BGP and Interdomain routing

COMPUTER NETWORKS A.A. 24/25

Leonardo Maccari, DAIS: Ca' Foscari University of Venice,
leonardo.maccari@unive.it

Venice, fall 2024

Sect. 1 Introduction to BGP

Introduction to Internet Architecture

- the Internet is a network of networks.
- This literally means that it is made of separate networks that are connected with each other and exchange traffic
- We know how intradomain routing works, we are now going to introduce how routing works networks that are owned by different administrative entities

Autonomous Systems (AS)

Autonomous System: An autonomous system is a collection of connected Internet Protocol (IP) routing prefixes under the control of one or more network operators on behalf of a single administrative entity or domain [Hawkinson and Bates, 1996]

There are two kinds of ASes...

Stub AS [1]

Stub AS: This type of AS does not provide transit services to others. They are end customers, such as universities, organizations, and most companies. Some stub ASes only connect to one other AS (typically an upstream provider). They are called single-homed stub ASes. Some stub ASes connect to multiple ASes, and they are called multi-homed stub ASes. A multi-home stub AS will not allow traffic from one AS to pass through to another AS, i.e., it does not provide transit service.

[1] Definitions from Kevin Du Book.

Transit AS

Transit AS: Transit AS: This type of AS connects to multiple ASes, and offer to route data from one AS to another AS. It provides transit services.

ASes, in practice

- Each AS is a collection of networks using their own interal routing protocol
- What makes them an AS is that they are owned and managed by the same entity: An ISP, a company, government, an university...
- So there is not a strict technical definition of AS, it references to some administrative entity.
- Each AS need to be assigned a unique number from the IANA, the Internet Assigned Numbers Authority

ASes, in practice

- ASes have Point of Presence (POP), that is, places where they are physically accessible.
- ASes are responsible for some prefixes, i.e. they manage sets of IP addresses
- These IP addresses must be reached by the rest of the Internet, so ASes need to connect to each other.

Connecting ASes

There are mainly two ways of connecting ASes

Private Peering: This is a Direct Point-to-point connection between two routers of two ASes. The owners of the ASes pay for the infrastructure needed to connect the two ASes, that is: fiber, wireless links, submarine cables, and the necessary hardware. Private peering is often not publicly declared, we don't know all the private peering (and thus the physical connections) that exist.

Connecting ASes

Internet Exchange Points (IXP): A shared data-center where ASes put their POPs and connect to each other. ISPs are more efficient, as with one single POP an AS can connect to many other ASes.

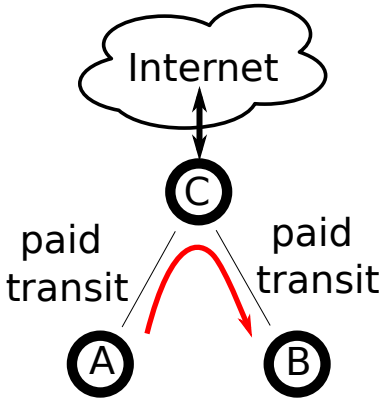
Connecting ASes

Transit Agreement: A commercial agreement between AS A and B in which B offers to connect A to the Internet. B behaves as a *gateway* for A.

Peering Vs Transit

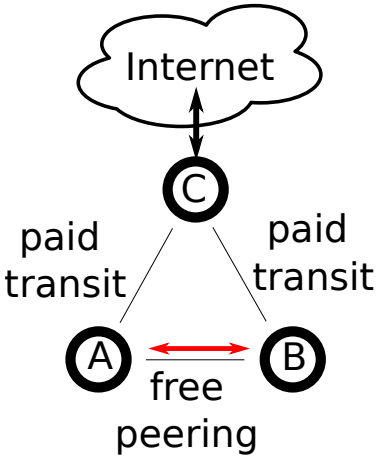
- A transit relationship generally includes a fee: entity A pays entity B so that traffic from A can reach the Internet
- A peering agreement (also called *shared cost* or *free peering*) is generally a free deal between A and B so that they can exchange traffic without cost
- Peering agreements are made to save money

No Peering



- A and B both pay C to send and receive traffic from/to the Internet
- When A sends traffic to B, they both pay C

