External Routing
BGP
Jean-Yves Le Boudec
2017
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

Textbook
Section 5.1.1, The control plane
A. 1. Inter-Domain Routing

Why invented?
The Internet is too large 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, IGPs): inside ASs:
  RIP, OSPF (standard), IGRP (Cisco)
- **exterior**: between ASs:
  BGP (today)
  EGP (old) and BGP-1 to BGP-4 (today), IDRP (tomorrow, maybe)
What is an ARD? An AS?

ARD = Autonomous Routing Domain
    = routing domain under one single administration
    - one or more border routers
    - all subnetworks inside an ARD should be connected
    - should learn about other subnetwork prefixes - the routing tables of internal routers should contain entries of all destination of the Internet

AS = Autonomous System
    = ARD with a number (“AS number”)
    - AS number is 32 bits denoted with dotted integer notation e.g. 23.3456
    - 0.559 means the same as 559
    - Private AS numbers: 0.64512 – 0.65535

ARDs that are do not need a number are typically those with default route to the rest of the world

Examples
    - AS1942 - CICG-GRENOBLE, AS2200 - Renater
    - AS559 - SWITCH Teleinformatics Services
    - AS5511 – OPENTRANSIT
    - EPFL: one ARD, no number
**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 \((\text{path}: \text{dest})\) 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, 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
AS A is a stub; A2 has default route to D6; could AS A avoid using BGP?

A. No because interdomain routing requires an interdomain routing protocol
B. Yes, if A2 injects the destinations 0/0 and ::/0 into its IGP
C. Yes but using another method
D. I don’t know
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 onlink of each other.
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 → A : D3 — D2 — X : n1
2. D → E : D2 — X : n1
3. C → 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: NAP (Network Access Point), MAE (Metropolitan Area Ethernet), CIX (Commercial Internet eXchange), GIX (Global Internet eXchange), IXP, SFINX, LINX

Mainly 3 types of relations

- Customer-provider: EPFL is customer of Switch. EPFL pays Switch
- Shared Cost peer: EPFL and CERN are peers: costs of interconnection is shared
- Siblings: EPFL-Ecublens and EPFL-IMT are inside the same organization
What is the Goal of Policy Routing?

Example:

- ISP3-ISP2 is transatlantic link, cost shared between ISP2 and ISP3
- ISP3-ISP1 is a local, inexpensive link
- Ci is customer of ISPi, 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 ISP 2 and ISP3

The goal of “policy routing” is to support this and other similar requirements
Policy Routing can be implemented by Import and Export filters

Example:

Import Filter for ISP1:
- from C1 accept C1
- from ISP2 accept ISP2
- from ISP2 accept ISP2 C2
- from ISP3 accept ISP3
- from ISP3 accept ISP3 C3

ISP1 does not accept
ISP2 ISP3 C3
because ISP1 does not want to use ISP2-ISP3 as transit

Export Filter for ISP1:
- to C1 announce ANY
- to ISP2 announce ISP1
- to ISP2 announce ISP1 C1
- to ISP3 announce ISP1
- to ISP3 announce ISP1 C1

ISP1 does not announce ISP2 or ISP2 C2 to ISP3
because ISP1 does not want to offer a transit service to ISP2 or ISP3
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
- In steady state, 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

- BGP Adj-RIB-In
  - Peer[N]
  - BGP Mgs from Peer[N]
  - Import filter
  - Attribute manipulation

- BGP Loc-RIB
  - All acceptable routes

- BGP Decision Process
  - One best route to each destination

- BGP Adj-RIB-Out
  - Peer[N]
  - Export filter
  - Attribute manipulation
  - BGP Mgs to Peer[N]

- FIB
  - Routing table

- routes obtained locally (redistributed)
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)
- Attributes
  - Well-known Mandatory: ORIGIN (route learnt from IGP, BGP or static), AS-PATH, NEXT-HOP
  - Well-known Discretionary: LOCAL-PREF (see later), ATOMIC-AGGREGATE (= route cannot be dis-aggregated)
  - Optional Transitive: AGGREGATOR (who aggregated this route)
  - Optional Nontransitive: WEIGHT (see later), MULTI-EXIT-DISC (MED, see later)

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. The default 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
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 → R1
10.1/16 AS = y
10.2/16 AS = y

