Border Gateway Protocol v4

Why Do We Need an EGP?

- Exterior Gateway protocol (EGP)
- Scaling to large network
  - Hierarchy
  - Limit scope of failure
- Define administrative boundary
- Policy
  - Control reacheability to prefixes

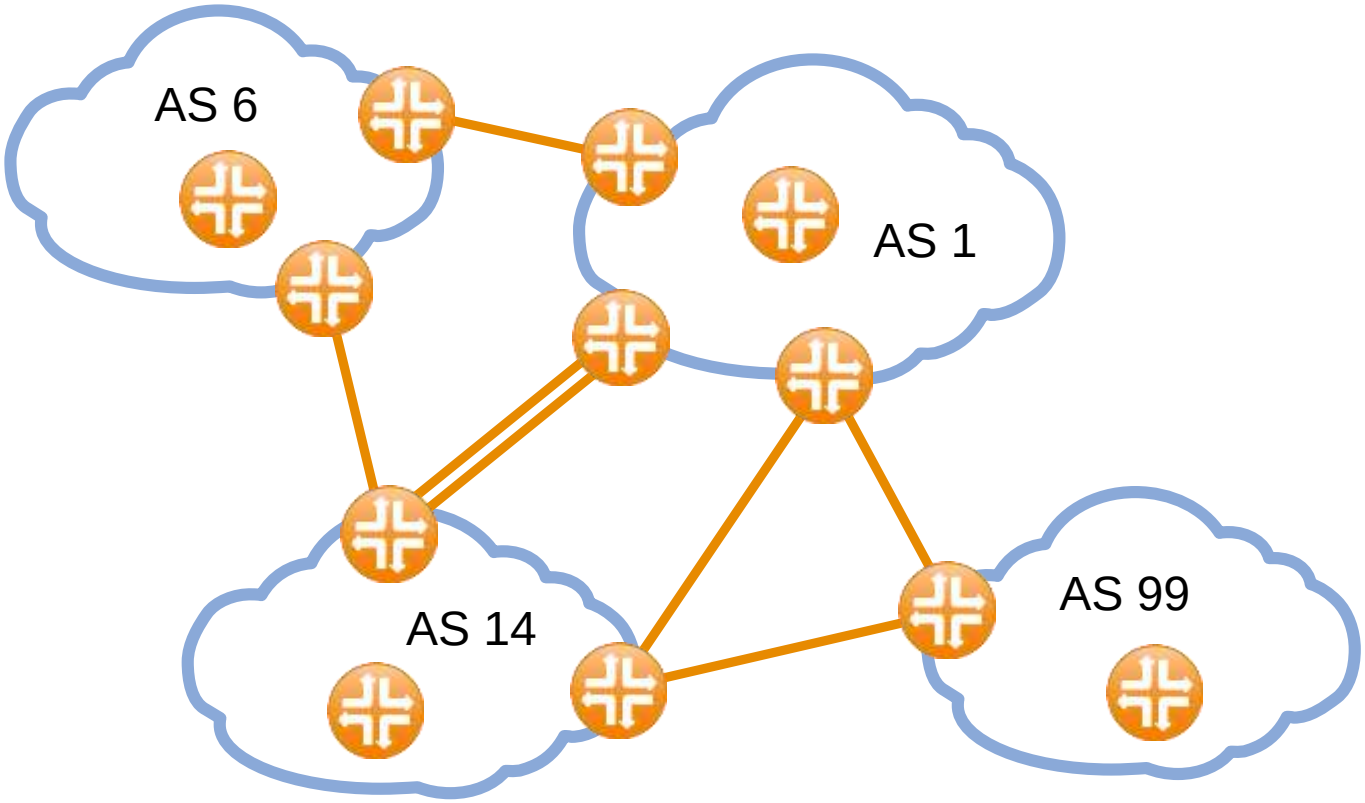
Benefits of BGP?

