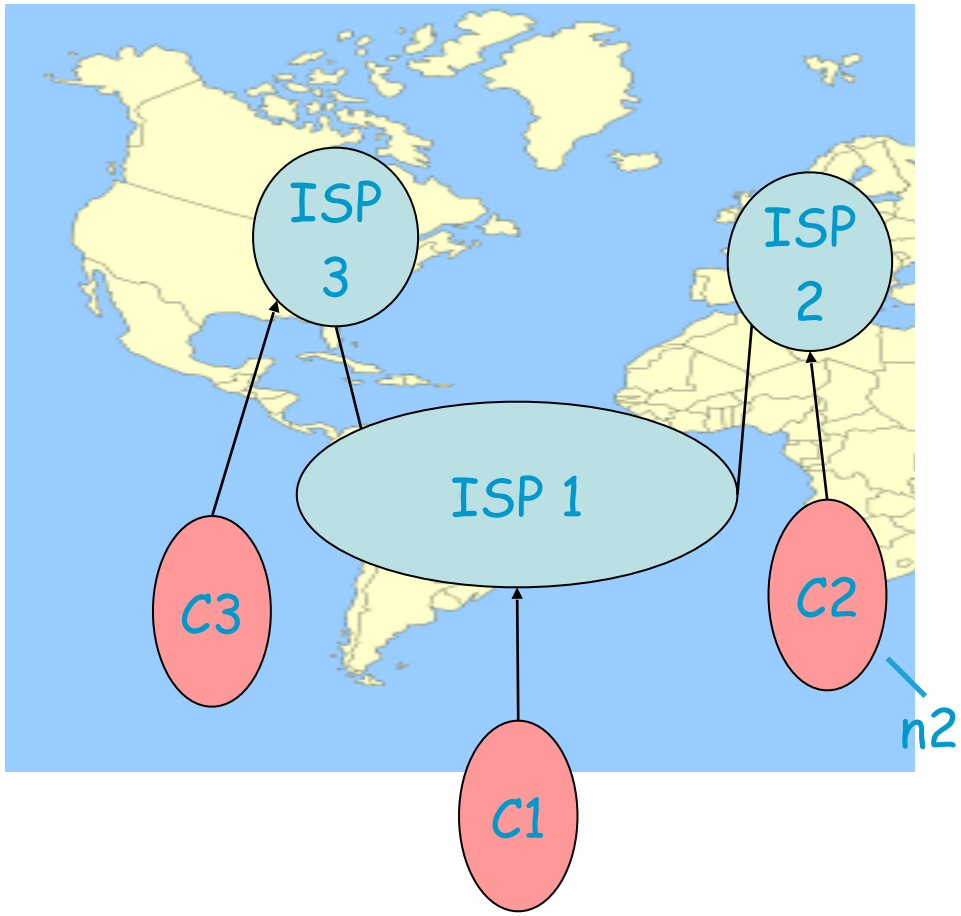
- All ISPs are shared-cost peers; Ci is customer of ISPi.
- ISP3-ISP2 is a transatlantic link, cost-shared between ISP2 & ISP3, but it is expensive;
- ISP3-ISP1 is a local, inexpensive link;
- Problem: It is advantageous for ISP3 to send traffic to n2 via ISP1; but… ISP1 may not agree to carry traffic from C3 to C2.
  How can ISP1 apply such a *policy*:
  - "transit service" to C1 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