With Peering

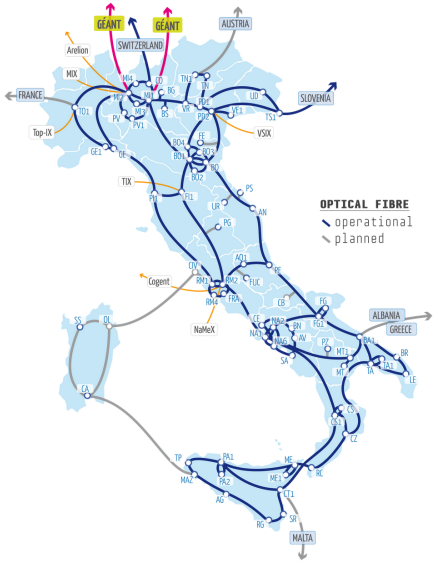


- A and B settle a free peering agreement
- Now they can exchange traffic without paying C.
- As long as peering is possible due to physical proximity, it is convenient to do.

Case Study: the GARR AS

- GARR is the national consortium that interconnects universities and research institutions in Italy.
- GARR is the entity that connects the physical networks that every University has
- GARR manages the IP addresses of the Universities
- Every university has a distinct network prefix

The Physical Network



- The GARR network has many PoP along the whole nation
- It is connected with the GÉANT network, that is the European network of research, with peering agreements.
- GARR is a transit network, as it offers connectivity from GÉANT to other countries (Albania, Greece, Malta, France. . .).
- GARR buys upstream bandwidth by two commercial providers (Arelion and Cogent)

GARR Presence in IXPs

- GARR is present in 5 IXP in Italy
- Within those IXP GARR makes peering agreements with tens of commercial entities

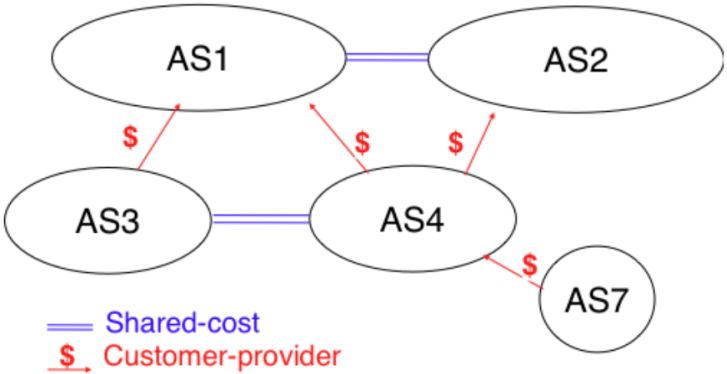
BGP

- While inside a single AS each network administrator can use the routing protocol that he/she prefers, the routing among ASes must use a single protocol
- This protocol is BGP, the Border Gateway Protocol.

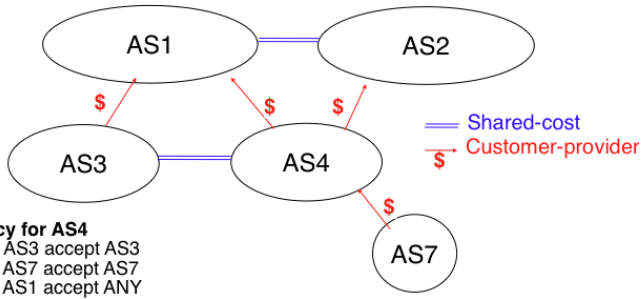
Exporting Prefixes

- We know that the role of a routing protocol is let routers exchange prefixes with other routers, so they can build the routing tables to any existing network
- In intra-domain routing this actuall happens, a border OSPF router always exposes all the prefixes it knows.
- In iter-domain routing this is not always true, because there are commercial relationships. In general:
  - over a customer→provider relationship, the customer domain advertises to its provider its own prefixes and all the routes that it has learned from its own customers.
  - over a provider→customer relationship, the provider advertises all the routes that it knows to its customer.
  - over a shared-cost peering relationship a domain only advertises its internal routes/prefixes and the routes that it has learned from its customers.