1) Internet service advantages
   - Scalability: BGP was designed to be a robust, conservative routing protocol able to carry hundreds of thousands of IP prefixes
   - Flexibility: The large number of attributes can be attached to a route, complex route selection rules and BGP-specific filtering mechanisms available
2) Increasing core network stability
   - You should never carry your customers' routes in your core (IGP) routing protocol, as customer's internal problems could quickly affect the stability of your own network
3) Increasing core network security
   - Internal routes does not need to be propagated to customers
   - Private IP address space in the core network
4) VPN services
   - An extension to BGP, called Multi Protocol (MP-BGP), together with MPLS technology supports a variety of customer virtual private network services

BGP Overview

- Exterior Gateway protocol (EGP)
- BGP4 v4 is the protocol used on the Internet to exchange routing information between ISP providers, and to propagate external routing information through networks
- Each autonomous network is called an **Autonomous System**
  - One AS means typically one running IGP, collection of routers under the control of one entity
- Each AS has AS Number (ASN) which is injected to the route information
- Relies on ASNs to construct AS paths (IGP relies on IP addresses)
- Currently in version 4 (1992)
- Uses TCP on port 179 to send routing messages
- BGP is a **distance vector** protocol

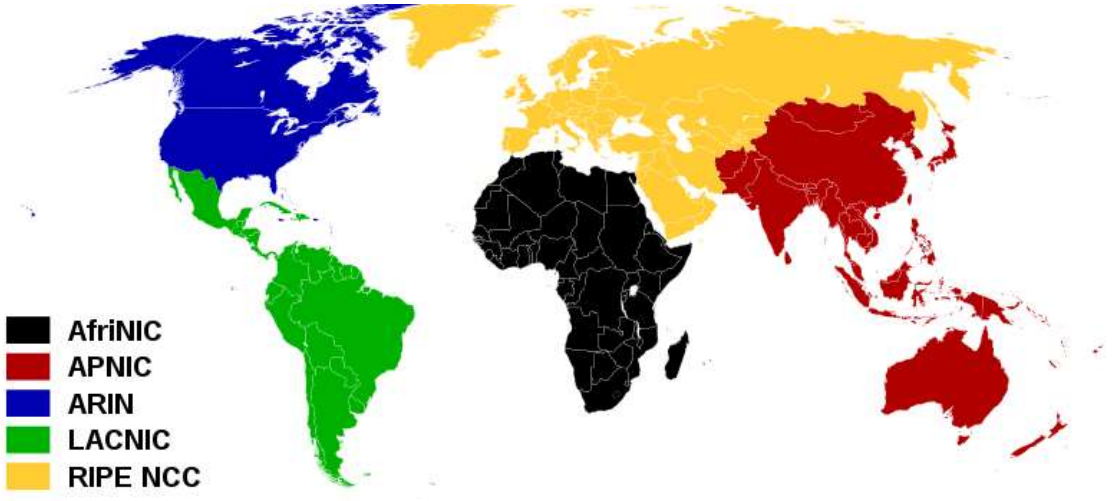
Autonomous Systems



Regional Internet Registry (RIR)

- Registrácia a správa IP adries a ASN je zabezpečená regionálnymi organizáciami, ktoré sa nazývajú *Regional Internet Registries (RIRs)*
- IANA prideluje adresne bloky regionalnym RIR organizáciam (Internet Assigned Numbers Authority)

  - Keď RIR obdrží blok adries od IANA, ktorá funguje ako najvyšia registračná organizácia, distribuuje RIR tieto svojim zákazníkom - *Local Internet Registries (LIRs).*
  - Adresy nie sú spoplatňované, avšak LIR platí ročné členské poplatky
  - RIR spravuje: IPv4 adresné bloky, IPv6 adresné bloky, Čísla autonómnych systémov (ASN)
  - 5 RIR organizácií



IANA and IPv4 address space status

- Currently IANA is a department of ICANN (from 1998), a nonprofit private American corporation, which oversees global IP address allocation, autonomous system number allocation, root zone management in the Domain Name System (DNS), media types, and other Internet Protocol-related symbols and numbers

Jon Postel

- Editor of the Request for Comment (RFC)
- Former Director of the names and number assignment clearinghouse, the Internet Assigned Numbers Authority (IANA)

http://www.cidr-report.org/as2.0/

https://www.ripe.net/internet-coordination/news/announcements/ripe-ncc-begins-to-allocate-ipv4-address-space-from-the-last-8

http://www.potaroo.net/tools/ipv4/

https://www.ripe.net

www.six.sk

The BGP Route

- The **BGP route** is a “container” of attributes
- The section of IP address space is composed of
  - the network address **Prefix** attribute of the route and
  - the **Length** of the prefix
  - Example 192.168.1.0/24
- As a BGP route travels from AS to AS, the ASN of each AS is stamped on it when it leaves that AS. Called the **AS_PATH** attribute
- In addition to the prefix, the as-path, and the **Next-Hop**, the BGP route has many other attributes

