

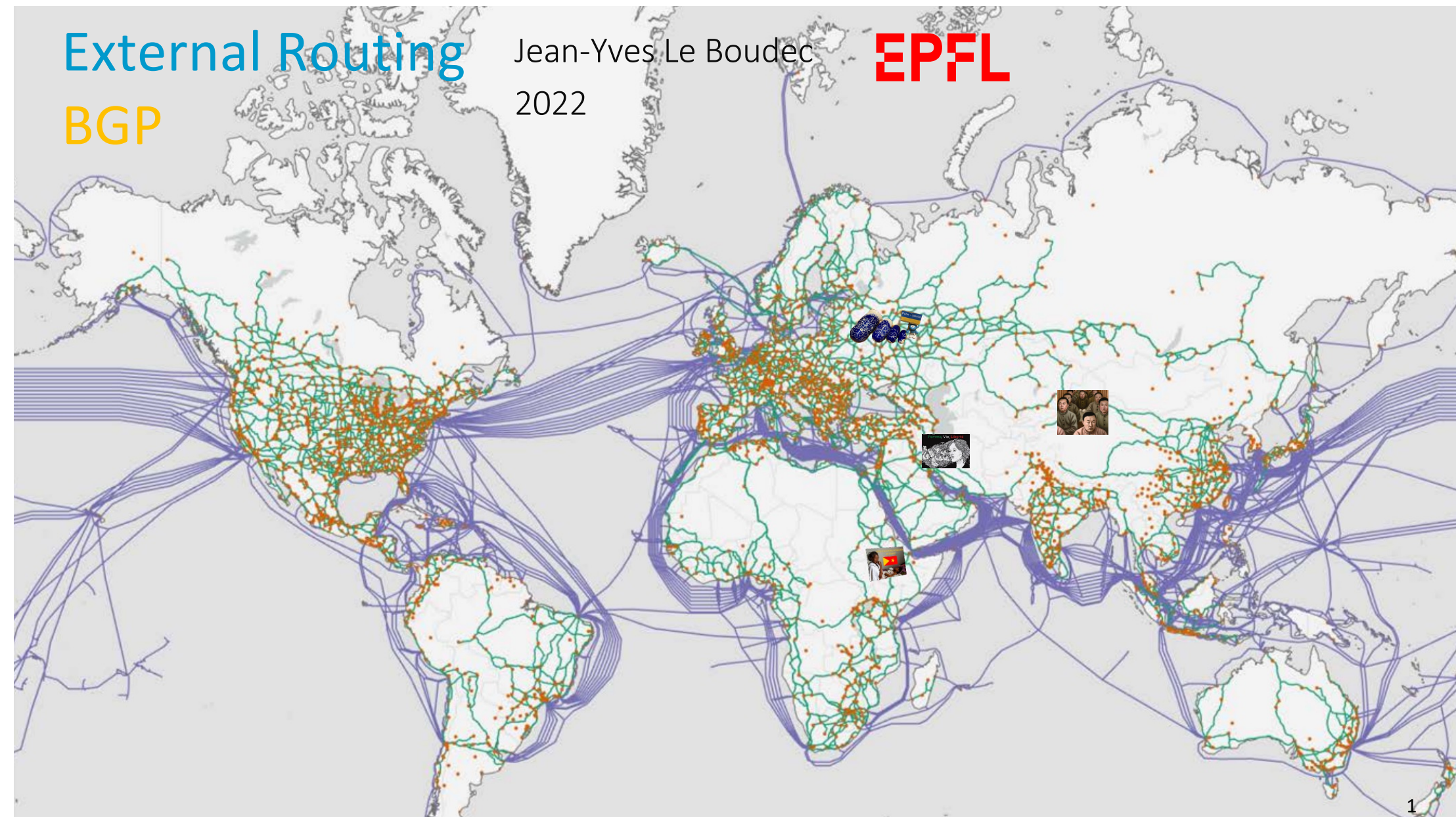
External Routing

BGP

Jean-Yves Le Boudec

2022

EPFL



Contents

- A. What Inter-Domain Routing does
 - 1. Inter-Domain Routing
 - 2. Policy Routing
- B. How BGP works
 - 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

Cover 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).

A. 1. Inter-Domain Routing

Why invented ?

The Internet is too large + heterogeneous to be run by one routing protocol

Hierarchical routing is used

the Internet is split into Domains, or Autonomous Systems
with OSPF: large domains are split into Areas

Routing protocols are said

interior: (Internal Gateway Protocols, IGP): inside ASs:
RIP, OSPF (standard), IGRP (Cisco)

exterior: between ASs: BGP

What is an ARD ? An AS ?

ARD = Autonomous Routing Domain
= routing domain under one single administration

AS = Autonomous System = ARD with a number (“AS number”)

AS number is 32 bits denoted with dotted 16 bit integer notation e.g.
23.3456

0.559 means the same as 559

Private AS numbers: 0.64512 – 0.65535

ARDs that do not need a number are typically served by one single ISP

Examples: AS1942 - CICG-GRENOBLE, AS2200 - Renater

AS559 - SWITCH Teleinformatics Services

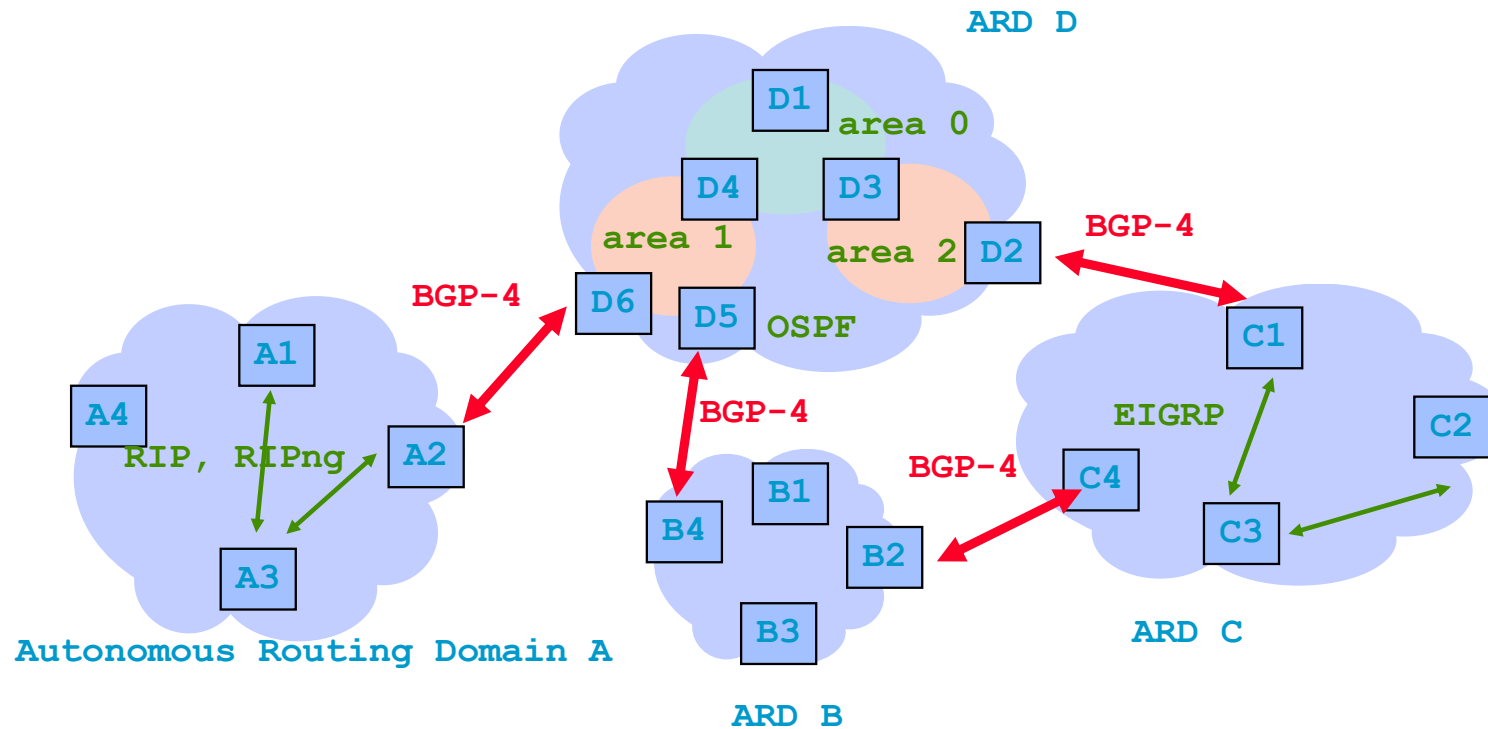
AS5511 – OPENTRANSIT

EPFL: one ARD, no number; all external traffic goes via Switch

BGP and IGP

ARDs can be transit (B and D), stub (A) or multihomed (C). Only non stub domains need an AS number.

An IGP is used inside a domain, BGP is used between domains



What does BGP do ?

What does BGP do ?

BGP is a routing protocol between ARDs. It is used to compute paths from one router in one ARD to any network prefix anywhere in the world

BGP can handle both IPv4 and IPv6 addresses in a single process

The method of routing is

Path vector

With policy

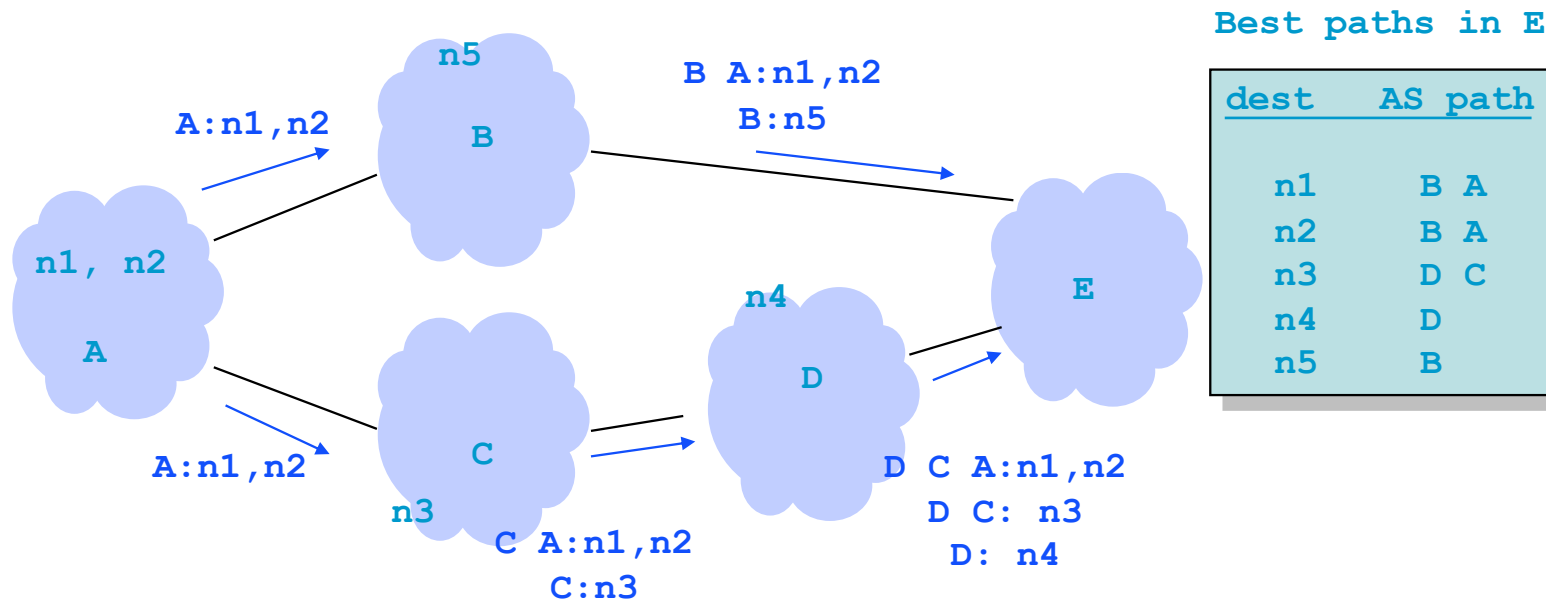
Path Vector Routing

What? Find best routes, in a sense that can be decided by every ARD using their own criteria

How? a route between neighbours is (path: dest) where path is a sequence of AS numbers and dest is an IP prefix. Example: B A:n1

Every AS appends its number to the path it exports

Every AS uses its own rules for deciding which path is better



Border Gateways, E- and I-BGP

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

At the **boundary** between 2 ARDs there are 2 BGP speakers, one in each domain

Q: compare to OSPF

Inside one ARD there are usually several BGP speakers

They all talk to each other, in order to exchange what they have learnt

Using “**Internal BGP**” (I-BGP)

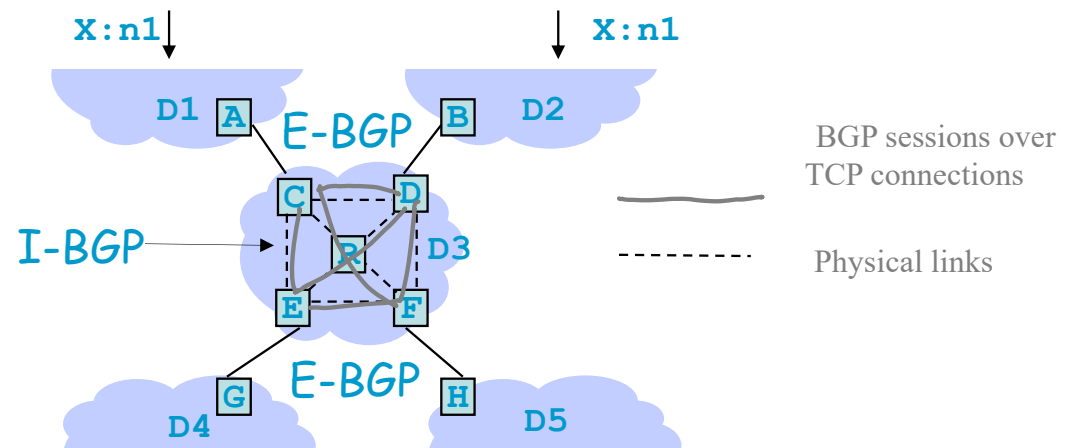
Over TCP connections, full mesh called the “**BGP mesh**”