Adj-RIB-in

<table>
<thead>
<tr>
<th>From R3</th>
<th>10.1/16 AS = y NEXT-HOP=1.1.1.2</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>From R3</td>
<td>10.2/16 AS = y NEXT-HOP=1.1.1.2</td>
<td>Best</td>
</tr>
</tbody>
</table>

Adj-RIB-out

<table>
<thead>
<tr>
<th>To R2</th>
<th>10.1/16 AS = y NEXT-HOP=1.1.1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>To R2</td>
<td>10.2/16 AS = y NEXT-HOP=1.1.1.2</td>
</tr>
</tbody>
</table>

(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.
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
Fundamental Example, Continued

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 → 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 → R1

10.2/16 AS = y NEXT-HOP = 2.2.2.1

---

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

---

**Adj-RIB-in**

<table>
<thead>
<tr>
<th>From</th>
<th>10.1/16 AS = y NEXT-HOP = 1.1.1.2</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>R3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R22</td>
<td>10.2/16 AS = y NEXT-HOP = 2.2.2.1</td>
<td>Best</td>
</tr>
</tbody>
</table>

**Adj-RIB-out**

<table>
<thead>
<tr>
<th>To</th>
<th>10.1/16 AS = y NEXT-HOP = 1.1.1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>R21</td>
<td></td>
</tr>
<tr>
<td>R22</td>
<td>10.1/16 AS = y NEXT-HOP = 1.1.1.2</td>
</tr>
</tbody>
</table>
Step 3

\[ R21 \rightarrow R1 \]

10.2/16 AS =y NEXT-HOP=3.3.3.1

Adj-RIB-in

<table>
<thead>
<tr>
<th>From R3</th>
<th>10.1/16 AS =y NEXT-HOP=1.1.1.2</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>From R22</td>
<td>10.2/16 AS =y NEXT-HOP=2.2.2.1</td>
<td>Best</td>
</tr>
<tr>
<td>From R21</td>
<td>10.2/16 AS =y NEXT-HOP=3.3.3.1</td>
<td></td>
</tr>
</tbody>
</table>

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. Both in parallel
D. I don’t know
How are routes originated?

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

Several methods

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

- **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.
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?

1. Yes and the aggregated prefix is 2001:baba:bebe/47
2. Yes and the aggregated prefix is 2001:baba:bebf/48
3. Yes but the aggregated prefix is none of the above
4. No
5. 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 known
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:bebe/48
There is none
The routing table entry 2001:baba:bebf/48 is removed
3. Forwarding Entries learnt by BGP are written into Routing Tables

So far, we have seen how BGP routers learn about all the 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 quagga (in the lab): this is always done.
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

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. RIP)

If redistribution is avoided, only *injection* is used, i.e. BGP routes are written directly into the forwarding table.
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: 2.2.2.3  
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

<table>
<thead>
<tr>
<th>To</th>
<th>next hop</th>
<th>interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.1/16</td>
<td>2.2.2.63</td>
<td>N/A</td>
</tr>
<tr>
<td>2.2.2.63</td>
<td>2.2.2.33</td>
<td>eth0</td>
</tr>
</tbody>
</table>
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
**BGP with MPLS**

Alternative to redistribution or running I-BGP in all backbone routers:
- Associate MPLS labels to exit points
- MPLS labels are somewhat 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 R3 support IP and MPLS
  - R2 creates a “label switched path” to 2.2.2.2
  - 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
  - Can also be used to provide quality of service

<table>
<thead>
<tr>
<th>To</th>
<th>NEXT- HOP</th>
<th>Layer- 2 addr</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.1/16</td>
<td>2.2.2.2</td>
<td>MPLS label 23</td>
</tr>
</tbody>
</table>

*RIB and LIB at R2*
Injection Conflicts

In quagga and cisco, BGP always injects routes into forwarding table, even if BGP is redistributed into IGP. This causes injection conflicts: a route may be injected into the forwarding table by e.g. both OSP and BGP.

To solve the conflicts, every route in the forwarding table has a 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 smaller administrative distance is selected to forward IP packets.