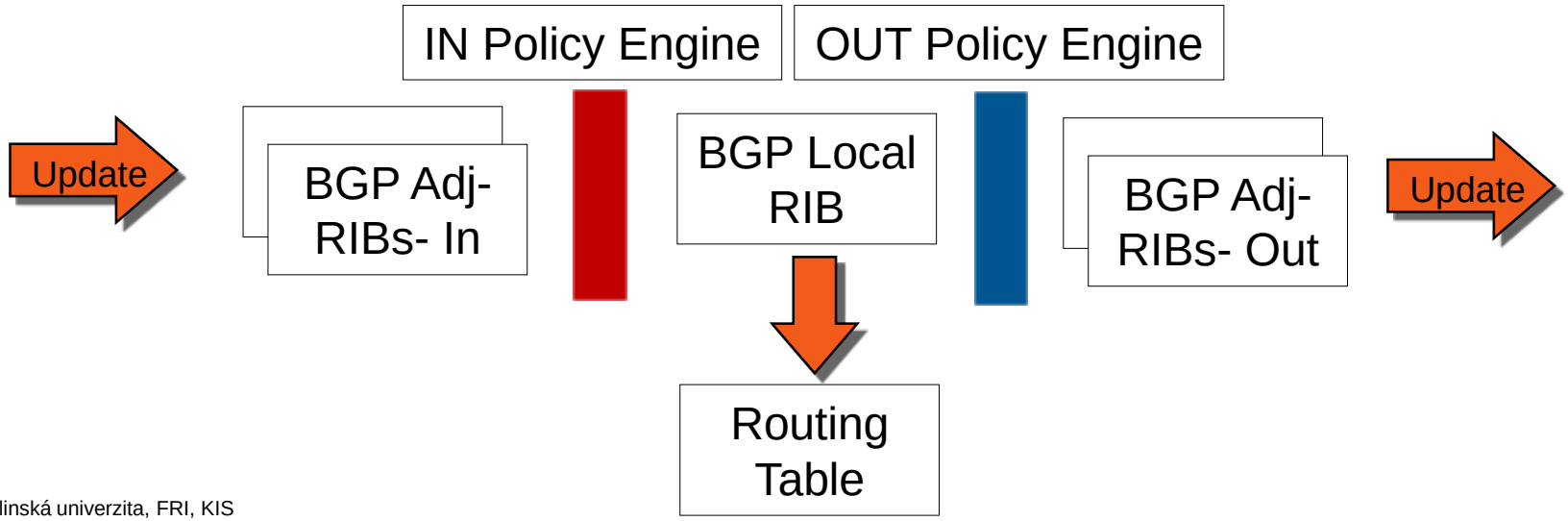
BGP Operations

- Two BGP routers exchanging information on a connection are called **peers**
- Initially, BGP peers exchange the entire BGP routing table
- A BGP router holds the current version of the entire BGP routing tables of all of its peers for the duration of the connection
- Subsequently, only incremental **Updates** are sent as the routing tables change
- **Keepalive** messages are sent periodically to ensure that the connection between the BGP peers is alive
- **Notification** messages are sent in response to errors or special conditions