I-BGP is the same as E-BGP except:

routes learned from I-BGP are not repeated to I-BGP

router does not prepend own AS number over I-BGP

NEXT-HOP is not modified (see later)

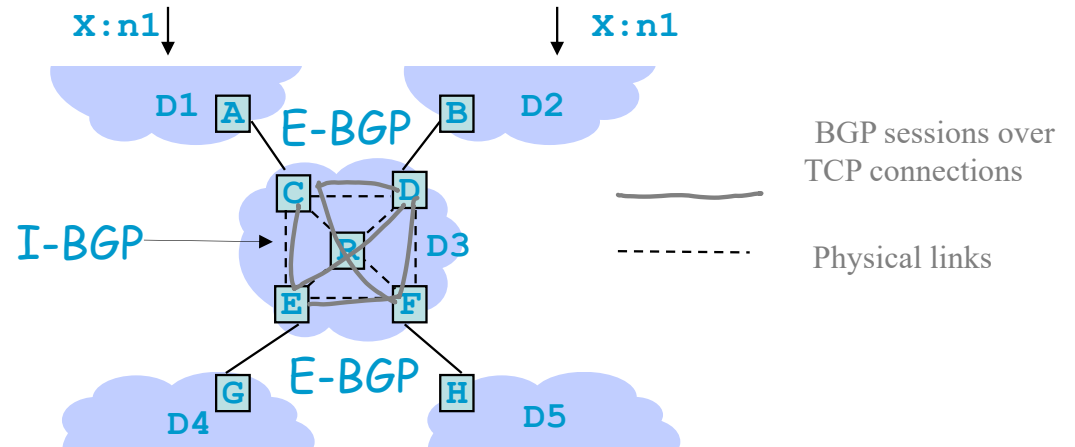


Say what is always true

- 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

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. $C \rightarrow E : D2 - X : n1$

2. Policy Routing

Why invented ?

Interconnection of ASs (= peering) is self-organized

point to point links between networks: ex: 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...

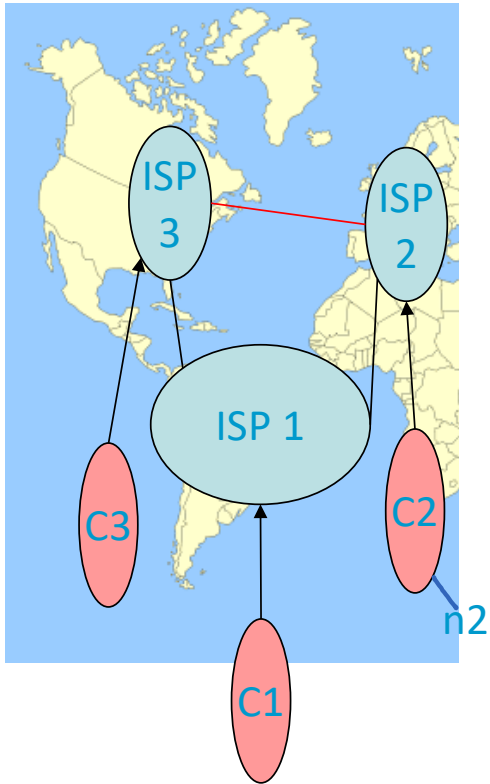
Mainly 2 types of relations

Customer-provider: EPFL is customer of Switch. EPFL pays Switch

Shared Cost peer: Swisscom and Switch are peers. They collaborate to serve their customers

Plus many others, depending on (private) business agreements

What is the Goal of Policy Routing ?



Example:

ISP3-ISP2 is transatlantic link, cost shared between ISP2 and ISP 3

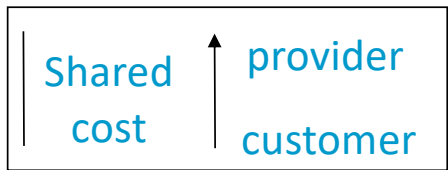
ISP3- ISP1 is a local, inexpensive link

C_i is customer of ISP_i , ISPs are peers

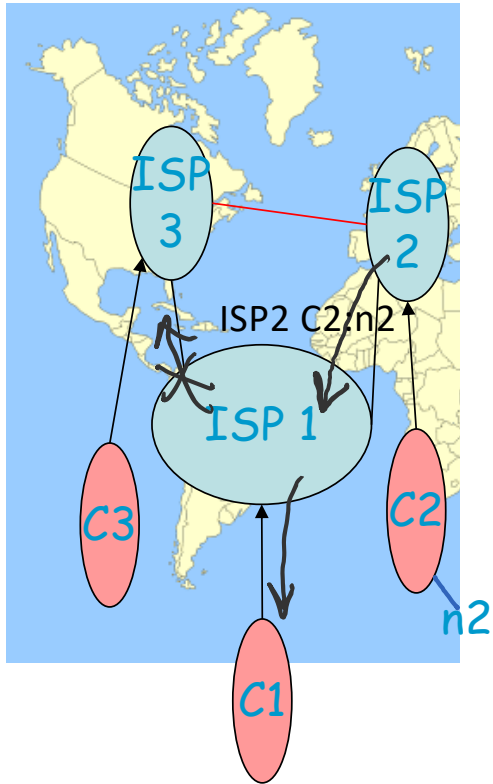
It is advantageous for ISP3 to send traffic to n2 via ISP1. But...ISP1 does not agree to carry traffic from C3 to C2

ISP1 offers a “transit service” to C1 and a “non-transit” service to ISP2 and ISP3

The goal of “policy routing” is to support this and other similar requirements



A Common Rule



Routes coming from peers and providers are not propagated to peers nor providers.

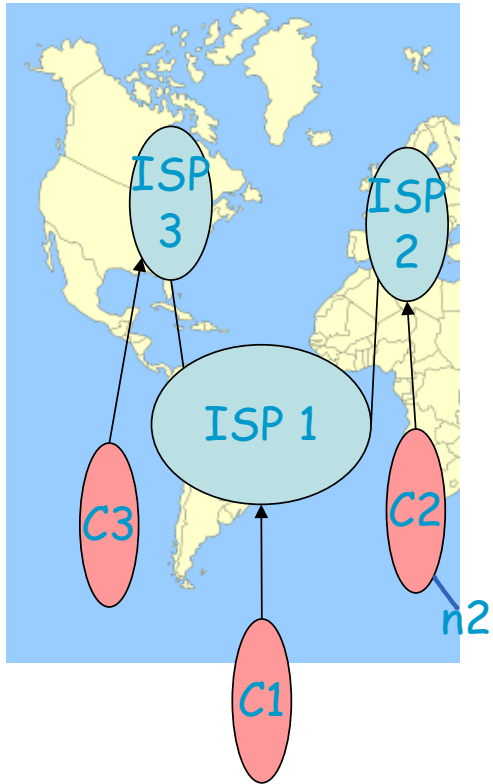
Example

ISP1 announces 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

Policy is implemented using import and export rules (by using **route-map**), and the decision process



ISP1-ISP2 and ISP1-ISP3 are peers; ISP2-ISP3 are *not* peers nor customers/providers

All apply the rule “Routes coming from peers and providers are not propagated to peers nor providers”.

What is a valid path from C2 to C3 ?

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

B. BGP (Border Gateway Protocol)

1. How it works, Fundamental Examples

BGP-4, RFC 4271

BGP routers talk to each other over TCP connections

BGP messages: OPEN, NOTIFICATION (= RESET), KEEPALIVE
UPDATE

UPDATE messages contains modifications

- Additions and withdrawals

- A BGP router **transmits only modifications**

A BGP Router ...

Receives and stores candidate routes from its BGP peers and from itself

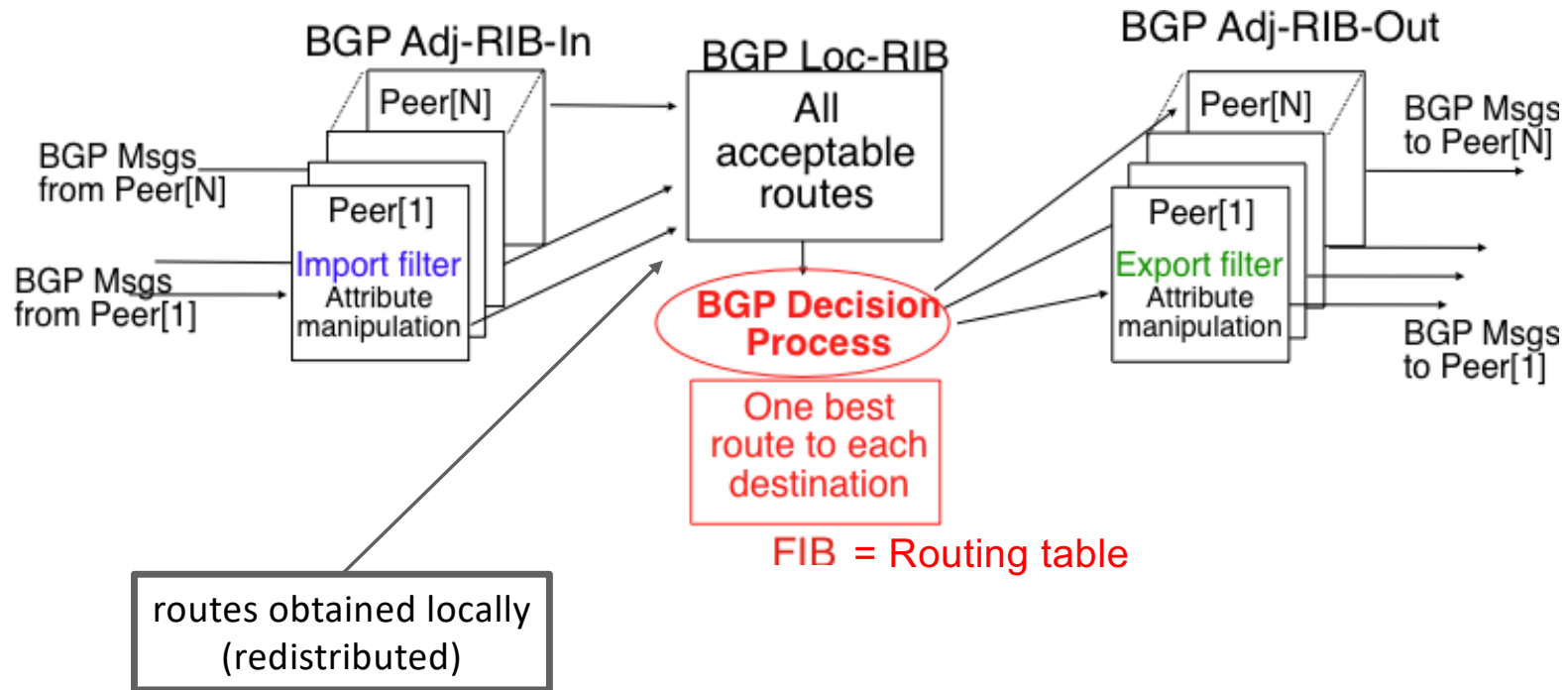
Applies the decision process to **select at most one route** per destination prefix

Exports the selected routes to BGP neighbours, after applying export policy rules and possibly aggregation.

Stores result in Adj-RIB-out (one per BGP peer) and sends updates when Adj-RIB-out changes (addition or deletion).

Only routes learnt from E-BGP are sent to an I-BGP neighbor.

Model of a BGP Router



Routes, RIBs, Routing Table

The records sent in BGP messages are called “**Routes**”. Routes + their attributes are stored in the Adj-RIB-in, Loc-RIB, Adj-RIB-out.

A route is made of:

- destination (subnetwork prefix)

- path to the destination (AS-PATH or BGPsec_Path (authenticated))

- NEXT-HOP (set by E-BGP, left unchanged by I-BGP)

- Origin: route learnt from IGP, BGP, static

- Other attributes

 - LOCAL-PREF (see later), ATOMIC-AGGREGATE (= route cannot be dis-aggregated), MED (see later) etc

In addition, 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** decides which route is selected;

At most one best route to exactly the same prefix is chosen

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

A route can be selected only if its next-hop is reachable

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

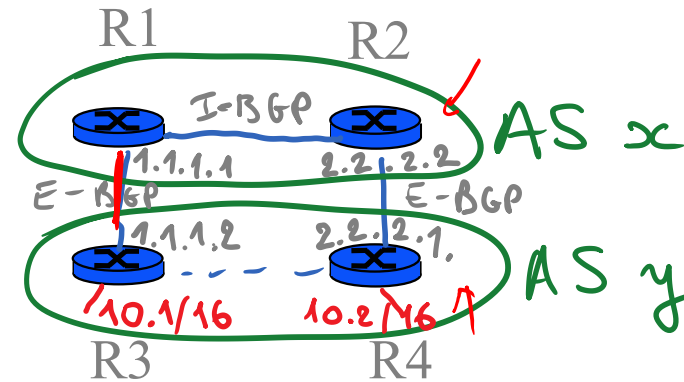
0. Highest weight (Cisco proprietary)
 1. Highest LOCAL-PREF
 2. Shortest AS-PATH
 3. Lowest MED, if taken seriously by this network
 4. E-BGP > I-BGP
 5. Shortest path to NEXT-HOP, according to IGP
 6. Lowest BGP identifier (router-id of the BGP peer from whom route is received)
- (The Cisco and FRR implementation of BGP, used in lab, have additional cases, not shown here)

Fundamental Example

In this simple example there are 4 BGP routers. They communicate directly or indirectly via E-BGP or I-BGP, as shown on the figure.