Furthermore, the decision process selects a BGP route only if there is no route with smaller administrative distance in the forwarding table.
Example

Assume R2 and R6 redistribute BGP into OSPF and assume R2 is first to do it.

at $t_1$: R2 redistributes 18.1/16 from BGP into OSPF
In R2’s forwarding table we see:
18.1/6, Admin Dist = 200, cost =1, Next-Hop 2.2.2.2

at $t_2 > t_1$: via OSPF R6 learns the route and puts it into forwarding table:
18.1/6, Admin Dist =110, cost =3, Next-Hop 2.2.2.2

at $t_3 > t_2$: R6 redistributes 18.1/16 from BGP into OSPF
in R6’s forwarding table we see:
18.1/6, Admin Dist = 20, cost =1, Next-Hop 2.2.2.2
18.1/6, Admin Dist =110, cost =3, Next-Hop 2.2.2.2

The first route is preferred, OSPF now propagates only the first

at $t_4 > t_3$: via OSPF R2 learns the route and puts it into forwarding table:
18.1/6, Admin Dist =110, cost =3, Next-Hop 2.2.2.2
18.1/6, Admin Dist =200, cost =1, Next-Hop 2.2.2.2
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

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. Both
D. None
E. 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 cloud
- Therefore there is no weight attribute in route announcements
MULTI-EXIT-DISC (MED)

One AS connected to another over several links

- ex: multinational company connected to worldwide ISP
- 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 ASy 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 to tell one provider AS which entry link to prefer; LOCAL-PREF is used to tell the rest of the world which AS path we want to use, by not announcing the other ones.
Communities

Other attributes can be associated with routes in order to *simplify* rules. They are called « communities »

- Pre-defined: Example: NO-EXPORT (a well known, pre-defined attribute) – see later for an example
- Defined by one AS (a label of the form ASN:x where AS= AS number, x = a 2 byte—number)
NO—EXPORT

Written on E-BGP by one AS, transmitted on I-BGP by accepting AS, not forwarded
Example: AS2 has different routes to AS1 but AS2 sends only one aggregate route to AS3
▶ simplifies the aggregation rules at AS2
Import Filter for ISP1:
from C1 accept C1
from ISP2 accept ISP2 ANY
from ISP3 accept ISP3 ANY

ISP1 accepts all routes from its peers ISP2 and ISP3 but would still like to avoid ISP2-ISP3 as transit, except as a backup.
Is that possible? How?
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...).

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 ⇒ penalty incremented;
- UPDATE ⇒ if penalty > suppress_th then SUPPRESS= true
- penalty is updated e.g. every<5 sec, with exponential decay; when penalty <suppress_th 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 UPDATE
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
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 can use a *private AS number* -- not usable in the global internet, used only between Client and Provider

Provider translates this number to his own when exporting routes to the outside world
Which way will packets from z to 2.1.1.1 go?

A. Via x
B. Via y
C. Both
D. I don’t know
Avoid 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
Avoid 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
Avoid 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) originates a route to 128.178/16 and a route to 128.179/16.

What will happen?

A. All traffic to EPFL will go to ISP3
B. Some fraction of traffic to EPFL will go to ISP3
C. All traffic to EPFL will go normally to C2
D. I don’t know
BGP Security

Forged AS paths, destination prefix, next-hop etc cause some fraction / all traffic to go to malicious ISP -> used to cut service / spy / forge malicious transactions

BGP security measures (under development):
- Authentication of ASs number and IP address blocks: RPKI: public key infrastructure rooted at IANA (internet number authority) used to authenticate originations
- BGPsec: all BGP messages are cryptographically signed: one signature at origination + one signature added for every modification
  all signatures in BGP messages are verified using RPKI
  requires crypto support in routers
C. Illustrations: The Switch Network

www.switch.ch
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.
Registry Browser (AS559)

aut-num: AS559

as-name: SWITCH

descr: SWITCH, Swiss Education and Research Network

admin-c: WH1101

tech-c: SNOC1-RIPE

mnt-by: RIPE-NCC-END-MNT

mnt-by: SWITCH-MNT

mnt-lower: AS559-MNT

Routing Status (AS559)

AS559 is visible by 100% of 98 IPv4 and 100% of 88 IPv6 RIS full peers.

First ever seen as origin announcing 152.96.0.0/16, on 2000-08-20 00:00:00Z.

Originated IPv4 prefixes: 95

Originated IPv6 prefixes: 5

Observed BGP neighbours: 48
Number of announced prefixes

IPv4 Prefixes Originated (1000 Days)

IPv6 Prefixes Originated (1000 Days)

seen by Hurricane Electric: bgp.he.net