BGP Operations

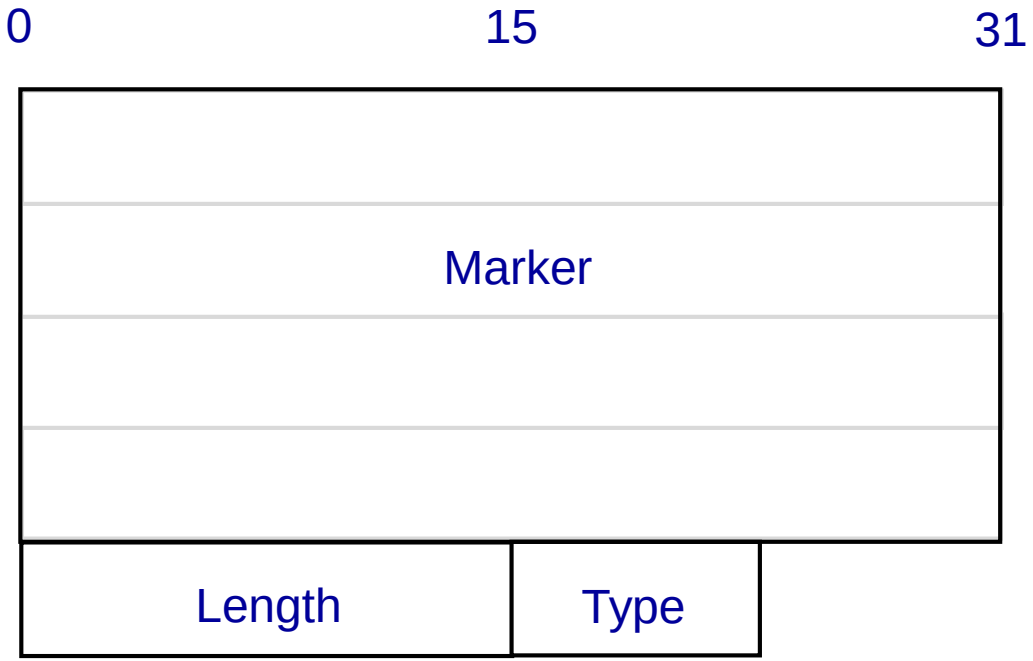
- The BGP Routes are stored in the BGP Routing Information Bases (RIBs)

- A RIB within a BGP router consists of three distinct parts:
  - **Adj-RIBs-In**: contains unprocessed routing information that has been advertised to the local BGP router by its peers
  - **Loc-RIB**: contains the routes that have been selected by the local BGP router's Decision Process
  - **Adj-RIBs-Out**: organizes the routes for advertisement to specific peers by means of the local speaker‘s UPDATE messages



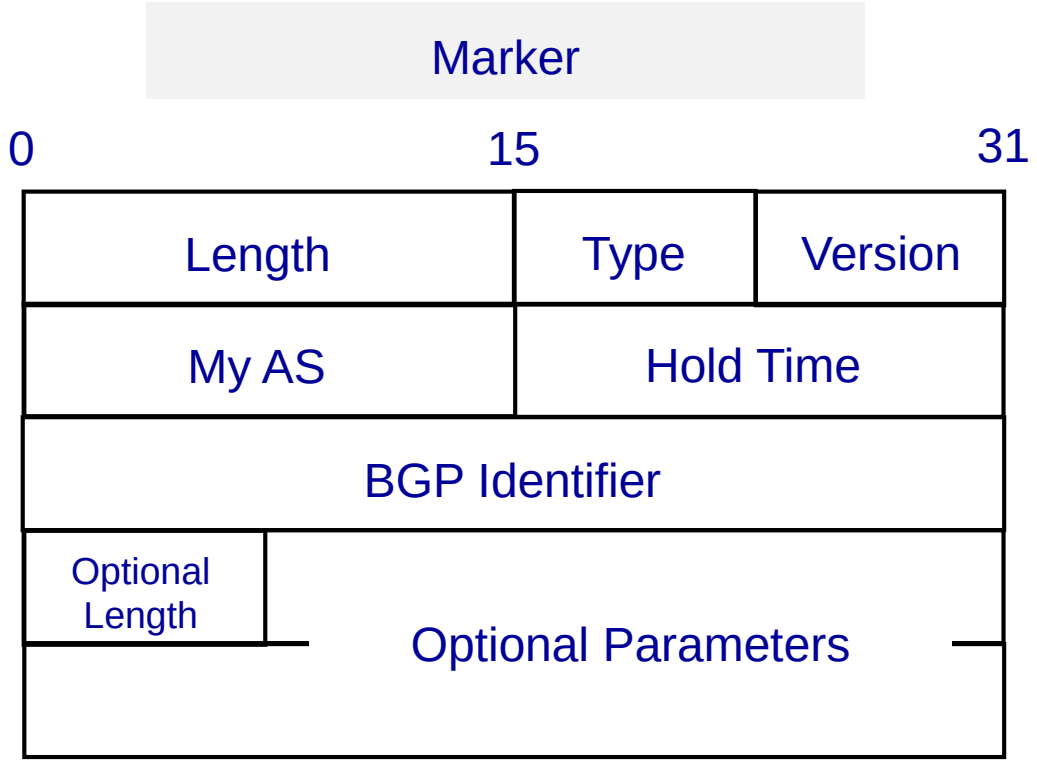
BGP Messages – Common Header Format

- **Marker** – all 1's, used to separate multiple messages in a single TCP stream
- **Length** - indicates the total length of the message in octets, including the BGP header
- **Type** - indicates the type of the message
  - 1- Open
  - 2- Update
  - 3- Notification
  - 4- Keepalive