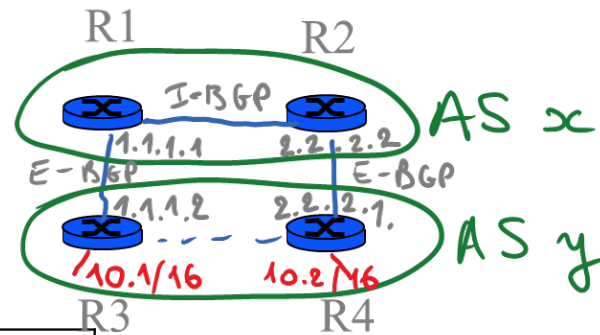
There are 2 ASs, x and y. We do not show the details of the internals of y. R3 and R4 send the BGP messages shown.

We show next only a subset of the route attributes (such as : destination, path, NEXT-HOP)



We focus on R1 and show its BGP information:

Step 1 $R3 \rightarrow R1$
 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

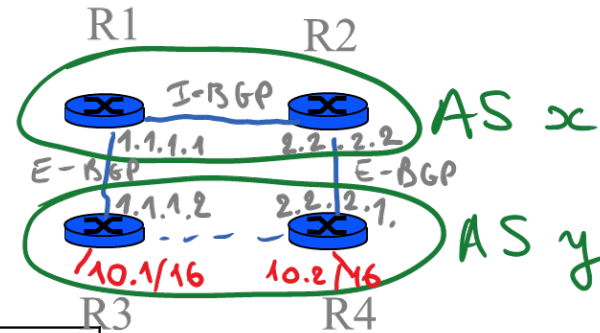
(import filters) R1 accepts the updates and stores them in Adj-RIB-In
 (Decision Process) R1 designates these routes as best routes
 (export filters) R1 puts updates into Adj-RIB-Out, which will cause them to be sent to BGP neighbours

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” ?

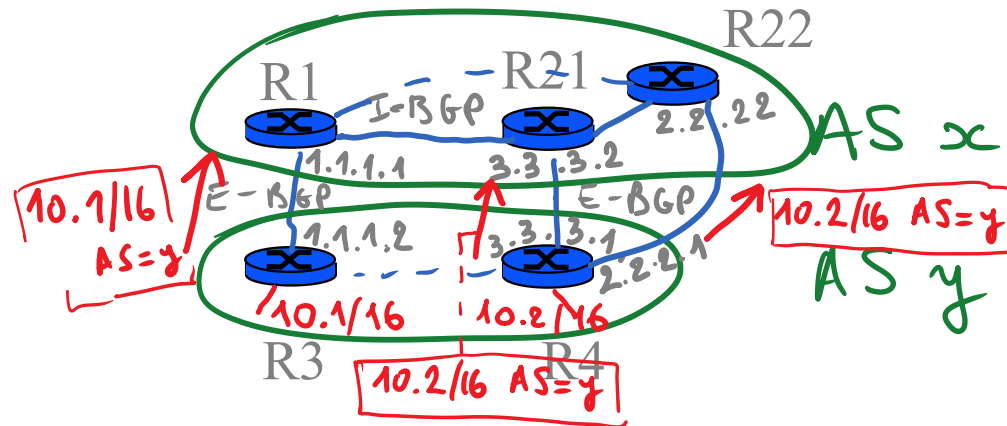
- A. The first one only
- B. The second one only
- C. Both
- D. None
- E. I don't know

Another Fundamental Example

There are now 3 BGP routers in AS x. Note that the 3 BGP in AS x routers must have TCP connections with each other (same in AS y, but not shown on figure).

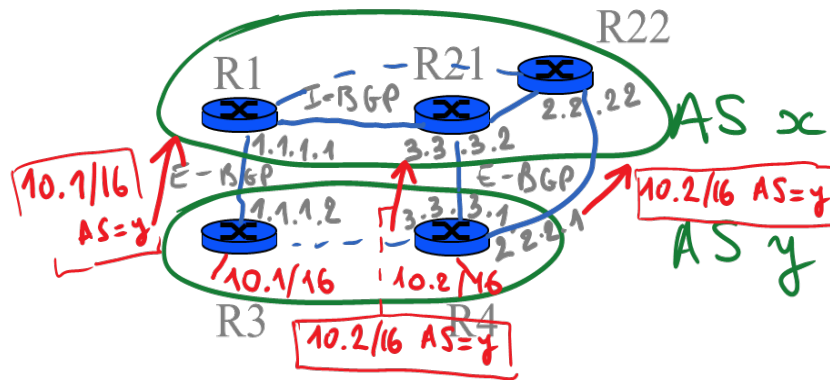
An IGP (for example OSPF) also runs on R1, R21 and R22. All link costs are equal to 1.

The announcements made by R3 and R4 are different, as shown on the figure.



We focus on R1 and show its BGP information:

Step 1 $R3 \rightarrow R1$
 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

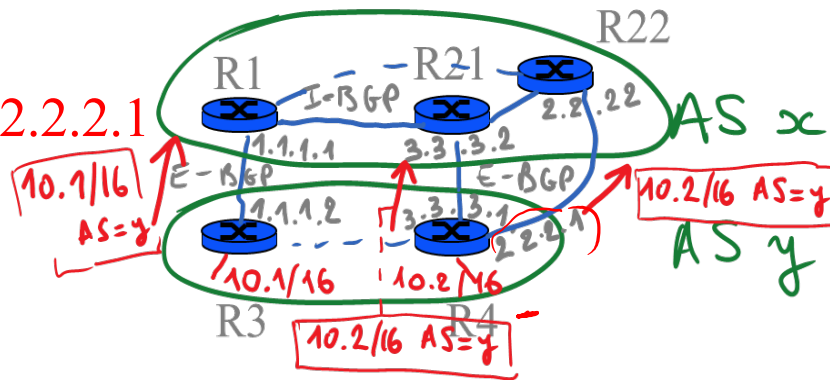
To R21	10.1/16 AS=y NEXT-HOP=1.1.1.2
To R22	10.1/16 AS=y NEXT-HOP=1.1.1.2

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 neighbours R21 and R22

Step 2 $R22 \rightarrow R1$
 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

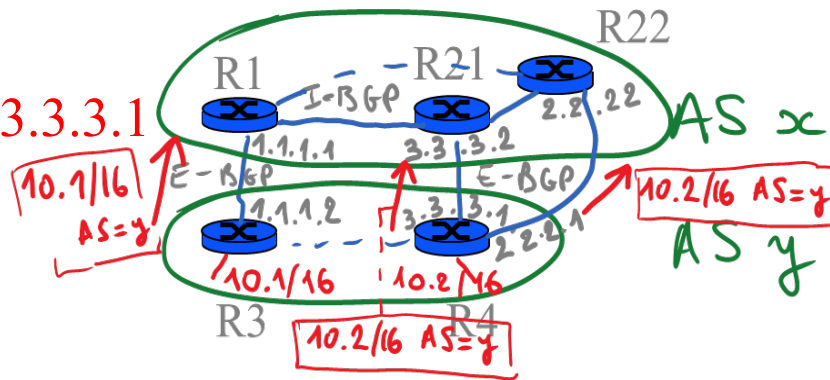
R1 designates this route as best route

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 $R21 \rightarrow R1$

10.2/16 AS = y NEXT-HOP=3.3.3.1



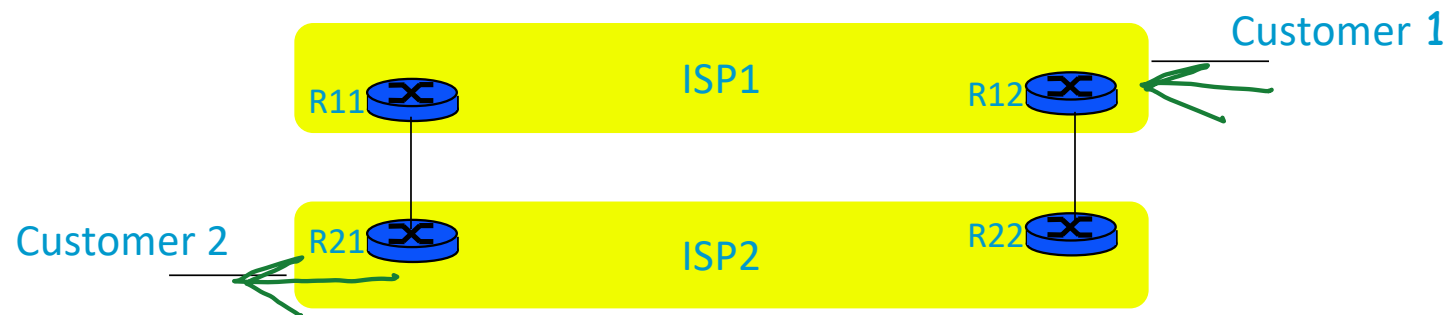
Adj-RIB-in

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” ?

- A. Yes
- B. No
- C. I don't know

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

How are routes originated (= sourced) ?

BGP propagates route information, but how is this bootstrapped ?
Some BGP routers must *originate the routes* that are in their domains.

Several methods for the source of a route:

Static configuration: tell this BGP router which are the prefixes to originated (“network” command in FRR)

Redistribute connected: tell this BGP router to originate all prefixes that are on-link with this router

(assumes that all routers in network run BGP)

Redistribute from IGP:

= tell this router to originate all prefixes that IGP has learnt

Example: redistribute OSPF into BGP

With OSPF, in principle, only internal prefixes should be redistributed

In BGP such routes have attribute ORIGIN=IGP.

When originated, the BGP NEXT-HOP of a route is its IGP next-hop.

2. Aggregation

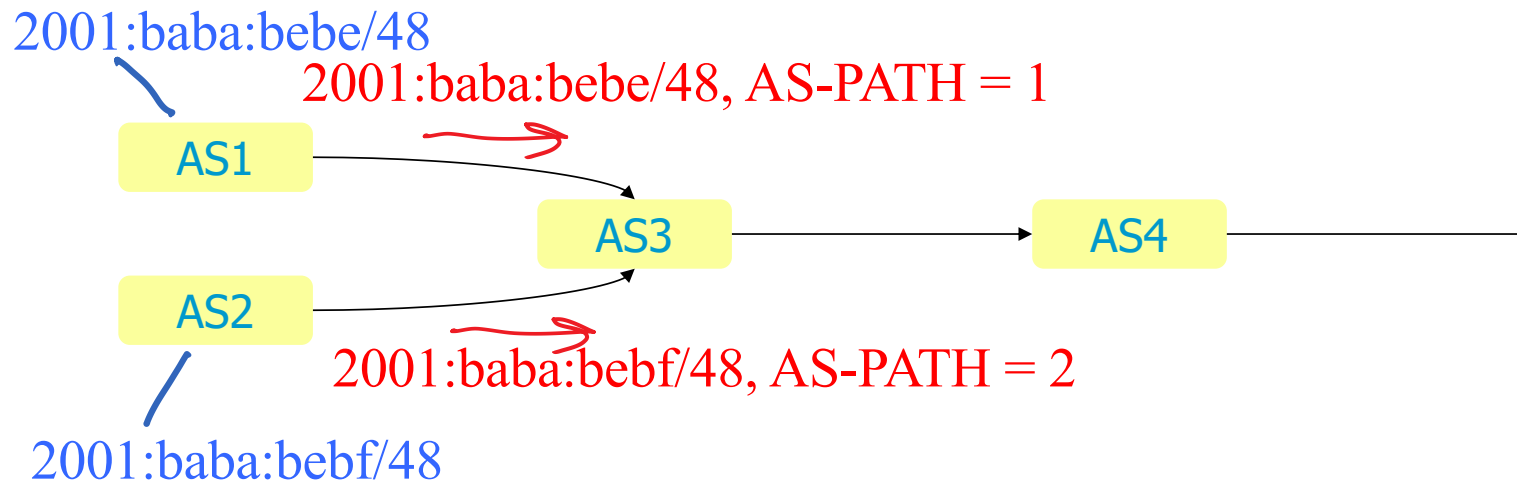
Domains that do not have a default route (i.e. all transit ISPs) must know all routes in the world (several hundreds of thousands of prefixes)

- in IP routing tables unless default routes are used
- in BGP announcements

Aggregation is a way to reduce the number of routes

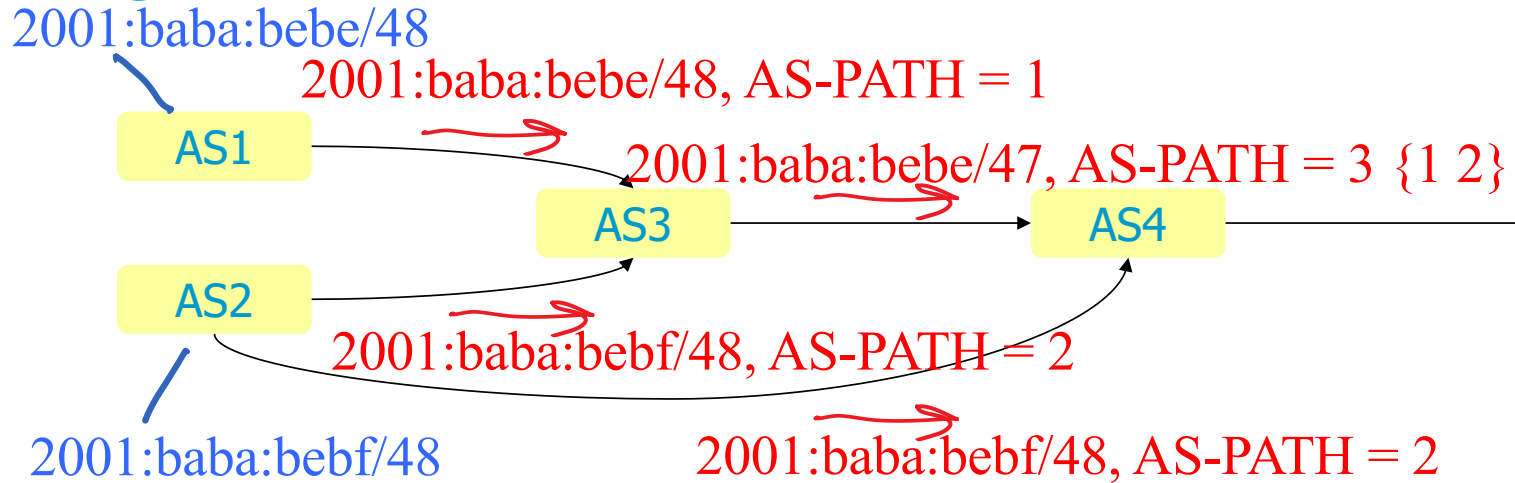
Aggregation is expected to be very frequent with IPv6, less with IPv4

Can AS3 aggregate these routes into a single one ?