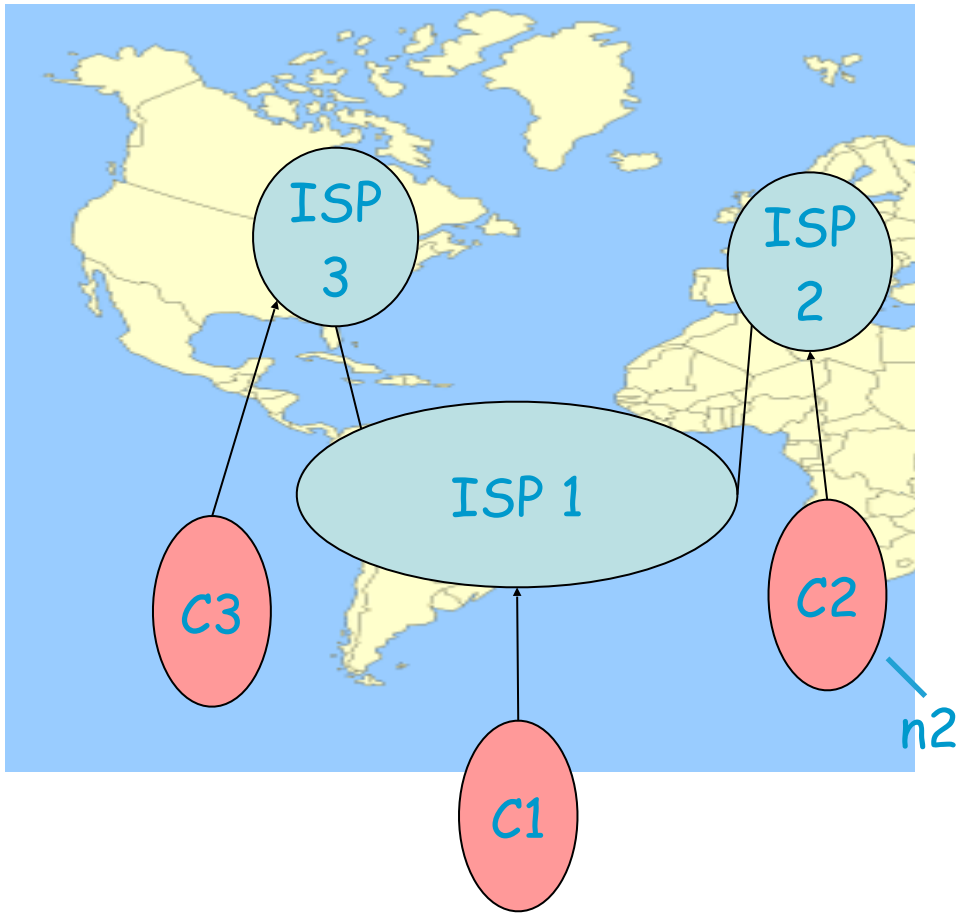


ISP1-ISP2 and ISP1-ISP3 are peers;
ISP2-ISP3 are *not* peers nor customers/providers.
All apply the rule “Routes coming from peers or providers are not propagated to peers or providers”.
What is a valid path from C2 to C3 ?

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

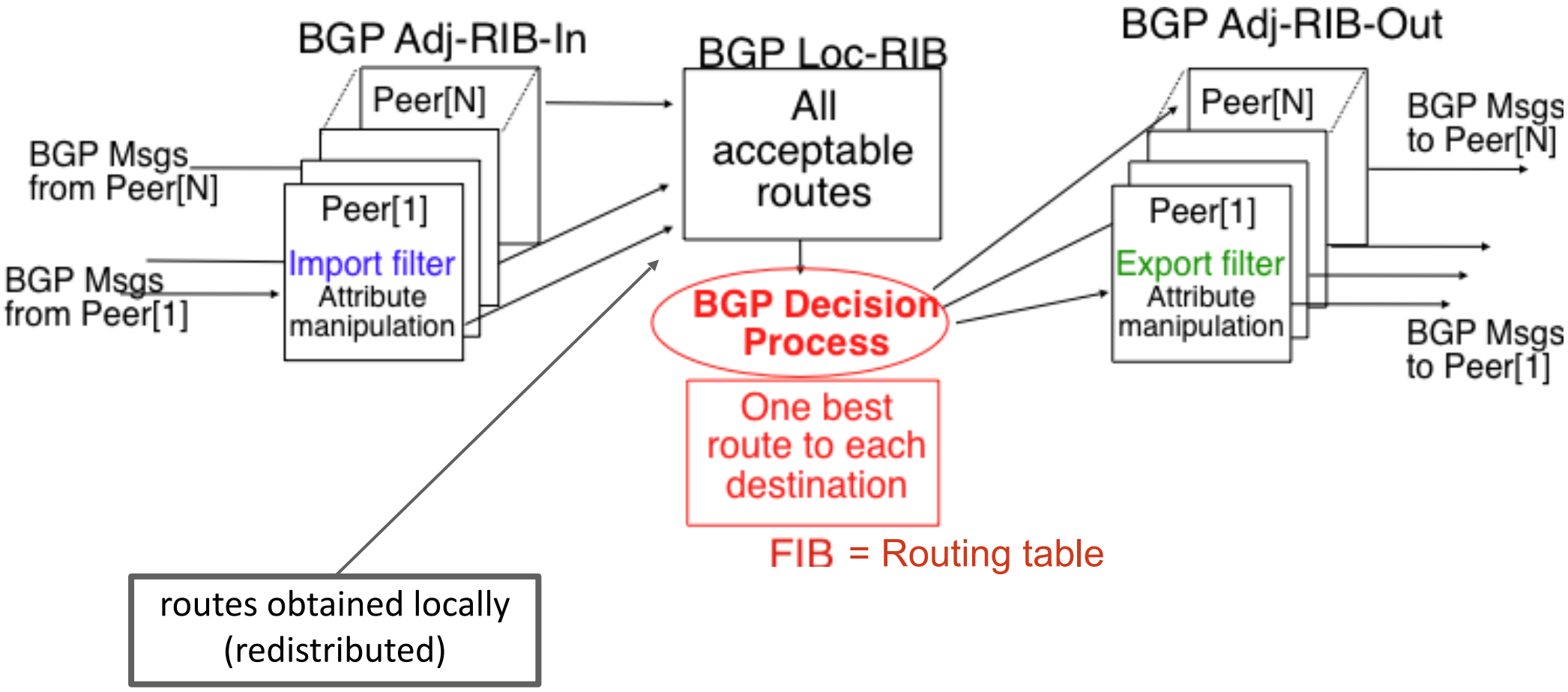
Part B.