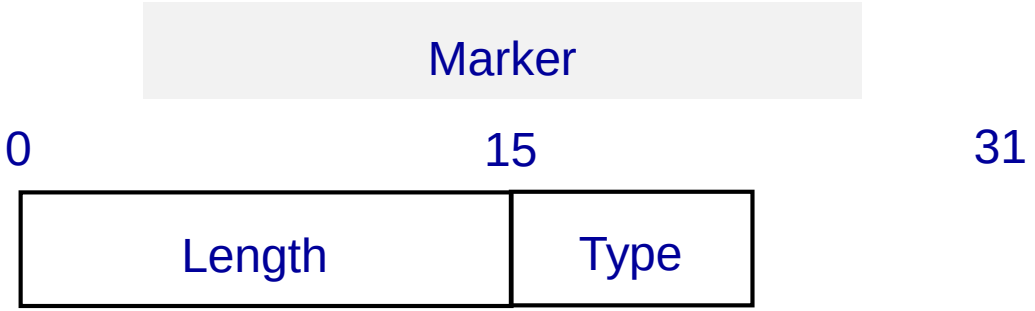
OPEN Message

- **Hold time -** the number of seconds between the transmission of successive KEEPALIVE/UDATE messages, indicates to the peer the length of time that it should consider the sender valid
  - 180 sec default
- **BPG identifier -** the sending BGP router
- **Optional parameter** - a list of optional parameters, encoded in TLV structure (Authentication, ..), RFC 5492



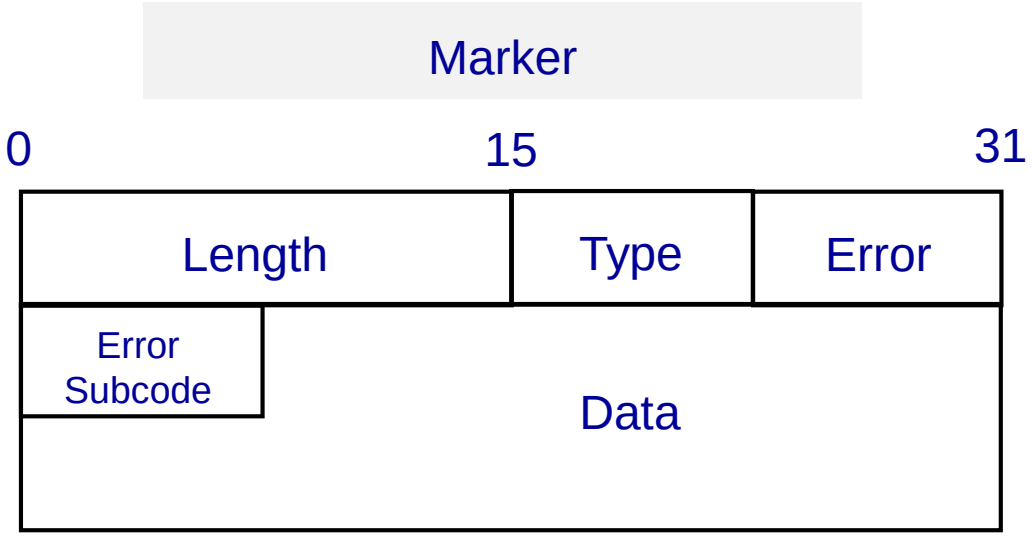
KEEPALIVE Message

- BGP does not use any TCP-based, keep-alive mechanism to determine if peers are reachable
- Time between KEEPALIVE messages typically one third of the Hold Time, not more than one per second
- If the hold time is zero, then KEEPALIVE messages will not be sent