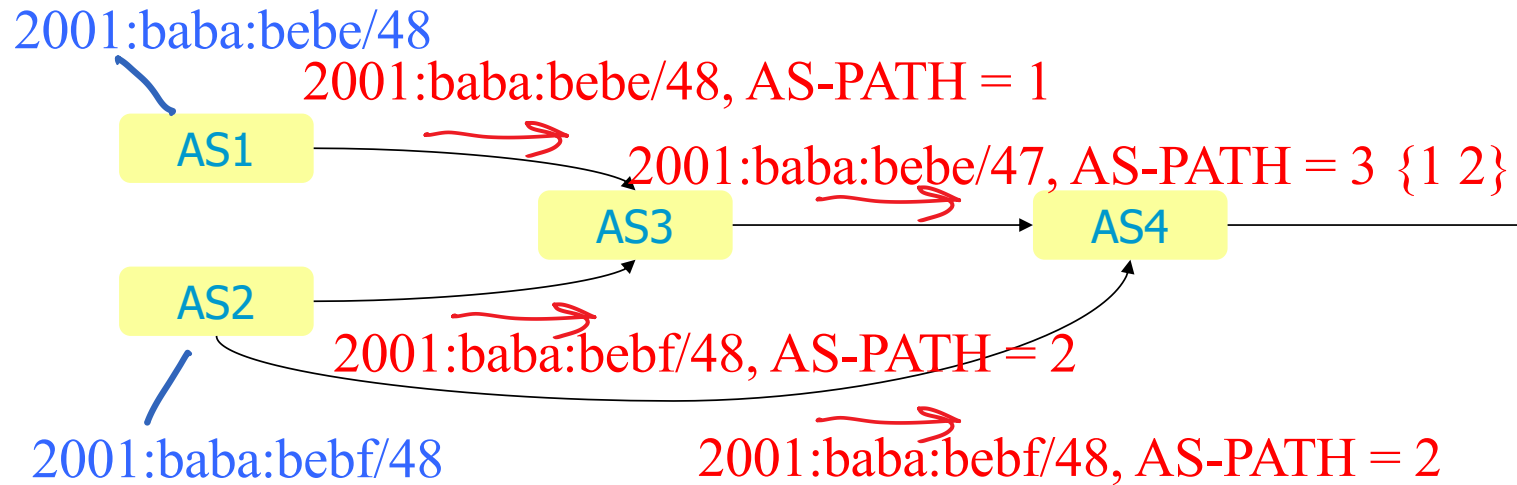
- A. Yes and the aggregated prefix is 2001:baba:bebe/47
- B. Yes and the aggregated prefix is 2001:baba:bebf/48
- C. Yes but the aggregated prefix is none of the above
- D. No
- E. I don't know

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



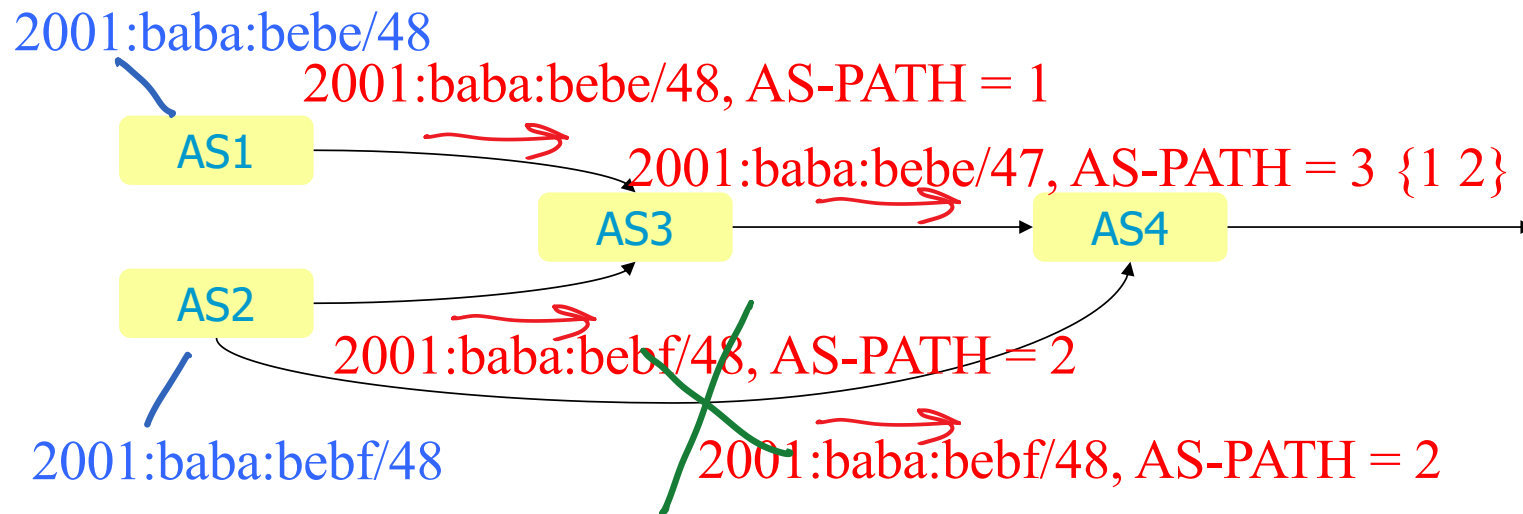
- A. The top route
- B. The bottom route
- C. Both
- D. I don't know

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



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

Assume the link AS2-AS4 breaks ...



At AS4: keepalive detects loss of AS2

Adj-RIB-In routes are declared invalid

Decision process recomputes best route to 2001:baba:bebf/48

There is none

The routing table entry 2001:baba:bebf/48 is removed

A packet to 2001:baba:bebf/48 matches the route 2001:baba:bebe/47 and goes via AS3

3. Forwarding Entries learnt by BGP are written into Routing Tables

So far, we have seen how BGP routers learn about all prefixes in the world. It remains to see how they write the corresponding entries in the forwarding tables (i.e. routing tables). There are two possible ways for this:

Redistribution of BGP into IGP : routes learnt by BGP are passed to IGP (ex: OSPF)

- Only routes learnt by E-BGP are redistributed into IGP (unless `bgp redistribute-internal` is used)

- IGP propagates the routes to all routers in domain

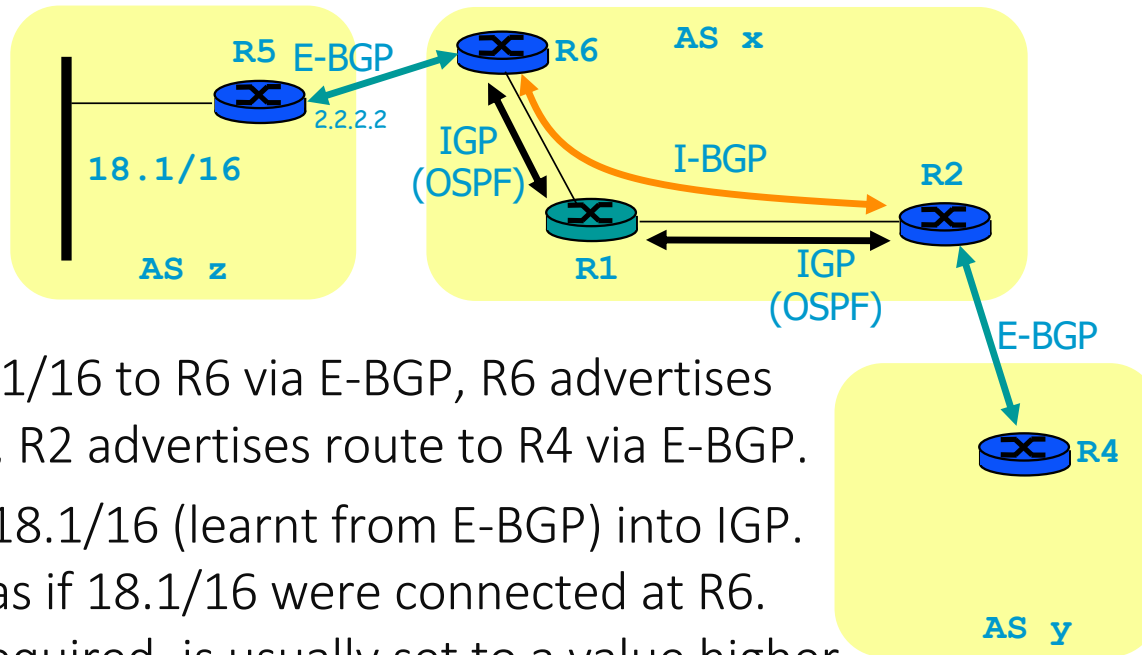
- Works with OSPF, might not work with other IGPs (table too large for IGP)

Injection of BGP routes into forwarding table of this router

- Routes do not propagate; this helps only this router

- With Cisco routers and in FRR (in the lab): this is always done.

Example



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 connected at R6. The IGP cost, if required, is usually set to a value higher than all IGP distances.

IGP propagates 18.1/16 (for OSPF: as a type 5 LSA); 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

Avoiding Re-Distribution of BGP into IGP

Many operators avoid re-distribution of BGP into IGP

- Large number of routing entries in IGP

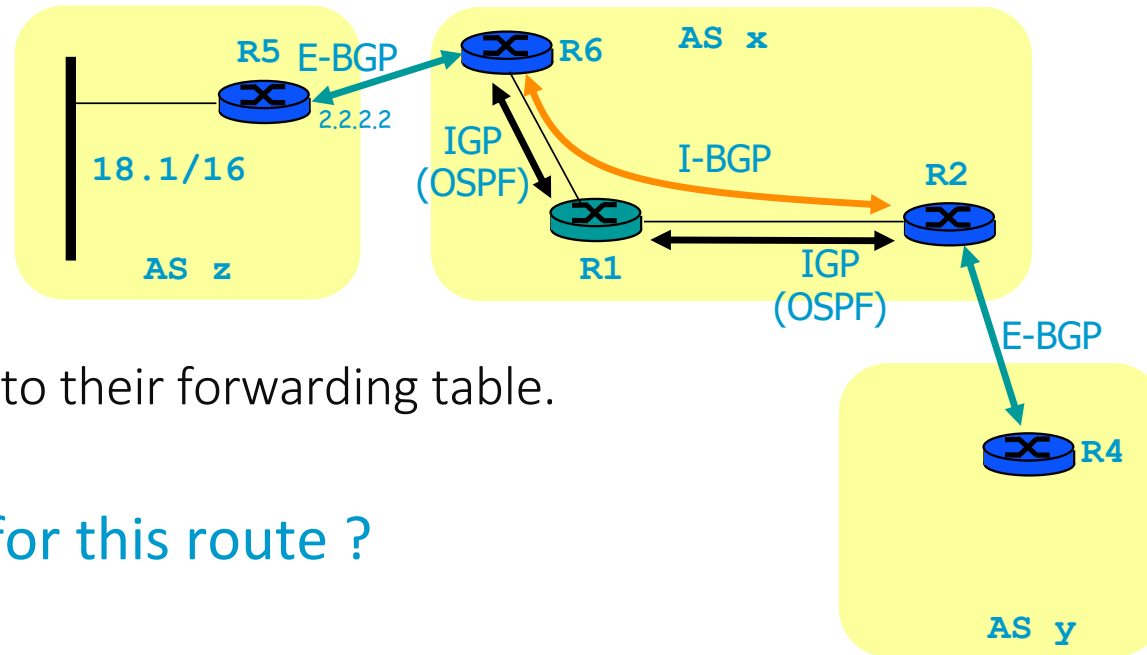
- Convergence time after failures is large if IGP has many routing table entries

- OSPF is able to handle large routing tables, other IGPs may not (e.g. distance vector routing protocols)

If redistribution is avoided, only *injection* is used, i.e. BGP routes are written directly into the forwarding table.

Example

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



What is the next-hop for this route ?