AS Relationships



AS Inport/Export Policy



Import policy for AS4
Import: from AS3 accept AS3
import: from AS7 accept AS7
import: from AS1 accept ANY
import: from AS2 accept ANY

Export policy for AS4
export: to AS3 announce AS4 AS7
export: to AS7 announce ANY
export: to AS1 announce AS4 AS7
export: to AS2 announce AS4 AS7

Import policy for AS7
Import: from AS4 accept ANY

Export policy for AS7
export: to AS4 announce AS7

Sect. 2 BGP

BGP Routers

- Every AS needs at least one router that supports the BGP protocol
- BGP routers announce network prefixes: an announcement is like saying *The network with prefix X is homed in this AS*, similarly to what happens with OSPF
- BGP routers announce their own prefixes, but can also announce other prefixes from other ASes

Border routers

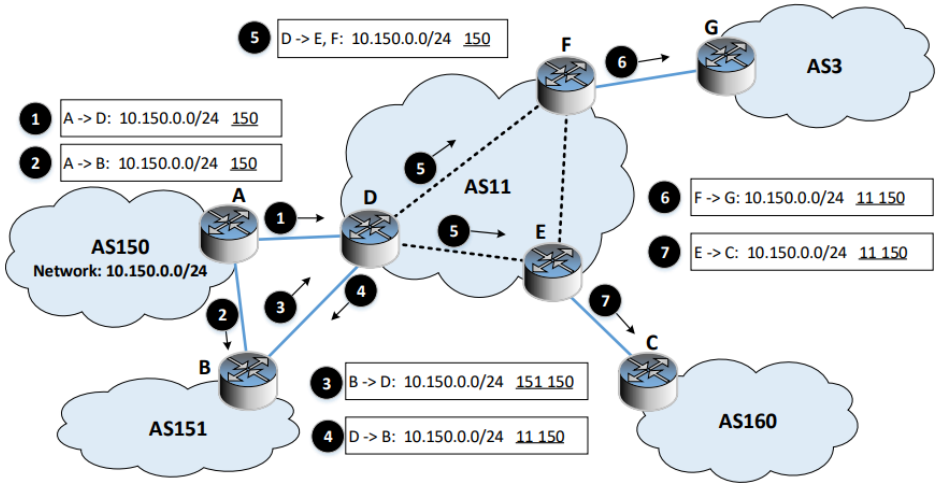
- A border router will run both BGP and the internal routing protocol (such as OSPF)
- The routing daemons (the software that implements the control messages) produce the routing tables
- The routing tables will then be merged into a forwarding table.

Forwarding Tables

- The decision is based on the content of a *forwarding table*.

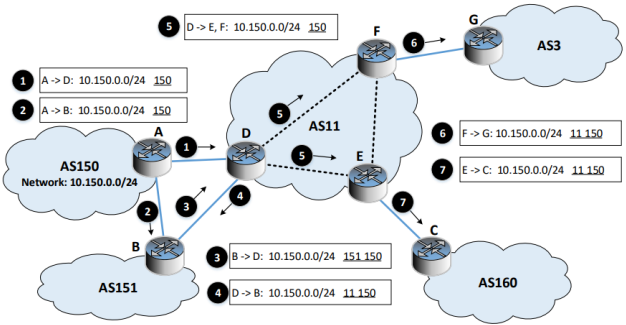
Forwarding table: A data structure that maps each destination address (or set of destination addresses) to the outgoing interface over which a packet destined to this address must be forwarded to reach its final destination. The router consults its forwarding table to forward each packet that it handles.

Step-by-step Example: a small Internet[2]



[2]Images by Wenliang Du, *Internet Security: A Hands-on Approach 3rd Edition*

AS 150

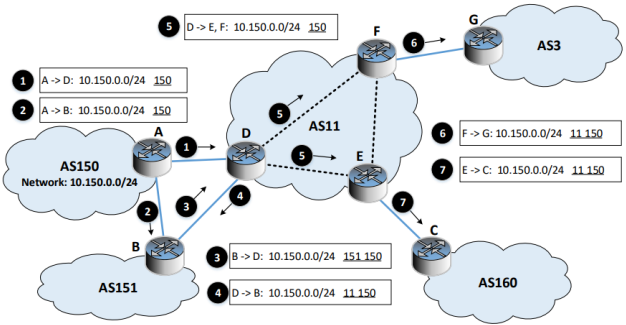


AS 150 owns network 10.150.0.0/24, a class C network. AS 150 is a multi-homed stub AS.