1. How does BGP work?

- [Recall:] 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
  - *exports* the selected routes to BGP neighbors,
    after applying *export* policy rules and possibly aggregation

- Routes are advertized via UPDATE messages that contain *only modifications*: ***additions/withdrawals***
- Other BGP messages are: OPEN, NOTIFICATION (= RESET), KEEPALIVE
- [Also recall:] routers advertize to an i-BGP neighbor *only routes learnt from e-BGP*

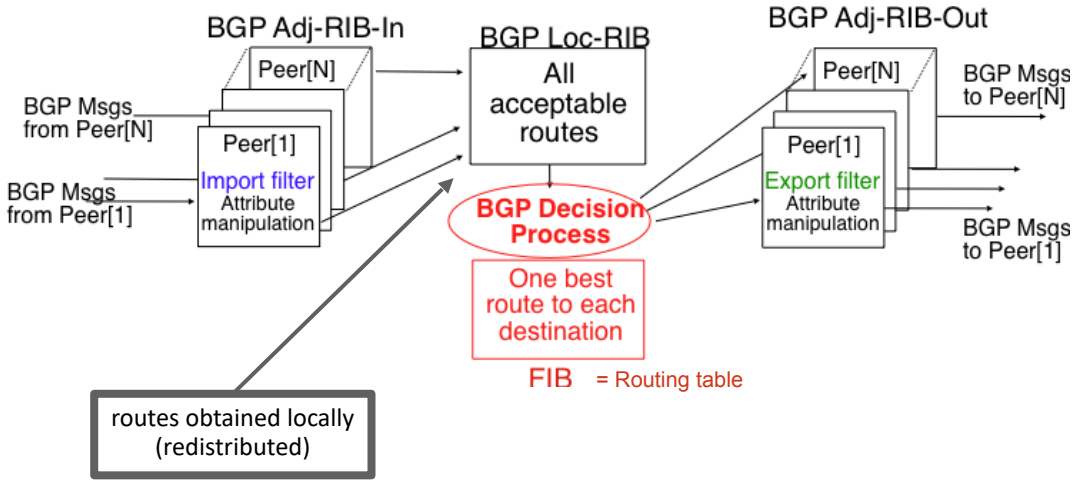
Model of a BGP Router



Routes, RIBs, Routing Table

An *advertized route* contains:

- destination (subnetwork 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

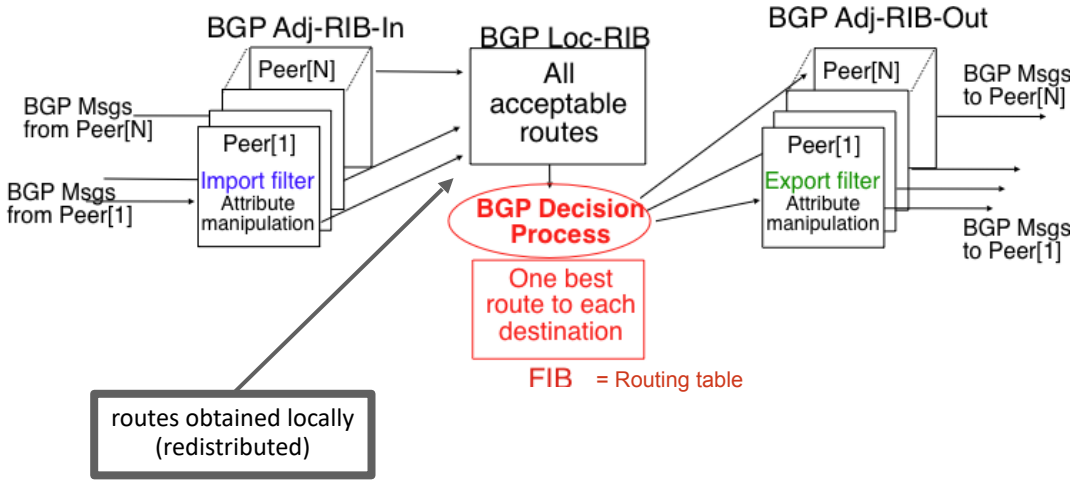
Used for packet forwarding, in real time