- 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

Recursive Table Lookup

When an IP packet is submitted to router, the forwarding table may indicate a “next-hop” which is not on-link with this router

A second table lookup needs to be done to resolve the next-hop into an on-link neighbour

in practice, second lookup may be done in advance – not in real time– by pre-processing the routing table

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

Example of Recursive Table Lookup

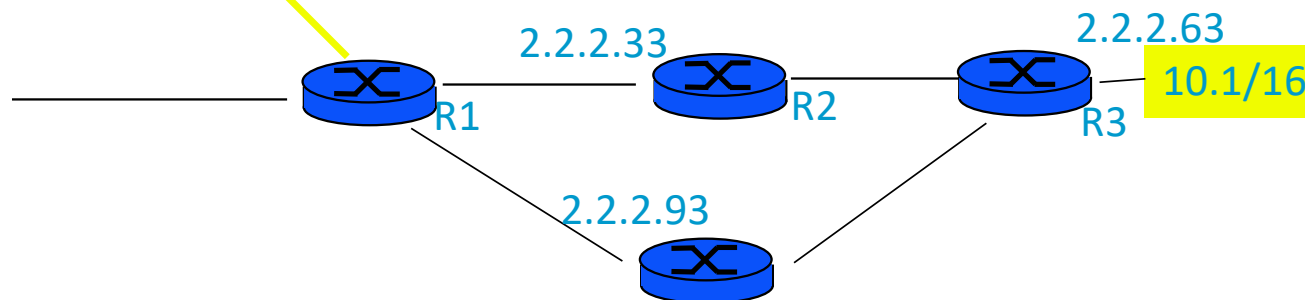
At R1, data packet to 10.1.x.y is received

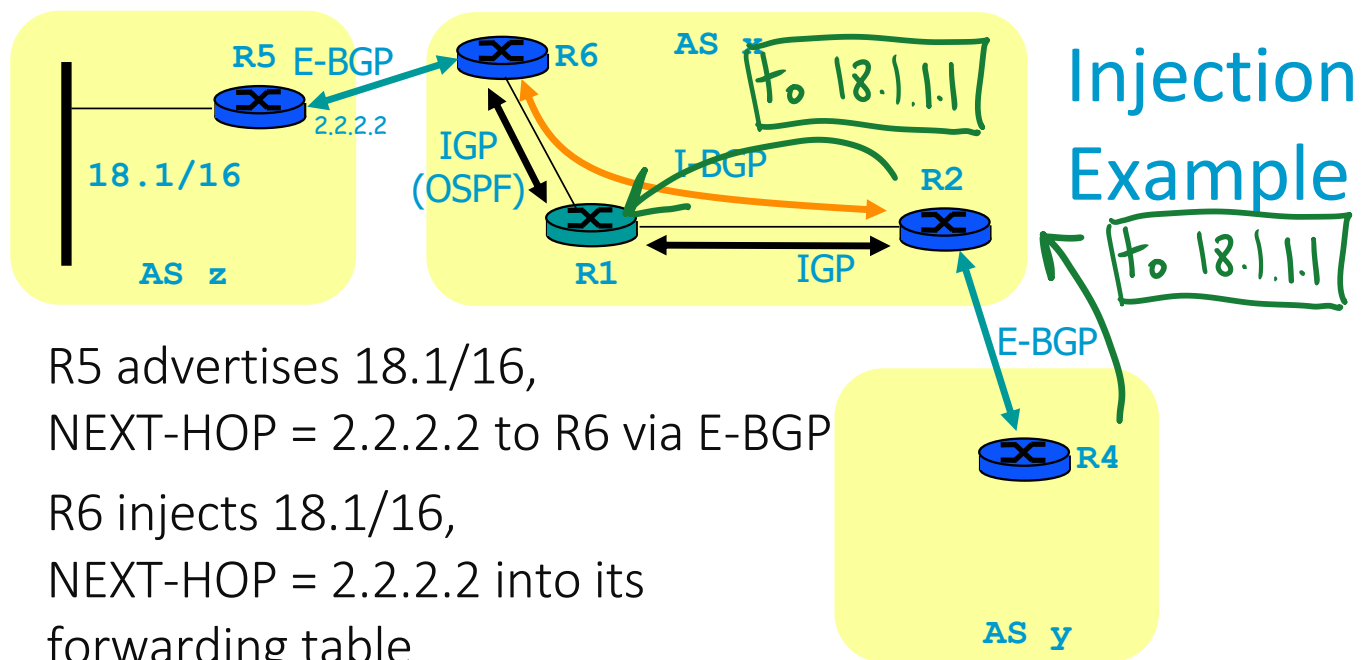
The forwarding table at R1 is looked up first, next-hop 2.2.2.63 is found;

A second lookup for 2.2.2.63 is done; the packet is sent to 2.2.2.33 over eth0

Forwarding Table at R1

<i>To</i>	<i>next hop</i>	<i>interface</i>
10.1/16	2.2.2.63	N/A
2.2.2.63	2.2.2.33	eth0





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

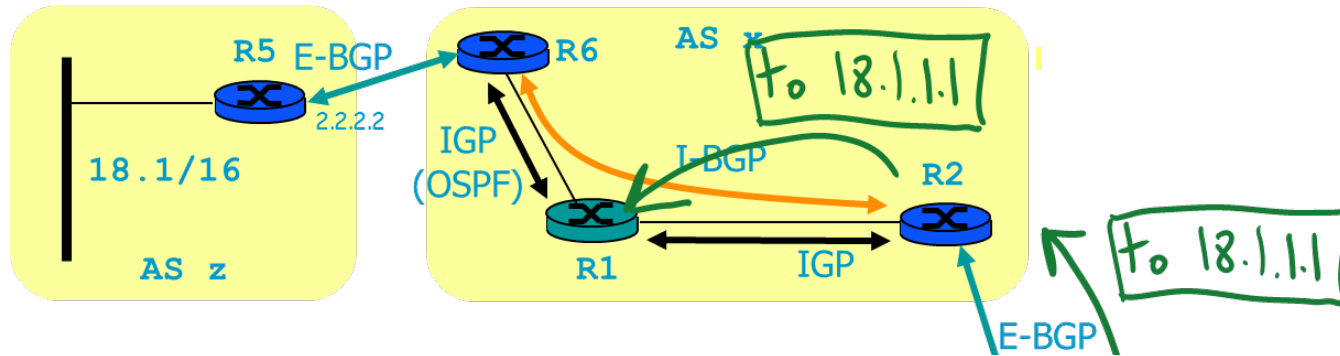
R2 injects 18.1/16, NEXT-HOP = 2.2.2.2 into its local forwarding table

Data packet to 18.1.1.1 is received by R2

Recursive table lookup at R2 can be used

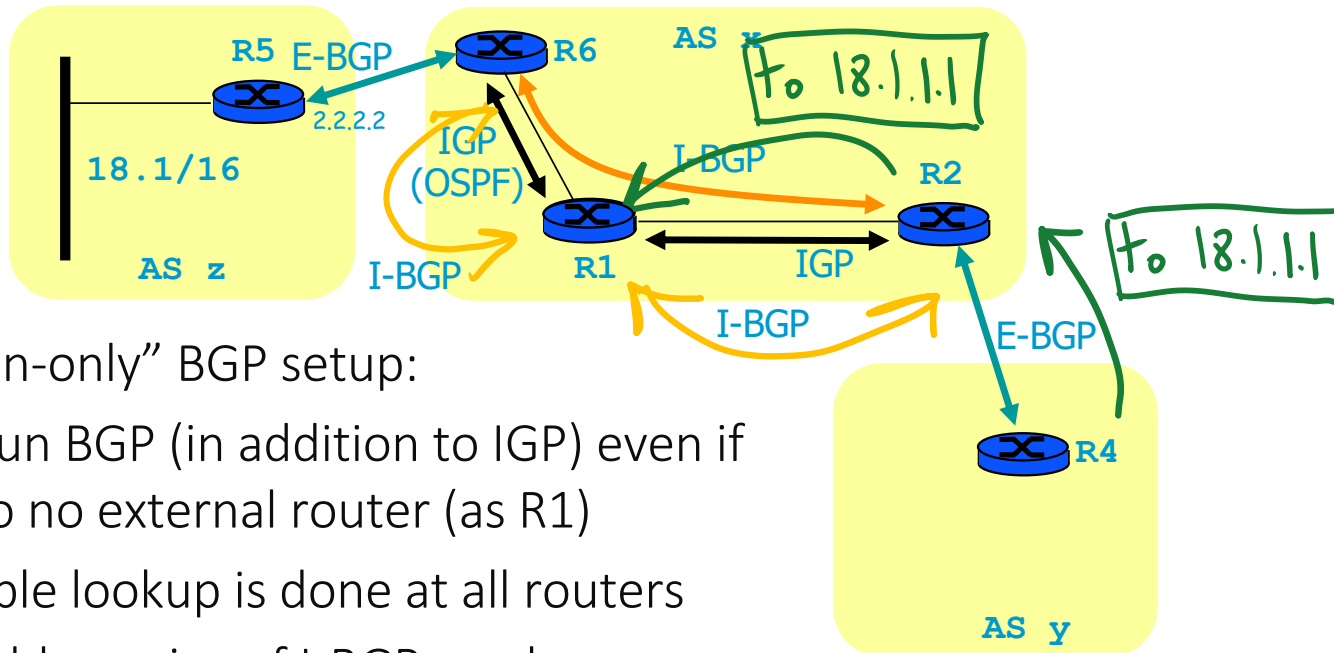
Packet is sent to R1

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

Injection in Practice Requires all Routers to Run BGP



The “injection-only” BGP setup:

All routers run BGP (in addition to IGP) even if connected to no external router (as R1)

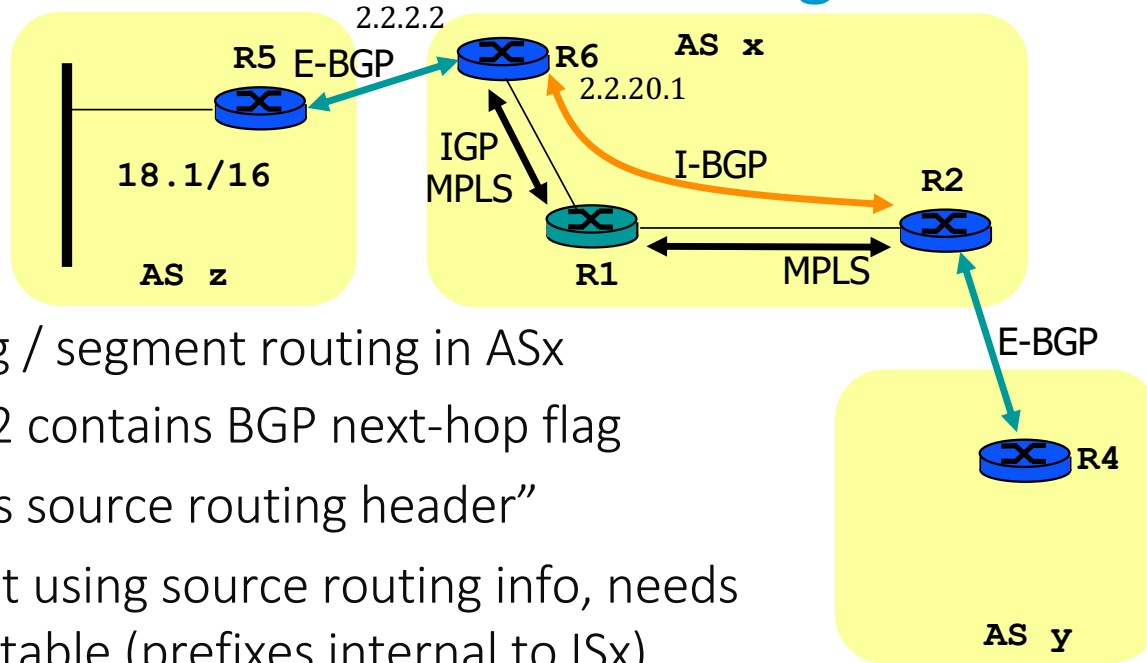
Recursive table lookup is done at all routers

Potential problem: size of I-BGP mesh -> use reflectors

IGP is still needed to discover paths to next-hops; but handles only internal networks – very few

Alternative : BGP with Source Routing

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



Use source routing / segment routing in ASx

Routing table at R2 contains BGP next-hop flag

“insert next-hop as source routing header”

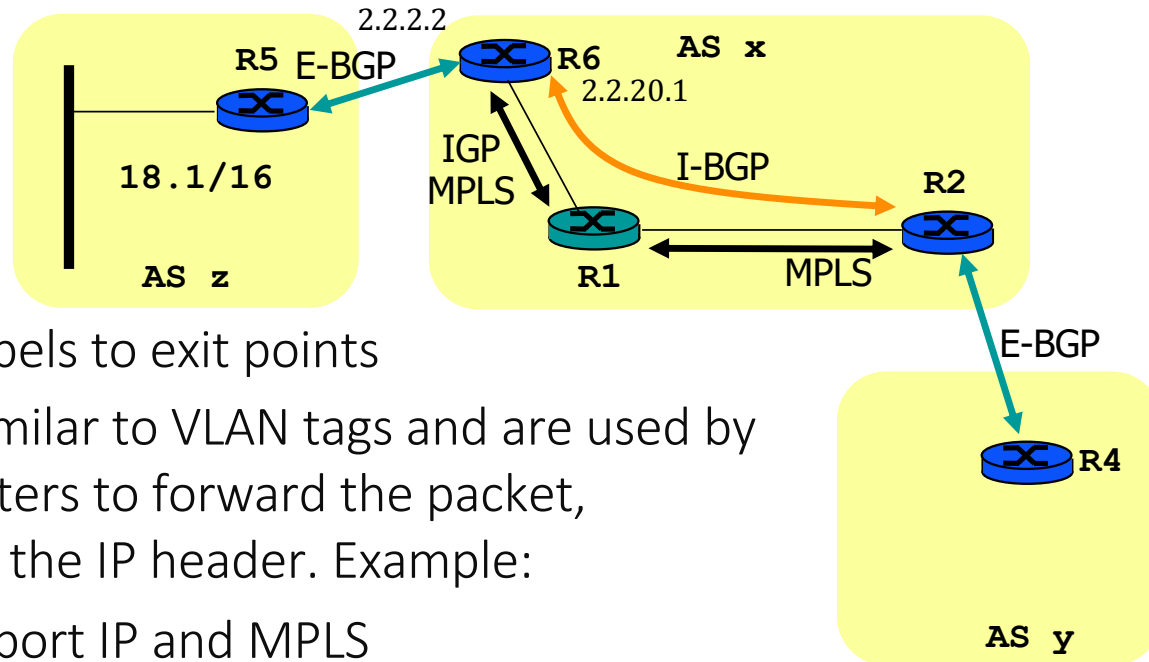
R1 forwards packet using source routing info, needs only small routing table (prefixes internal to ISx).