1. Router A in AS 150 announces 10.150.0.0/24 to router D in AS11
2. Router A in AS 150 announces 10.150.0.0/24 to router B in AS151

BGP announcements contain the prefix, and the path to reach the prefix. AS150 adds the path to itself: *150*.

AS 150



Both AS151 and AS11 offer transit to AS150. So they will announce the prefix from AS150

3. Router B announces 10.150.0.0/24 to D, adding 151 to the path
4. Router D announces 10.150.0.0/24 to B, adding 11 to the path

B and D do not announce 10.150.0.0/24 back to A.

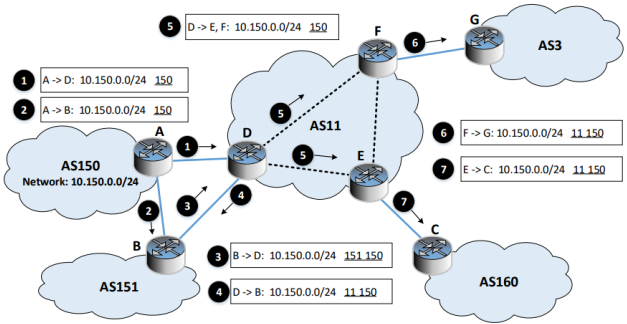
Route Selection

- At the end of step 4 both D and B have two different routes to reach A
- A direct one, with path made of only one hop
- An indirect one, with path made of two hops
- The choice of which one to use depends on many factors, among which, the path length.
- Let us assume the rule the routers apply is "use the shortest path".
- Now hosts in AS11 and AS151 know how to reach AS150, through a one-hop direct path.

Split Horizon

- An AS does not announce a prefix to the router from which it receives the announcement
- This eliminates the single link loop creation problem
- Moreover, BGP also includes in the announcements the full path to reach the destination
- This is a key difference from a pure distance vector protocol.
- Before announcing a prefix, the AS checks that it is already in the path: **this prevents loop longer than one link to be created.**

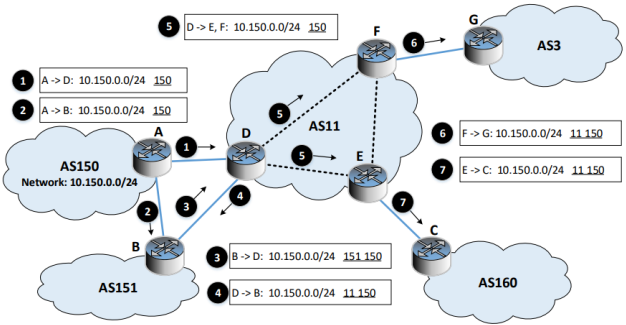
I-BGP



BGP can be used even inside a single AS, we refer to it as IBGP (Internal-BGP)

5. Router D, E, F are part of the same AS, and thus D announces 10.150.0.0/24 to E and F **but does not add anything to the path**

AS3 and AS160



AS3 and AS160 are single-homed stub ASes

6. Router F announces 10.150.0.0/24 to G prepending 11 to the path
7. Router E announces 10.150.0.0/24 to C prepending 11 to the path

Now C learned that to reach 10.150.0.0/24 it has to send traffic to router E, and go across AS 11 and then reach AS 150. Same for G.

Route is propagated

- At the end of this process, the prefix 10.150.0.0/24 has been propagated to the whole (small) Internet.
- Every router now has a route to 10.150.0.0/24
- Every other AS will announce its prefix(es) and with the same procedure, every router in the Internet will have a correct route to any prefix.

BGP is a Path Vector Protocol

BGP uses the principles of DV routing, with three main differences

- first, it exports not only the distance to the destination, but the whole *AS path*, it is called a *Path Vector* protocol
- second, it does not regularly send updates. It sends updates only when something changes, or when a neighbor explicitly asks for an update.
- third, BGP `UPDATE` messages contain information about some prefixes only, not the whole routing table. An `UPDATE` is sent about a prefix if the route is new, one of its attributes (for instance the AS Path) has changed or the route became unreachable and must be withdrawn.