NOTIFICATION Message

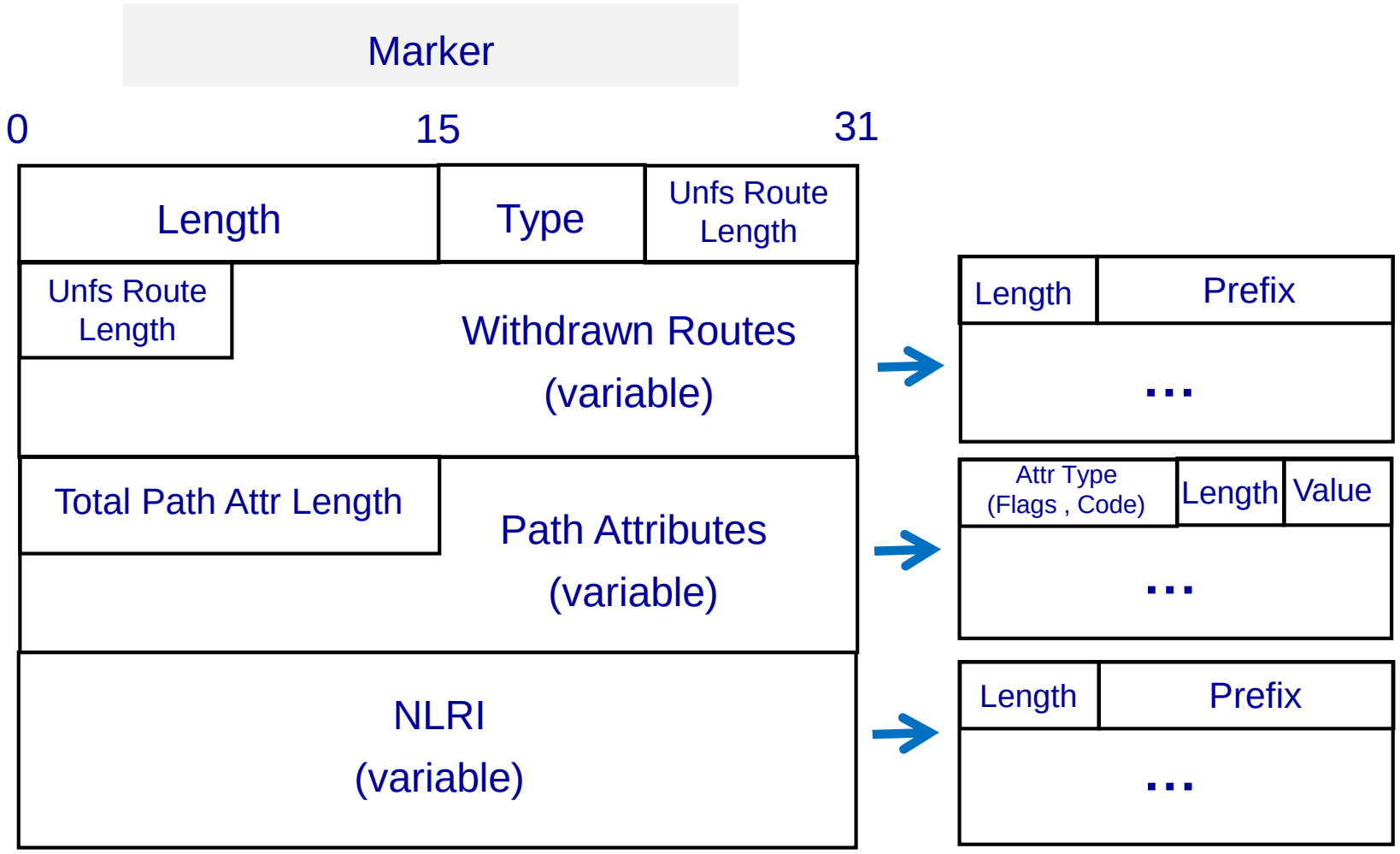
- Sent when an error condition is detected, the session is immediately closed
- **Error code** - the type of error condition
- **Error subcode** - specific information about the nature of the error
- **Data** - the reason for the notification



UPDATE Message

- **Unfeasible routes length** - the total length of the withdrawn routes field in octets
- **Withdrawn routes** - a list of IP address prefixes for the routes that need to be withdrawn from BGP routing tables
- **Total path attribute length** - the total length of the Path Attributes field in octets
- **Path attributes** - a variable length sequence of path attributes
- **NLRI** – Network Layer Reachability Information, a list of IP prefixes