	<i>To</i>	<i>NEXT-HOP</i>	<i>Flags</i>
<i>at R2</i>	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 (see in a later module)

	<i>To</i>	<i>NEXT-HOP</i>	<i>layer-2 addr</i>
at R2	18.1/16	2.2.2.2	MPLS label 23

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 e.g. both OSPF and BGP.

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

E-BGP = 20, OSPF = 110, RIP = 120, I-BGP = 200

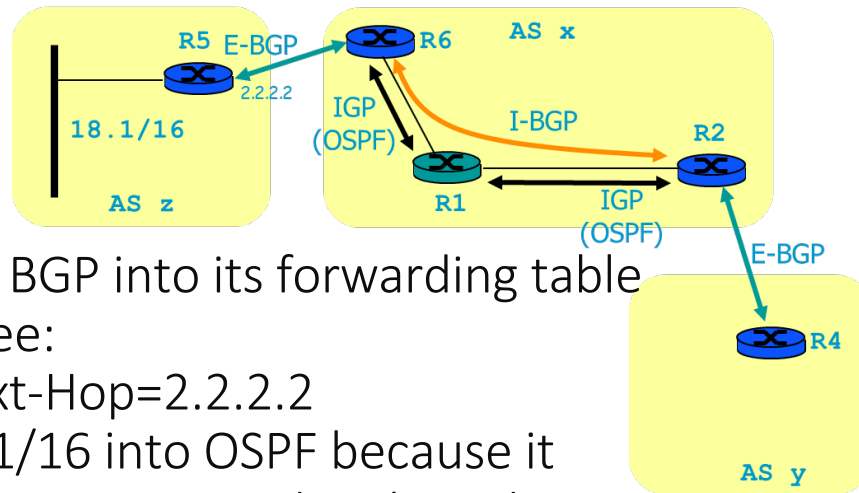
Only the route with the smallest administrative distance is selected to forward IP packets. Administrative distance is compared before the usual distance.

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

Furthermore, 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.

Example

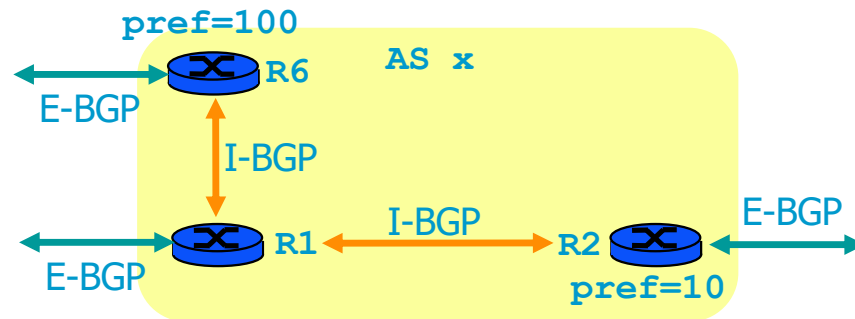
Assume R2 and R6 redistribute E-BGP into OSPF.



- 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



Used inside an AS to select a best *AS path*

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

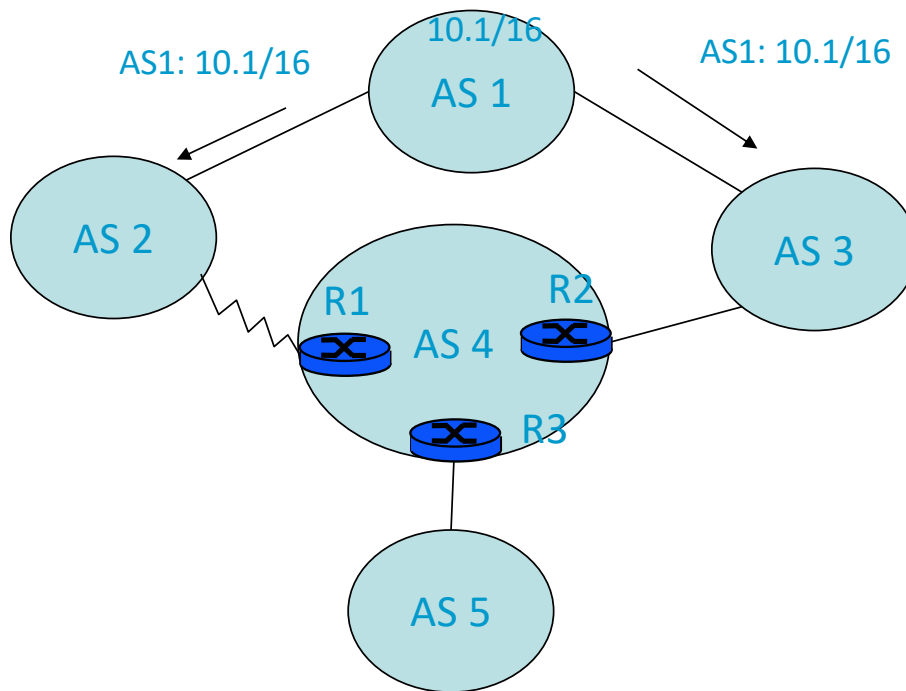
R1 chooses the largest preference

LOCAL-PREF Example: Link AS2-AS4 is expensive

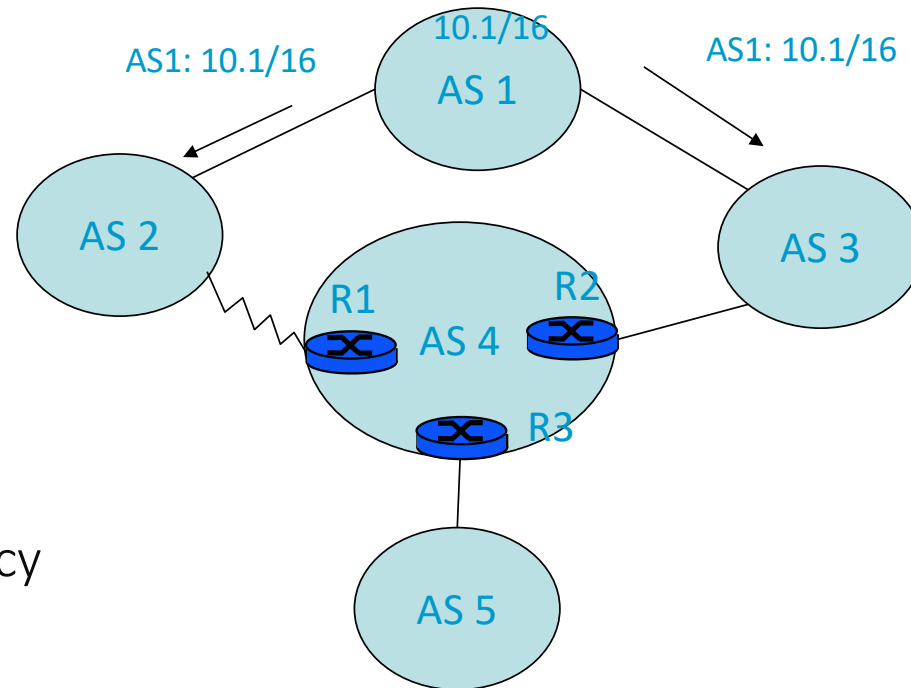
AS 4 sets LOCAL-PREF to 100 to all routes received from AS 3 and to 50 to all routes received from AS 2

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
- D. Both
- E. None
- F. I don't know

Weight

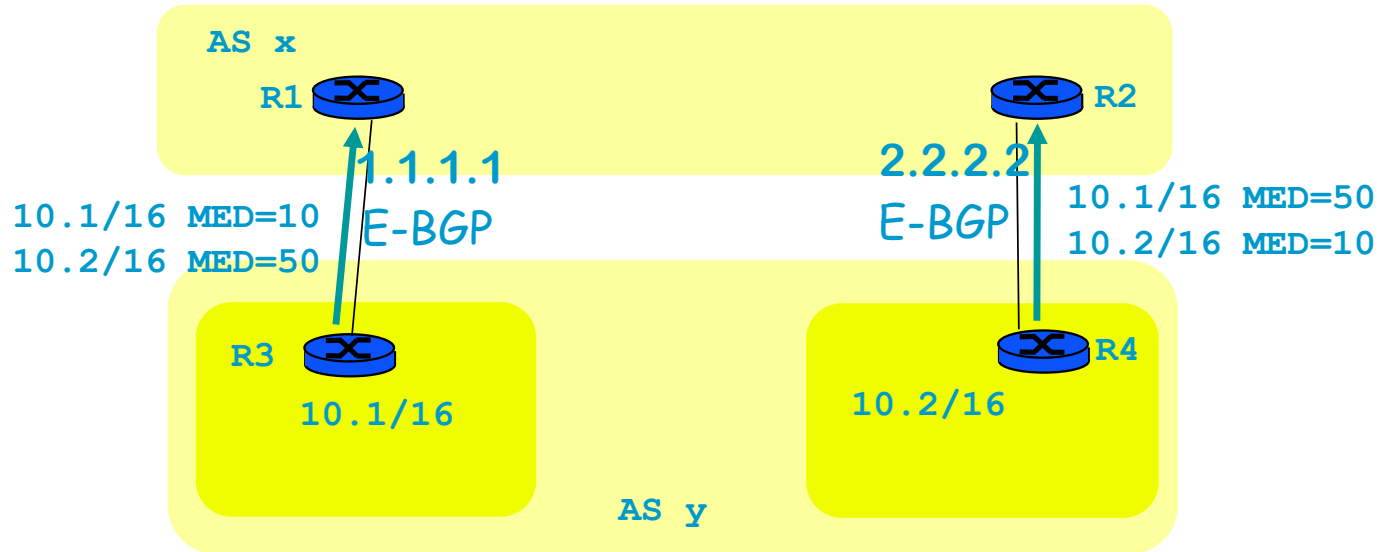
This is a route attribute given by Cisco or similar router

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

MULTI-EXIT-DISC (MED)



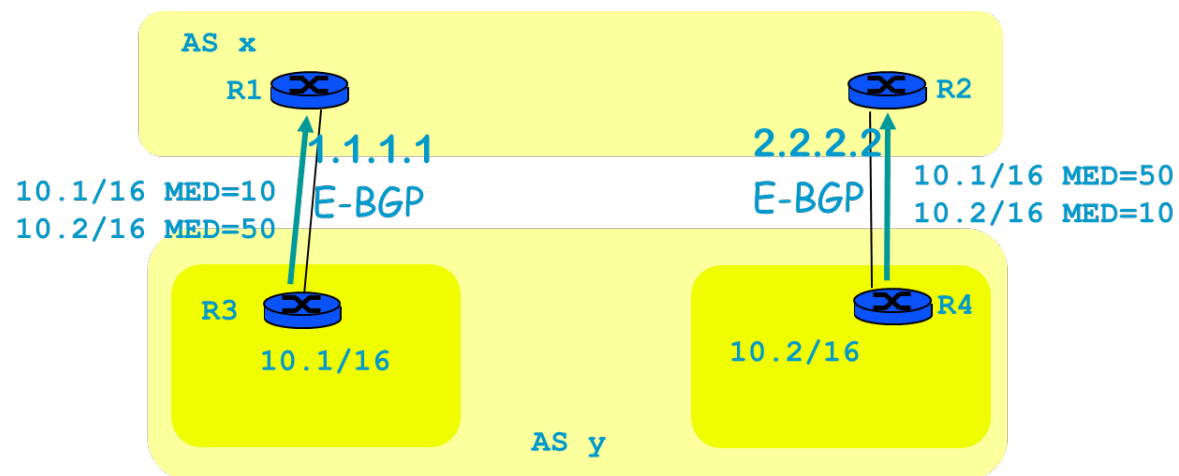
One AS connected to another over several links

ex: multinational company y connected to worldwide ISP x

AS y advertises its prefixes with different MEDs (low = preferred)

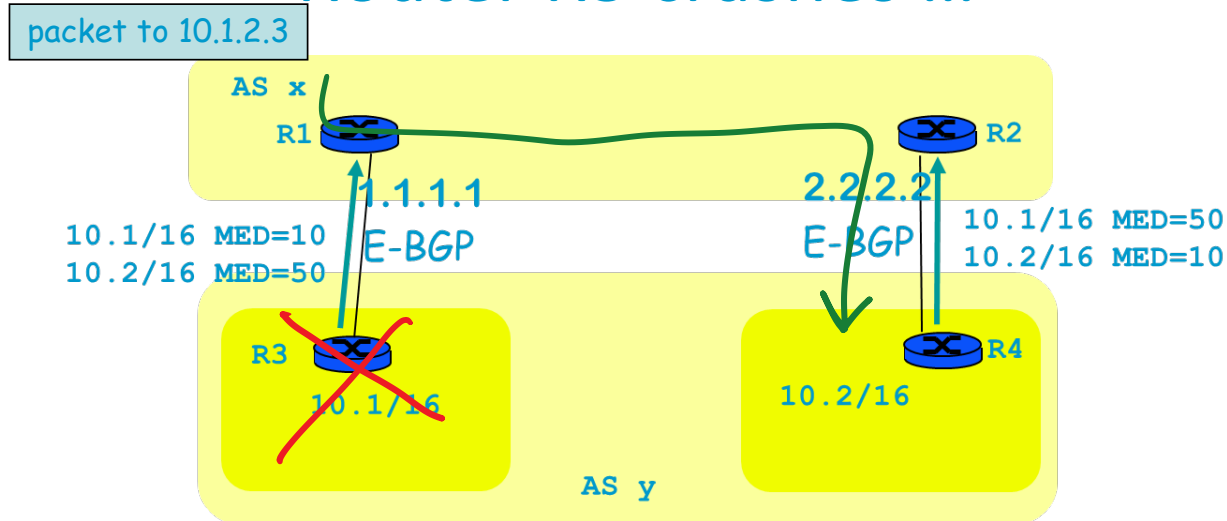
If AS x accepts to use MEDs put by AS y: traffic goes on preferred link