Path Withdrawal

- If, for some reason AS150 ceases to host the 10.150.0.0/24 prefix, it will send a withdrawal message
- The same sequence of routers that announced it at first, will also withdraw their announcement
- This will make the prefix disappear from the Internet

Link Failure

- Let's say that link between A and D breaks.
- Then D will withdraw 10.150.0.0/24 to B, E, F.
- However, D has an alternative path, passing through AS151
- So D will announce a new path to 10.150.0.0/24 passing through 11, 151, 150.
- This way, AS 160 and AS3 will still be able to reach AS150.

Let's look at the BGP Emulator from Seedlab

- Seedlab is a network security testbed that I use for the Network Security exam
- There is a very nice Network Emulator that supports a little Internet in a Box, based on a VirtualBox Virtual Machine.
- Let's run the BGP emulator in the VM and inspect the network.
- The emulator creates a network made of hosts and routers, routers run the `bird` software, that is an open source daemon implementing BGP
- We need two commands:
  - `birdc show route all`: shows the BGP routing table
  - `ip r`: shows the kernel routing table
- There is a key difference between these two

Let's look at the BGP Emulator

- `birdc show route all`: shows what `bird` knows about routing paths
- `bird` in a router receives `update` messages from its peers
- These messages help to construct the knowledge of the state of the network, and discover existing paths to some destination
- However, out of all the possible paths, only one will be chosen
- That path is converted in a route and installed in the kernel routing table, which you can read with `ip r`
- Note that a linux-based router is a software router, so it does not have a real *forwarding table*, meaning something separate from a routing table configured by software.

birdc show protocols

`birdc show protocols` will show all the protocols `bird` supports, and their state.

```
1  bird> show protocols
2  Name       Proto      Table      State  Since         Info
3  device1    Device     ---        up     12:39:18.085
4  kernel1    Kernel     master4    up     12:39:18.085
5  local_nets Direct     ---        up     12:39:18.085
6  pipe1      Pipe       ---        up     12:39:18.085  t_bgp <=> master4
7  pipe2      Pipe       ---        up     12:39:18.085  t_direct <=> t_bgp
8  u_as2      BGP        ---        up     12:39:20.422  Established
9  u_as4      BGP        ---        up     12:39:20.389  Established
10 p_as156    BGP        ---        up     12:39:21.146  Established
11 ospf1      OSPF       t_ospf     up     12:39:18.085  Alone
12 pipe3      Pipe       ---        up     12:39:18.085  t_ospf <=> master4
```

Note line 7-9 that show that `bird` has set up 3 peering sessions with other BGP peers, from which it will learn new routes, and one OSPF session too.

BGP

↳2.1 BGP Protocol Details

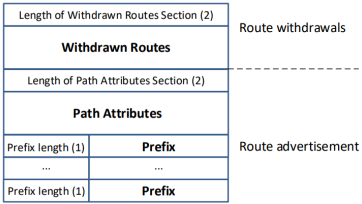
BGP connections

- When two BGP routers accept to share information, they need to be configured to do so
- Each router will accept a TCP connection on port 179 from one *peer*
- Note that in BGP terminology, a BGP *peer* is just a neighbor in the BGP graph, whatever the commercial relation they use (peering or transit)

BGP Connections

- BGP sends 5 kinds of messages:
  - Open message: for establishing BGP connections
  - Update message: for transferring routing information between BGP peers.
  - Keepalive message: for checking whether the peers are still reachable.
  - Notification message: for notifying BGP peers of errors.
  - Route-refresh message: a message type to support the Route Refresh Capability.

Focus on the Update



There are two sections:

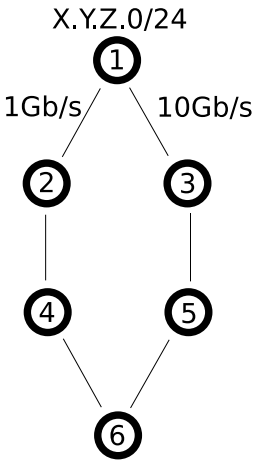
- Routes (prefixes) that one router wants to withdrawal
- Routes (prefixes) that one router wants to announce

There is no cryptography. No authentication, no secrecy...