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
- Routes' attributes are compared against each other using a sequence of criteria, until only one route remains.
  A common sequence is:

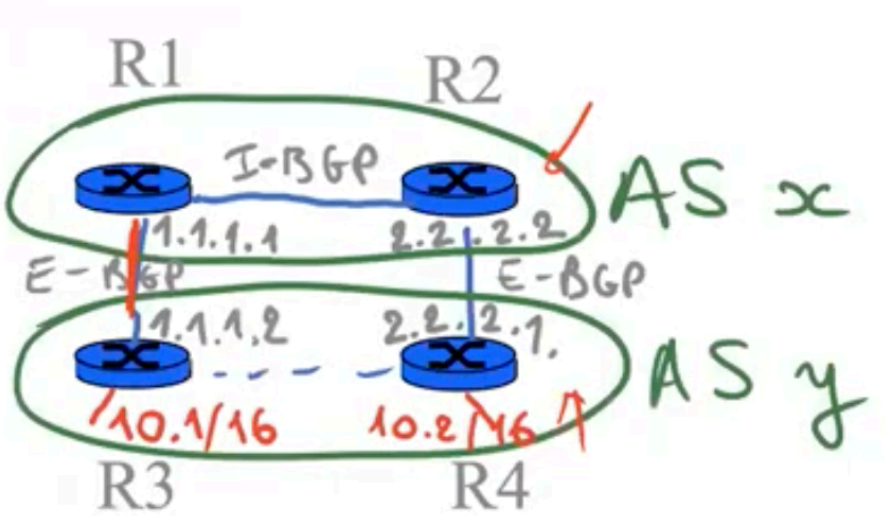
0. Highest weight (Cisco proprietary)
1. Highest LOCAL-PREF
2. Shortest AS-PATH
3. Lowest MED, if taken seriously by this network
4. e-BGP > i-BGP (= if route is learnt from e-BGP, it has priority)
5. Shortest path to NEXT-HOP, according to IGP
6. Lowest BGP identifier (router-id of the BGP peer from 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 Adj-RIB-out (*one route per destination for each BGP peer*) and 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