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

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

LOCAL-PREF vs MED

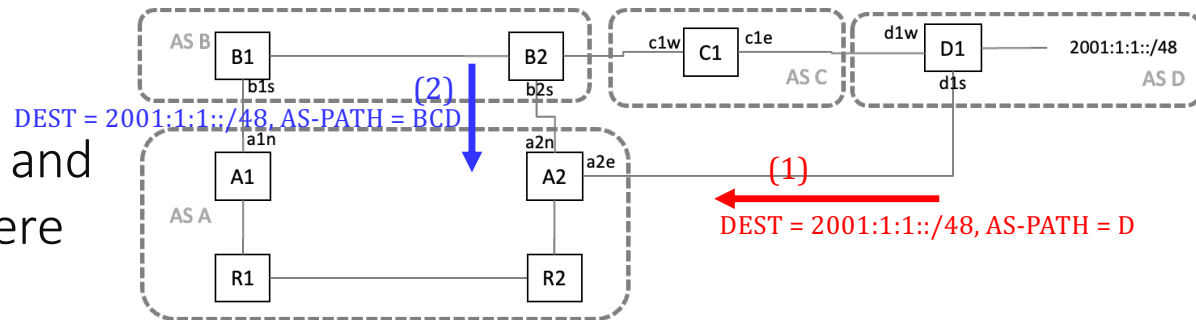
MED is used between ASs (i.e. over E-BGP); LOCAL-PREF is used inside one AS (over I-BGP)

MED is used e.g. to tell one provider AS which entry link to prefer;

LOCAL-PREF is used inside one AS to propagate AS path preferences -- this allows this AS to tell the rest of the world which AS path it wants to use, by not announcing the other paths.

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



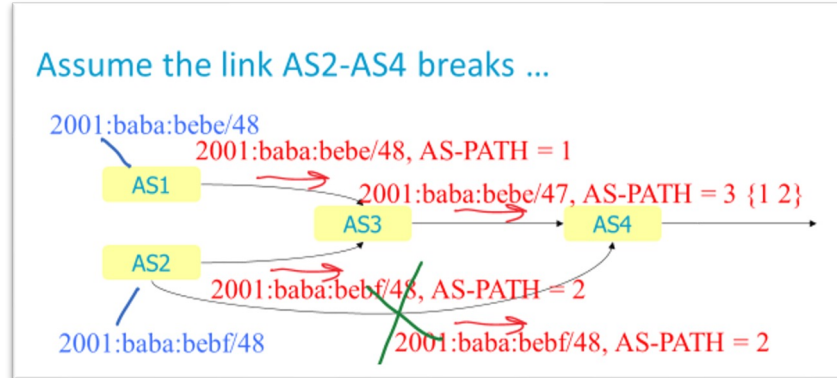
Example: A prefers B over D and sets LOCAL-PREF =100 to updates received from B

- 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, depending on message timings.

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



5. Other Bells and Whistles

Route Flap Damping

Recall that with BGP, routes are explicitly withdrawn (and updated).

Route flap : a route is successively withdrawn, updated, withdrawn, updated etc. The flap propagates to the AS and to other ASs. Causes CPU congestion on BGP routers.

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

Route flap damping (also called dampening) mitigates this:

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

- WITHDRAW \Rightarrow penalty incremented;

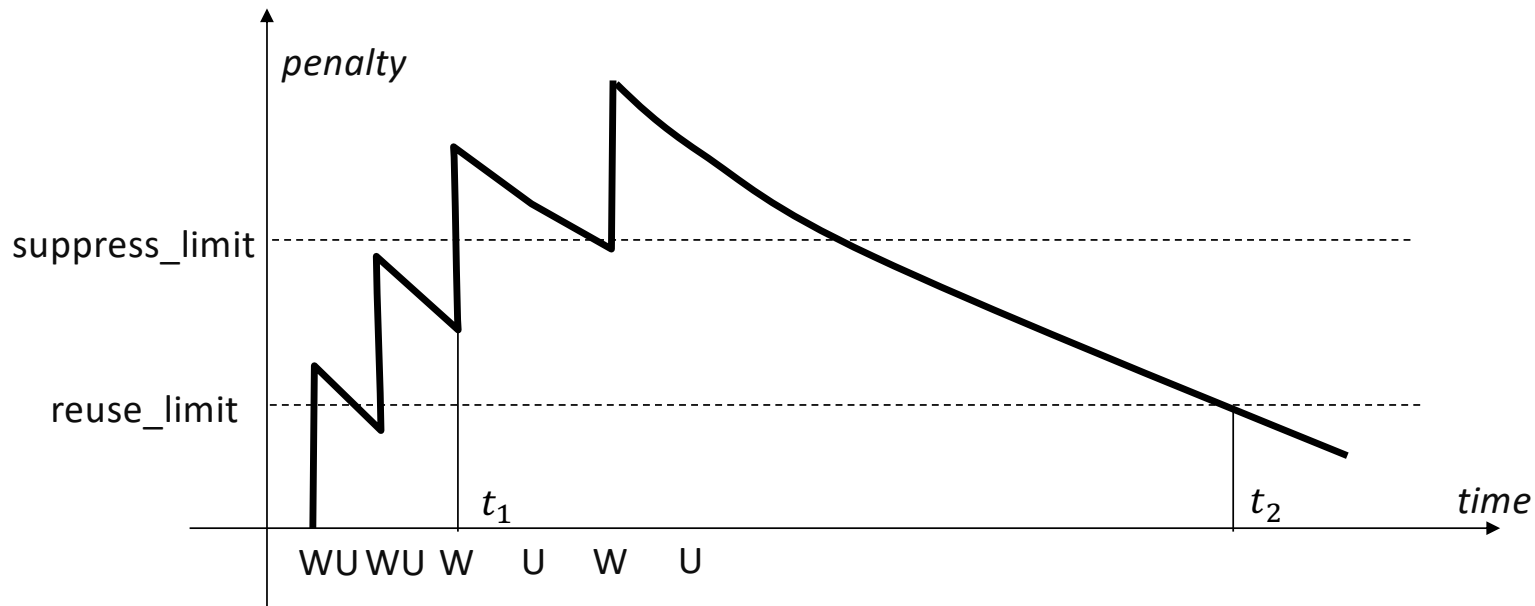
- updated ADVERTISEMENT \Rightarrow if penalty > suppress_limit then SUPPRESS= true

- 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



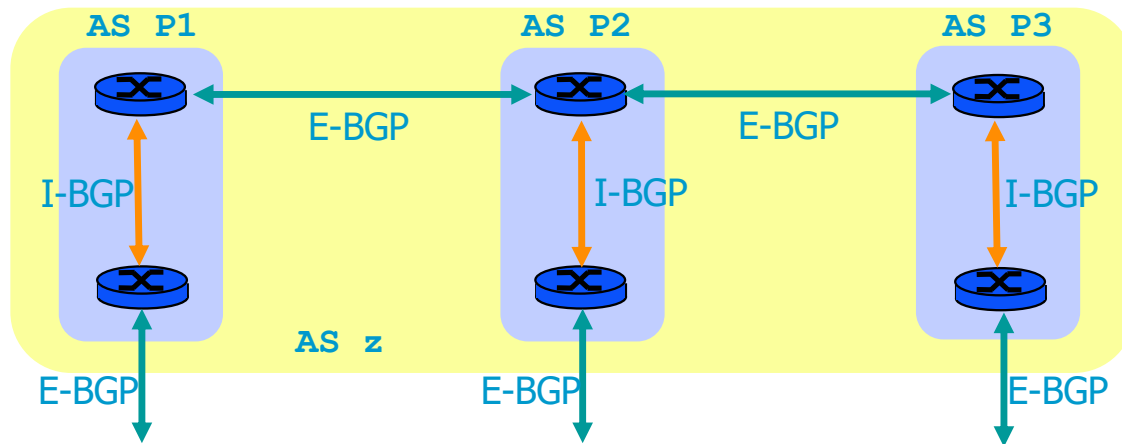
Client uses BGP with MED to control flows of traffic (e.g provider should use R1-R3 for all traffic to 10.1/16)

Client (e.g., EPFL) can use a *private AS number* -- not usable in the global internet, used only between Client and Provider (eg. 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

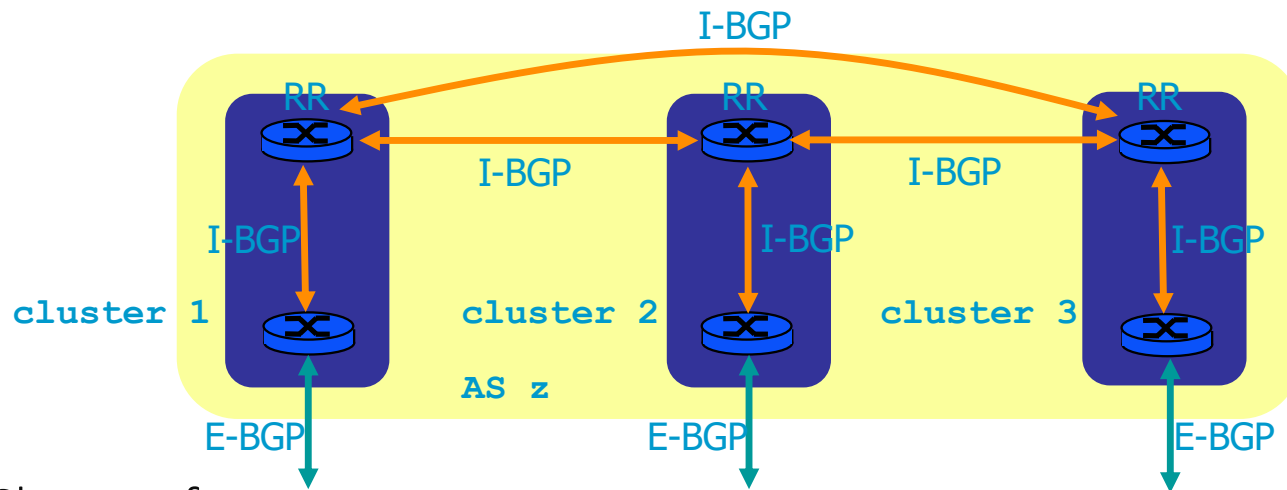
private AS number

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 one client and RR

Route reflector

re-advertises a route learnt via I-BGP

to avoid 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

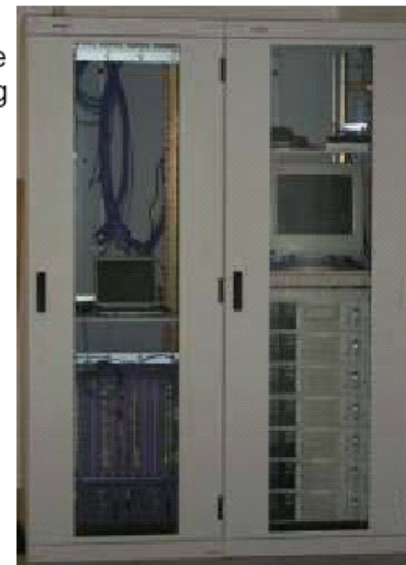
Links

Welcome to swissix

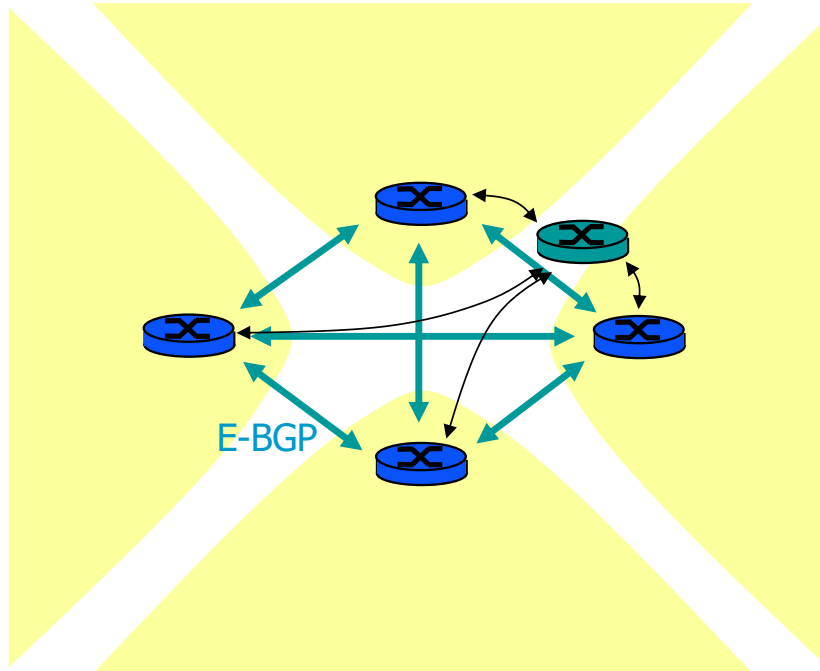
The Swissix (Swiss Internet Exchange) in Zurich, Switzerland, is now open. We are pleased to welcome ISPs and hosting companies as members and peering partners.