UPDATE Message



Example BGP Message

```
⊞ Ethernet II, Src: c2:01:14:bc:00:00 (c2:01:14:bc:00:00), Dst: c2:02:1e:d8:00:00 (c2:02:1e:d8:00:00)
⊞ Internet Protocol Version 4, Src: 10.39.1.1 (10.39.1.1), Dst: 10.39.1.2 (10.39.1.2)
⊞ Transmission Control Protocol, Src Port: bgp (179), Dst Port: 63912 (63912), Seq: 65, Ack: 84, Len: 89
⊟ Border Gateway Protocol
  ⊟ UPDATE Message
      Marker: 16 bytes
      Length: 51 bytes
      Type: UPDATE Message (2)
      Unfeasible routes length: 0 bytes
      Total path attribute length: 20 bytes
    ⊟ Path attributes
      ⊞ ORIGIN: INCOMPLETE (4 bytes)
      ⊟ AS_PATH: 1 10 (9 bytes)
        ⊞ Flags: 0x40 (Well-known, Transitive, Complete)
          Type code: AS_PATH (2)
          Length: 6 bytes
        ⊟ AS path: 1 10
          ⊟ AS path segment: 1 10
              Path segment type: AS_SEQUENCE (2)
              Path segment length: 2 ASs
              Path segment value: 1 10
      ⊟ NEXT_HOP: 10.39.1.1 (7 bytes)
        ⊞ Flags: 0x40 (Well-known, Transitive, Complete)
          Type code: NEXT_HOP (3)
          Length: 4 bytes
          Next hop: 10.39.1.1 (10.39.1.1)
    ⊟ Network layer reachability information: 8 bytes
      ⊟ 192.168.222.0/24
          NLRI prefix length: 24
          NLRI prefix: 192.168.222.0 (192.168.222.0)
      ⊞ 192.168.223.0/24
⊟ Border Gateway Protocol
  ⊞ KEEPALIVE Message
```

BGP Attribute Types

The path attributes fall in four categories:

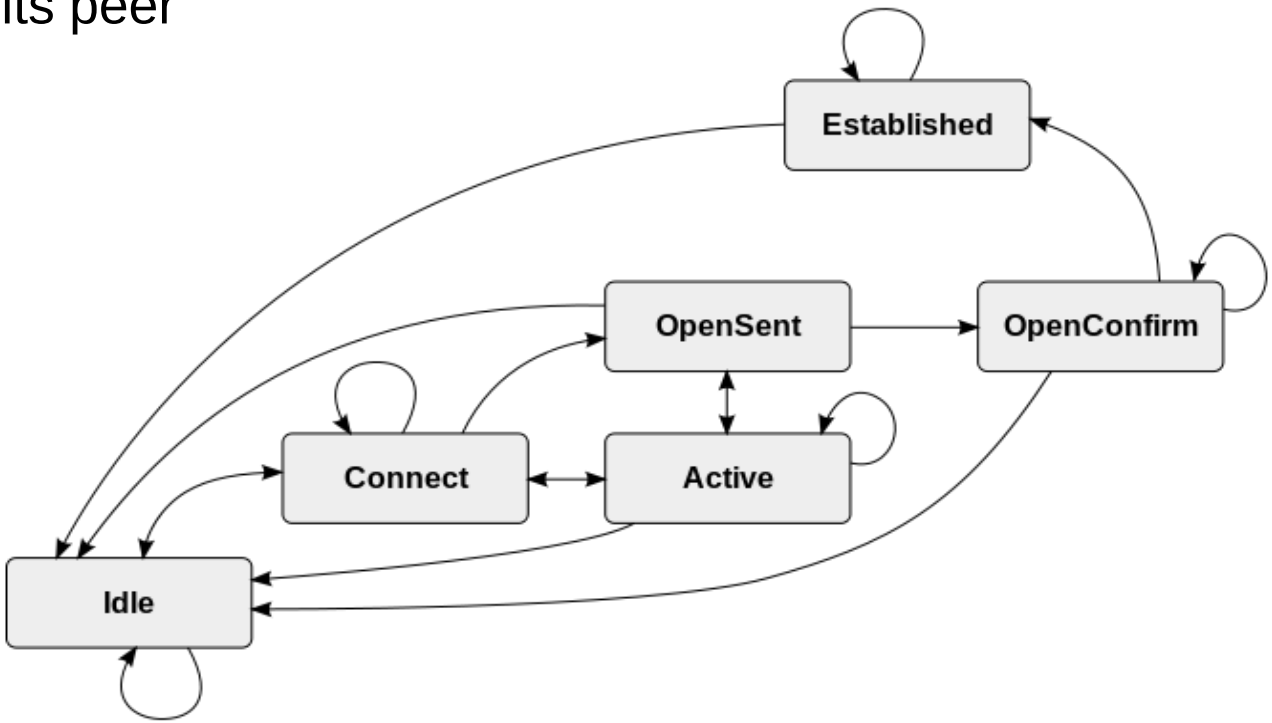
- **Well-known mandatory**
  - must appear in all BGP updates
  - AS-Path, Next Hop, Origin
- **Well-known discretionary (optional)**
  - does not have to be present in all BGP updates
  - Local Preference, Atomic Aggregate
- **Optional transitive**
  - a BGP process should accept the path in which it is included, even if it doesn't support the attribute, and it should pass the path on to its peers
  - Aggregator, Community
- **Optional non transitive**
  - a BGP process that does not recognize the attribute can ignore the Update in which it is included and not advertise the path to its other peers
  - Multi Exit Discriminator (MED)