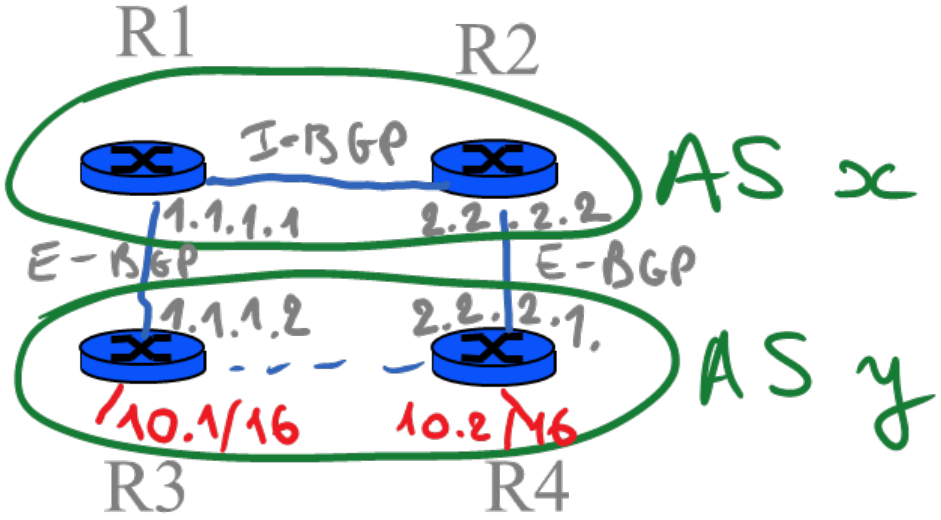
10.2/16 AS=y

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

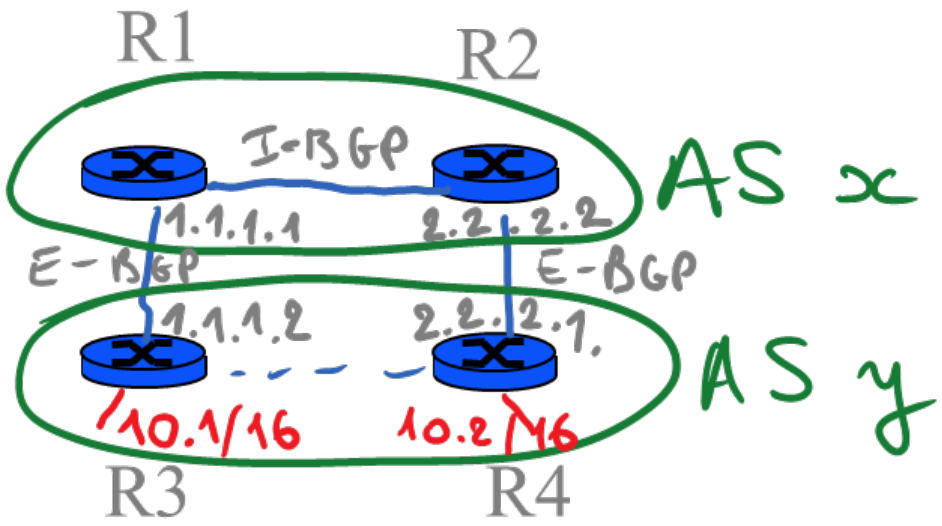
R2 —> R1

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
5. Shortest path to NEXT-HOP, according to IGP
6. Lowest BGP identifier (router-id of the BGP peer from whor

(The Cisco and FRR implementation of BGP, used in lab 6, have addition

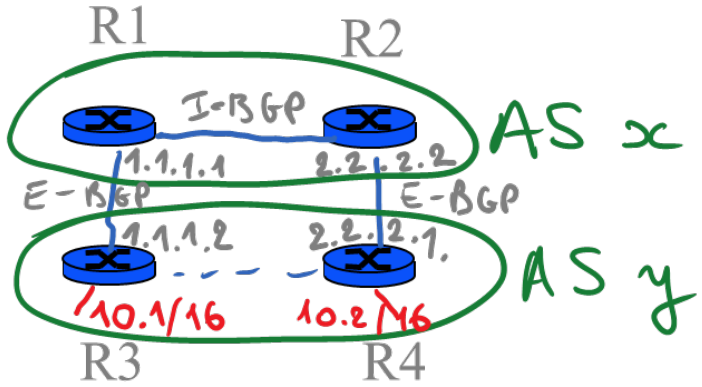
Step 2

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.

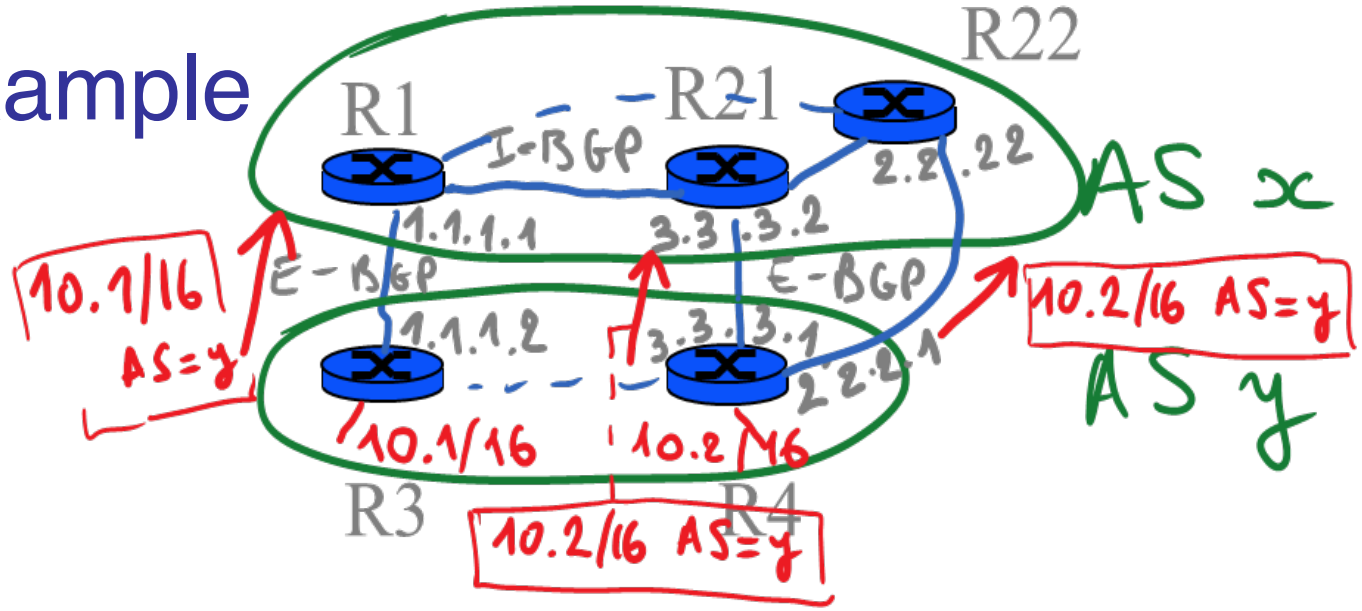
The first applicable rule in the decision process (slide “The Decision Process”) says that if a route is learnt from e-BGP it has precedence over a route learnt from i-BGP (e-BGP > i-BGP).