BGP

↳2.2 Path Prepending

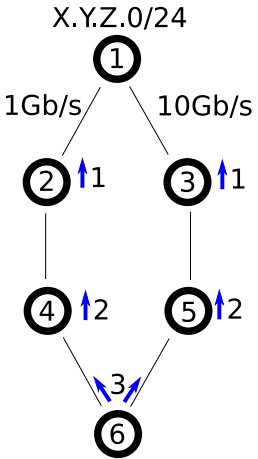
BGP Path Prepending[3]



- AS1 is a multi-homed stub AS, it has two connections:
  - a 10Gb/s one, to be used in normal conditions
  - a 1GMb/s one, to be used only as a backup

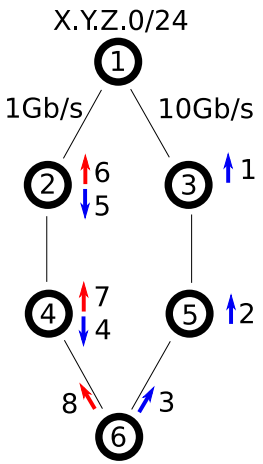
[3]See RFC draft
`https://datatracker.ietf.org/doc/draft-ietf-grow-as-path-prepending/13/`

BGP Path Prepending



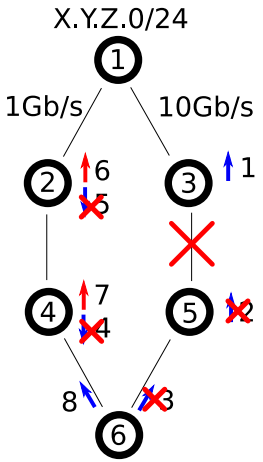
- AS1 is a multi-homed stub AS, it has two connections:
  - a 10Gb/s one, to be used in normal conditions
  - a 1Gb/s one, to be used only as a backup
- However in normal conditions this is the choice of each AS, 2 and 4 will surely pick the left path, and 6 with 50% probability
- How does AS 1 convince the others to use the right link?

BGP Path Prepending



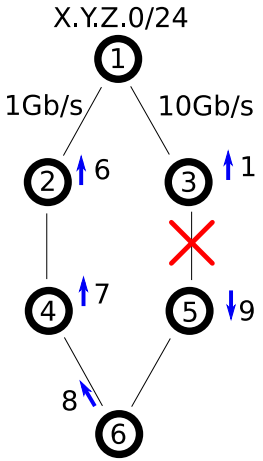
- It announces prefix `X.Y.Z.0/24` to AS 3 with path `1`
- It announces prefix `X.Y.Z.0/24` to AS 2 with path `1,1,1,1,1,1` (same, repeated 6 times)
- AS6 receives two different updates for network `X.Y.Z.0/24`:
  - One with path 5,3,1
  - One with path 4,2,1,1,1,1,1
- Then AS6 will use and propagate the first (shorter) one, as long as it is available
- AS 4 and 2 will choose the path that goes South, as it is shorter

BGP Path Prepending



- If the link between AS5 and AS3 breaks, AS5 will withdraw the prefix `X.Y.Z.0/24`
- The withdrawal messages are propagated and the "blue" paths disappear from all the ASes
- However, AS 2, 4 and 6 have an alternative one.

BGP Path Prepending



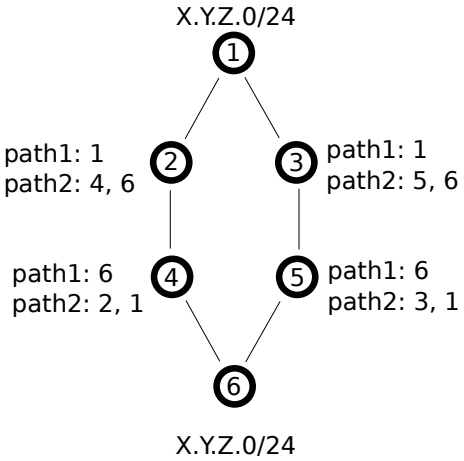
- AS6 announces the backup path
- Finally AS5 also receives the update and converges to the new path

Path prepending is used by an AS to force the choice of one of its links. It is used to make traffic engineering choices, not only for backups, so an administrator may export different networks on different links with different paths, so to balance the load on every link.