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



At interconnection point

Instead of $n(n-1)/2$ peer-to-peer E-BGP connections, n connections to Route Server

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

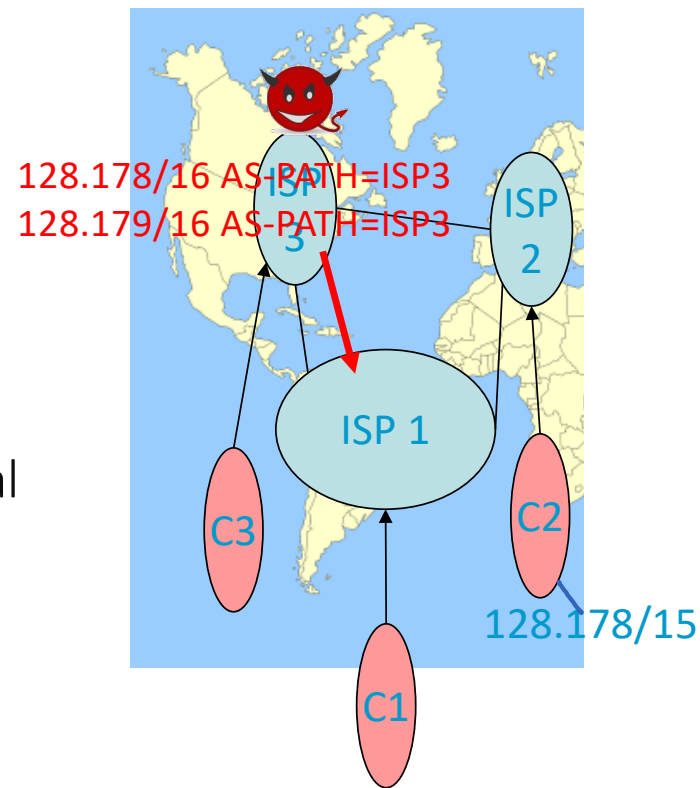
6. Security Aspects

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

Example 1: Assume ISP3 (malicious) announces to ISP1 a route to 128.178/16 and a route to 128.179/16

What will happen for traffic from C1 to EPFL?

- 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



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).
ASs can read Routing Registries and use it 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

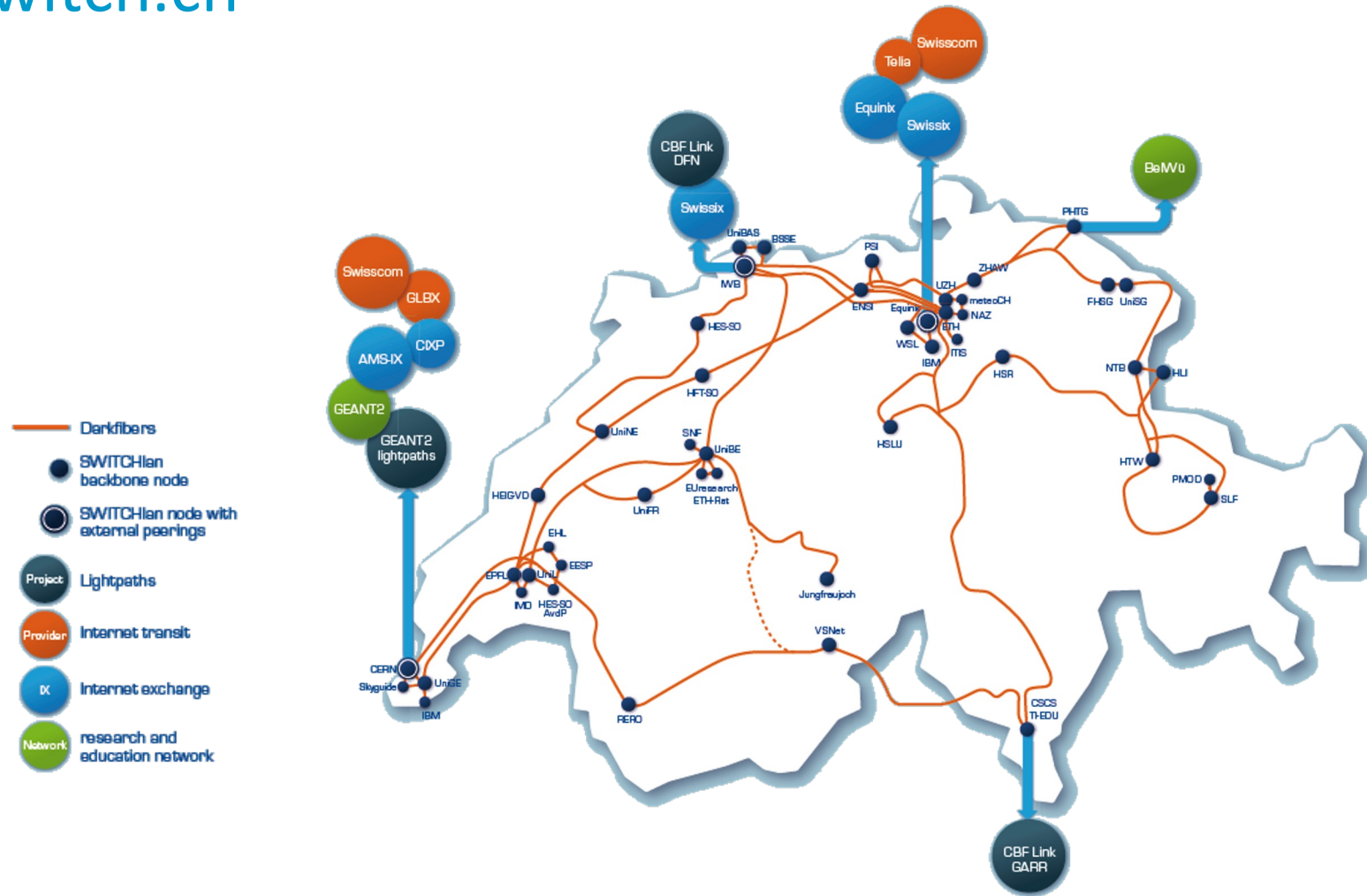
AS-PATH attribute replaced by BGPsec_Path attribute that contains the
AS path + signatures of every segment of the path performed by every
intermediate AS

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

ks

Network Info

Whois

DNS

IRR

[Home](#)

[Report](#)



[Report](#)

[Report](#)

[as](#)

[t](#)

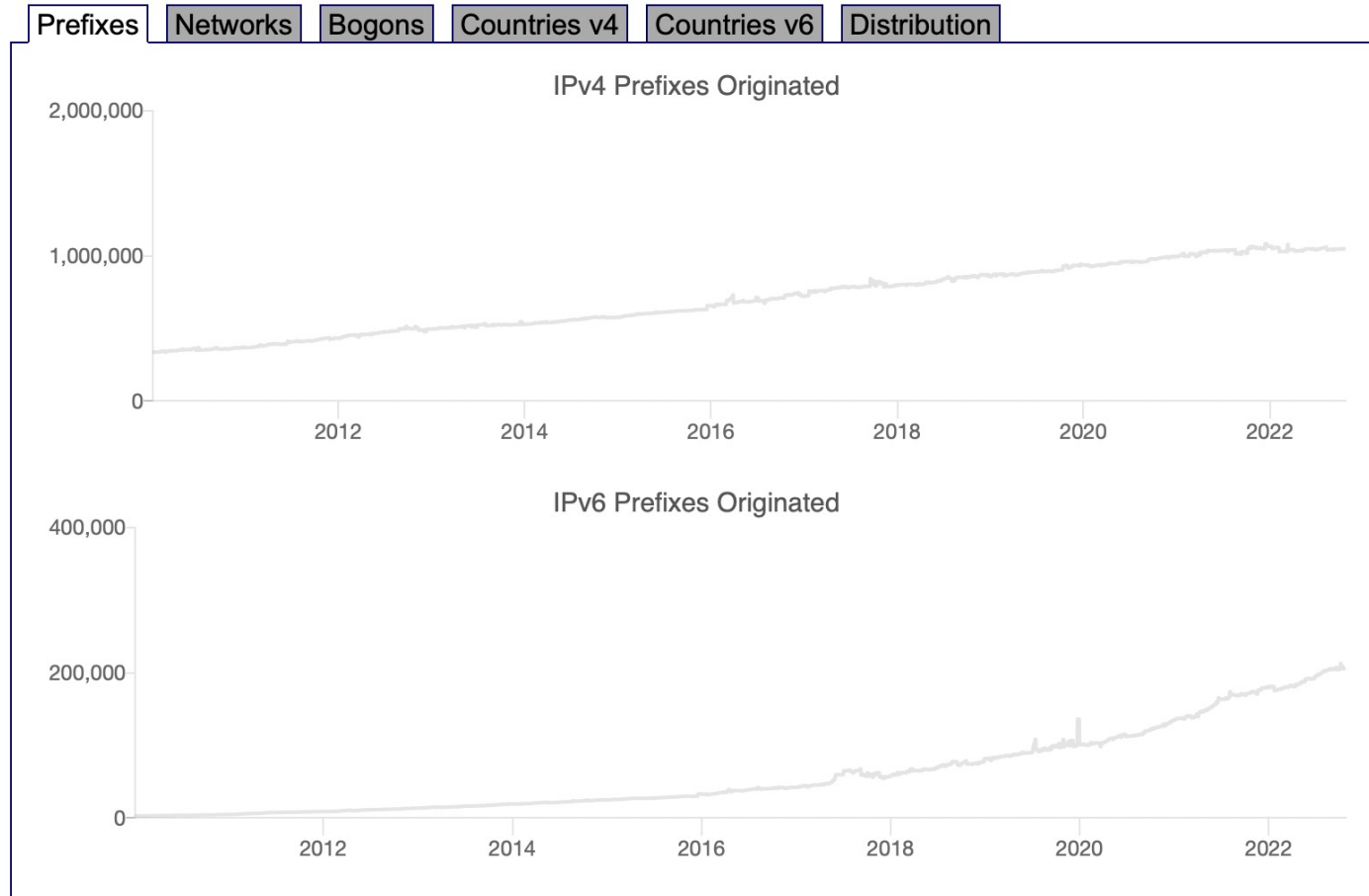
Announced By

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



ROA signed and valid

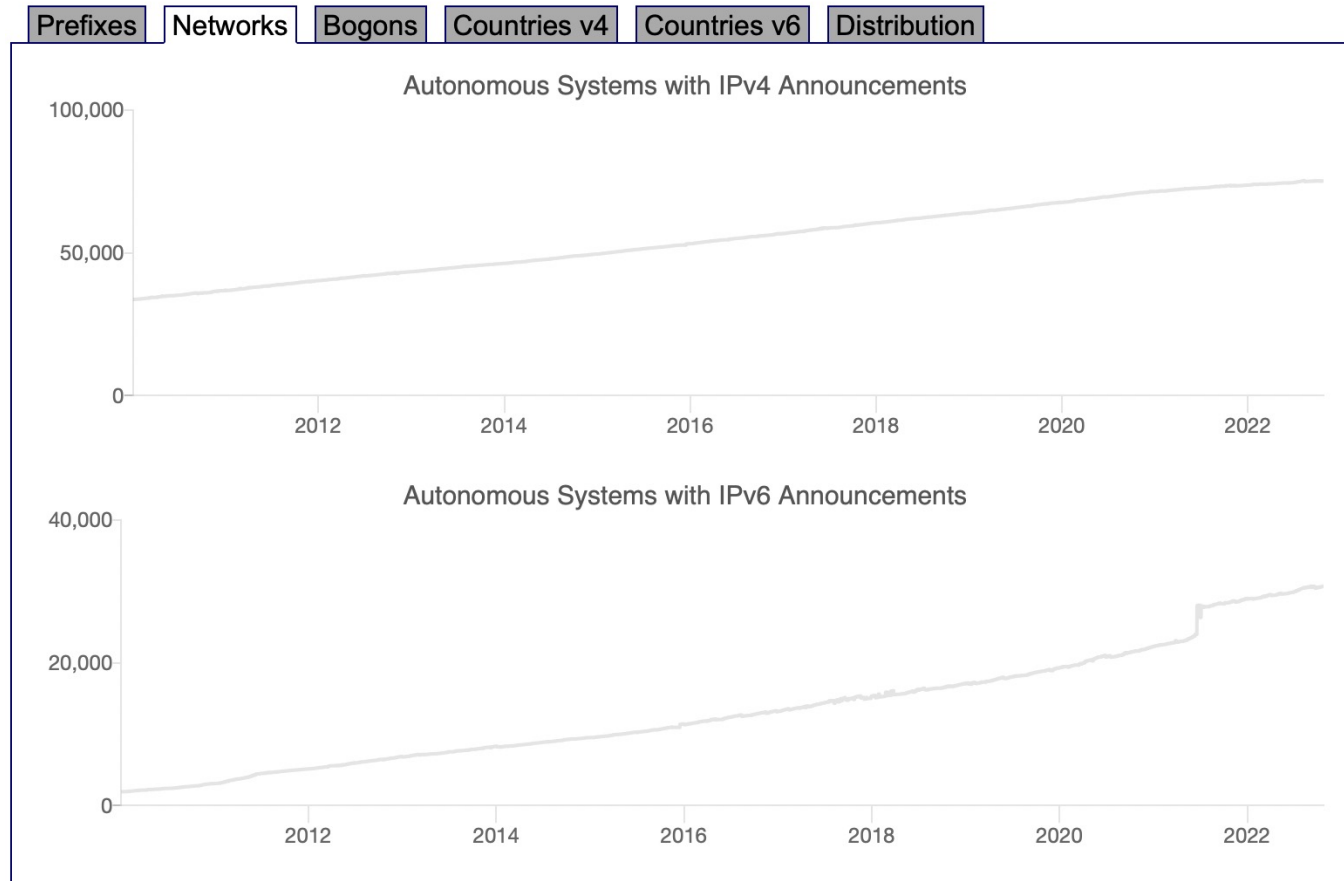
Number of announced prefixes



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.