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:

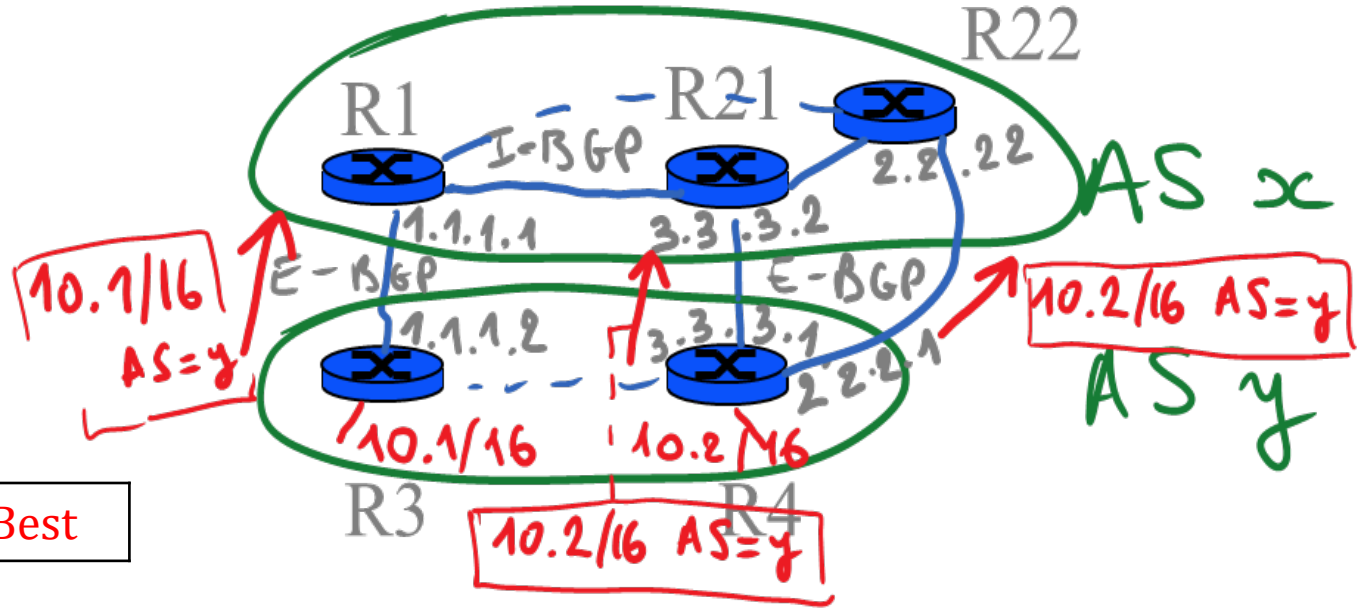
Note:

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

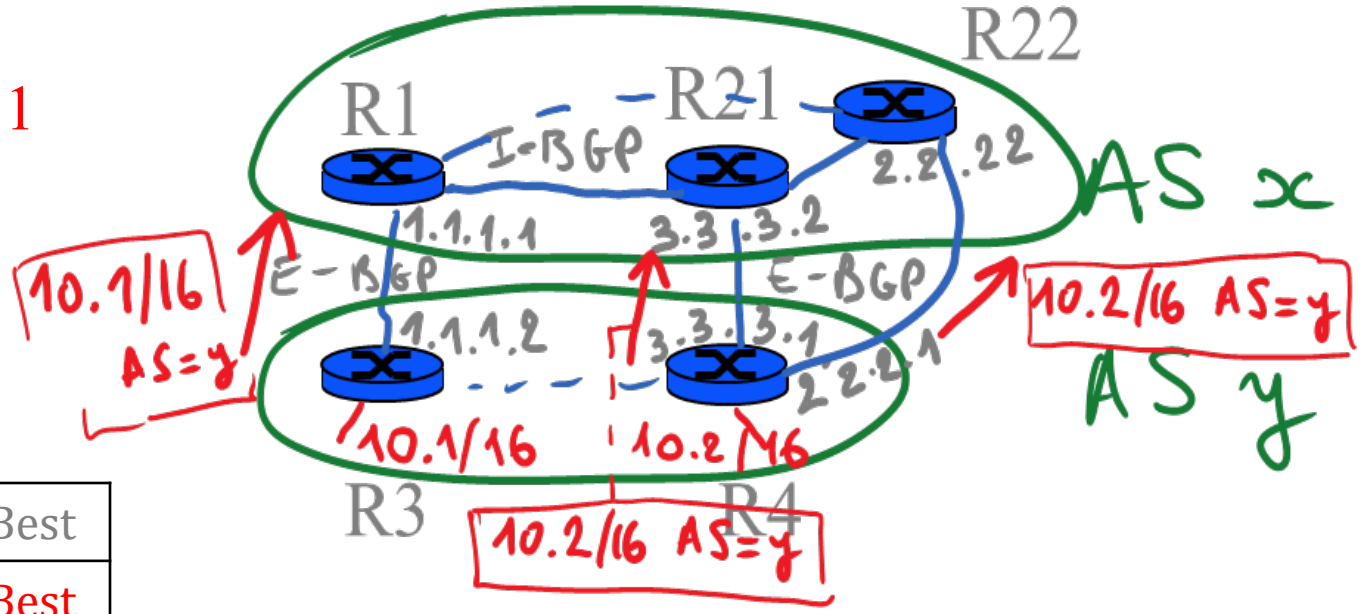
To R21	10.1/16 AS =y NEXT-HOP=1.1.1.2
To R22	10.1/16 AS =y NEXT-HOP=1.1.1.2