BGP

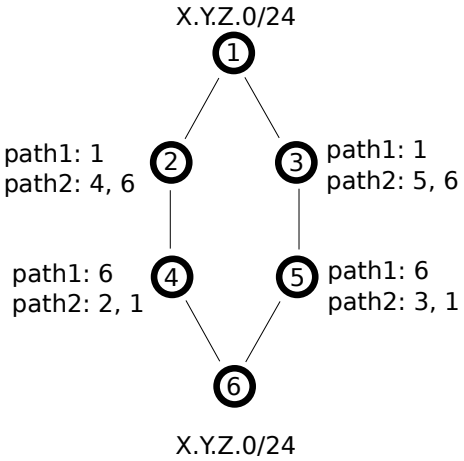
↳2.3 BGP Anycast

BGP Anycast



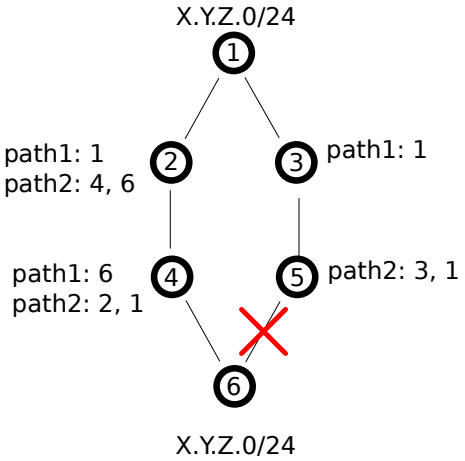
- In this network both AS 1 and 6 announce the **same** prefix
- However, the other ASes in the network will divide between those that route towards 1 (AS 2 and 3) and those that route towards 6 (AS 4 and 5)

BGP Anycast



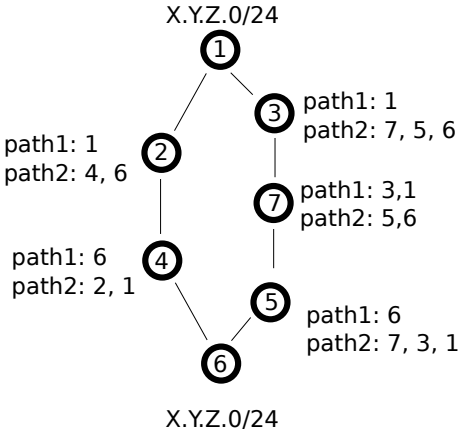
- On the Internet, the same prefix can be announced by more than one AS
- The other ASes will route towards the closest one
- This literally mean that for some application, more than one Internet host have the same IP address

BGP Anycast



- Anycast (route towards the closest IP) makes it possible to have redundant services
- The same IP X.Y.Z.1 is hosted in two servers
- If the best path fails, there is a backup one.

BGP Anycast



- However, what is the best path for AS 7?
- It is likely that AS7 may change its routing decision with time.
- So AS7 may send a packet to X.Y.Z.1 hosted at AS1 and a later time, send it to AS6

BGP Anycast

- For this reason, Anycast is never used with a stateful service, like a TCP connection
- It is used for services that have an request-response style
- The best known one is DNS

BGP Anycast for Root servers

- We have mentioned that on top of the DNS hierarchy, there are DNS root servers
- There are 13 of them, named from A to M
- However, each of them belongs to a network that is announced by tens of ASes around the world
- You can check the Root servers at: `https://root-servers.org/`
- When your browser requests a domain to, for instance, server A, BGP will deliver the request to the closest copy of DNS server

BGP Anycast for Root servers

- Servers are periodically updated and kept in sync, so their database is the same
- Since DNS is a request/response protocol, it does not matter if BGP changes its routing decision from time to time
- Every DNS request is independent from the previous one, and so the system works
- If you want to implement this with TCP it is more complicated, because you need to keep the state synchronized between N potential servers.