- 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

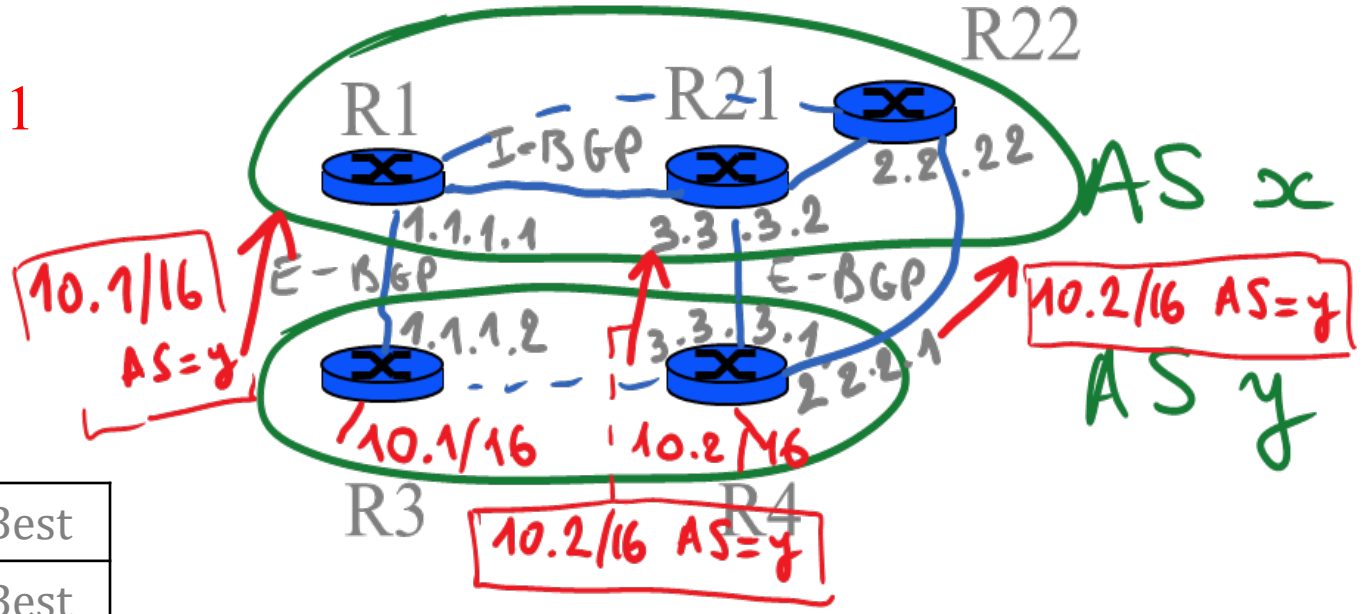
To R21	10.1/16 AS =y NEXT-HOP=1.1.1.2
To R22	10.1/16 AS =y NEXT-HOP=1.1.1.2

- 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	

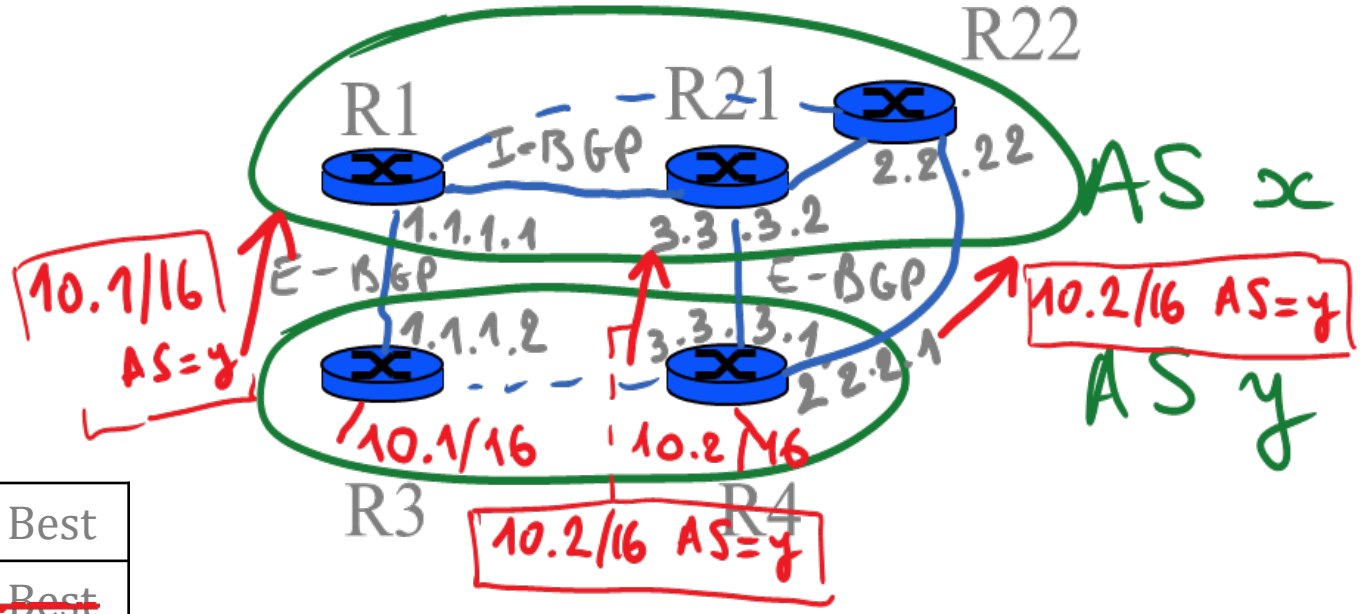
Will the decision process promote the new route to “best route” assuming that WEIGHT, LOCAL_PREF, MED are empty?

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
5. Shortest path to NEXT-HOP, according to IGP
6. Lowest BGP identifier (router-id of the BGP peer from whom ro

(The Cisco and FRR implementation of BGP, used in lab 6, have additional ca

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.

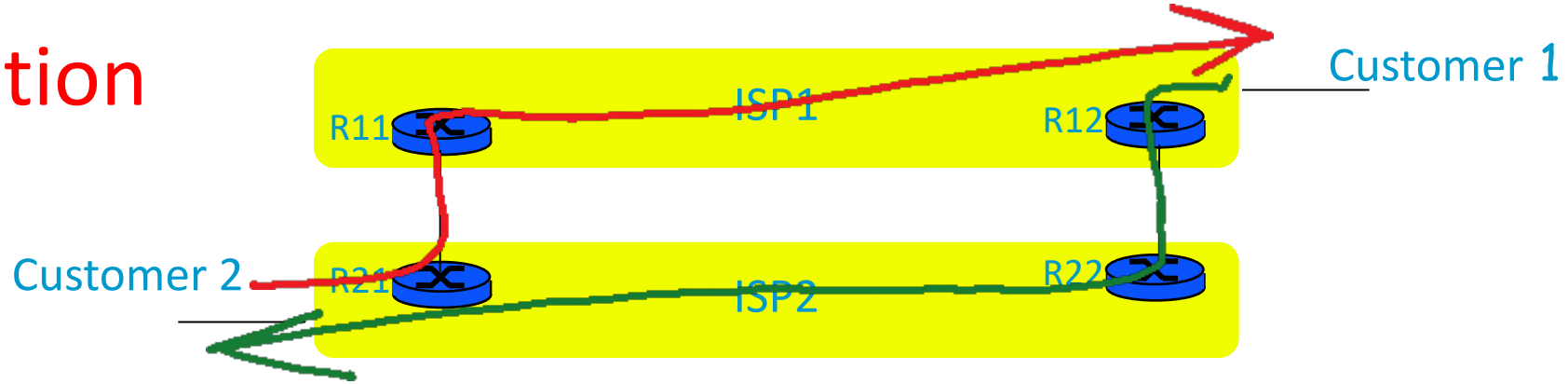
Thus the route that has NEXT-HOP=3.3.3.1 is preferred by the decision process, i.e. the new route is designated as “best”.

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

Solution



Answer C: It depends on the configuration.

If configuration in both ISPs 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* !