BGP Attribute Type Codes

- Type code 1 – Origin
- Type code 2 – AS-path
- Type code 3 – Next-hop
- Type code 4 – MED
- Type code 5 – Local preference
- Type code 6 – Atomic aggregate
- Type code 7 – Aggregator
- Type code 8 – Community
- Type code 9 – Originator-ID
- Type code 10 – Cluster list

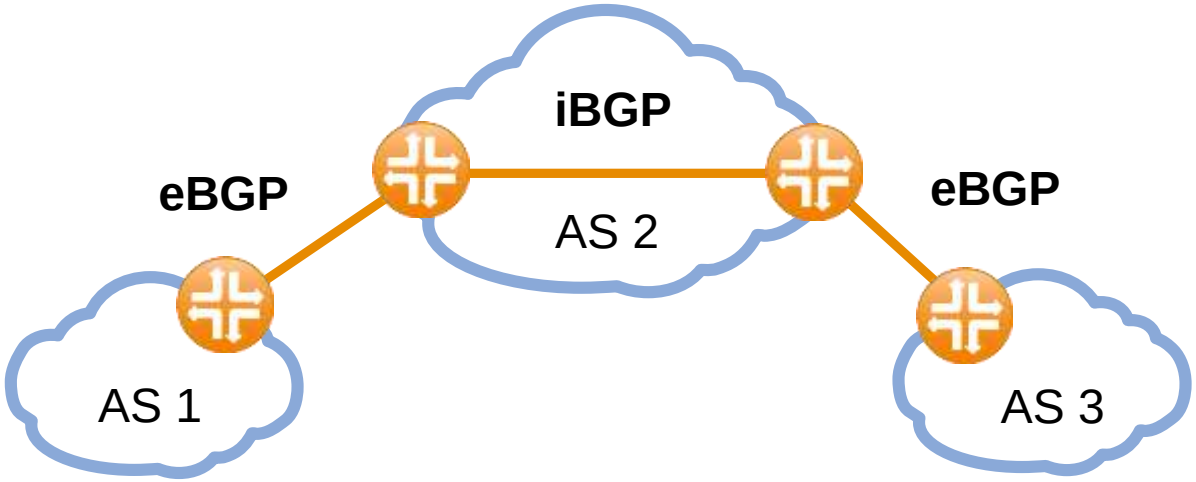
BGP Finite State Machine

- **Idle** state - In this state BGP refuses all incoming TCP BGP connections. No local resources are allocated to BGP peer
- **Connect** state - In this state BGP is waiting for the TCP connection to be completed
- **Active** state - It was unable to establish a successful TCP connection. In this state BGP is trying to acquire a peer by reinitiating a TCP connection
- **OpenSent** state – TCP connection is up. In this state BGP waits for an OPEN message from its peer
- **OpenConfirm** state - In this state BGP waits for a KEEPALV or NOTIF message
- **Established** state - In the Established state BGP can exchange UPDATE, NOTIF, and KEEPALV messages with its peer



iBGP and eBGP

- BGP can also be used within an AS. BGP connections inside an AS are called **internal BGP** (iBGP), and BGP connections between different ASes are called **external BGP** (eBGP)
- The purpose of iBGP is to ensure that network reachability information is consistent among multiple BGP routers in the same AS



iBGP versus eBGP Comparison

- iBGP and eBGP are the same protocol
  - just different rules

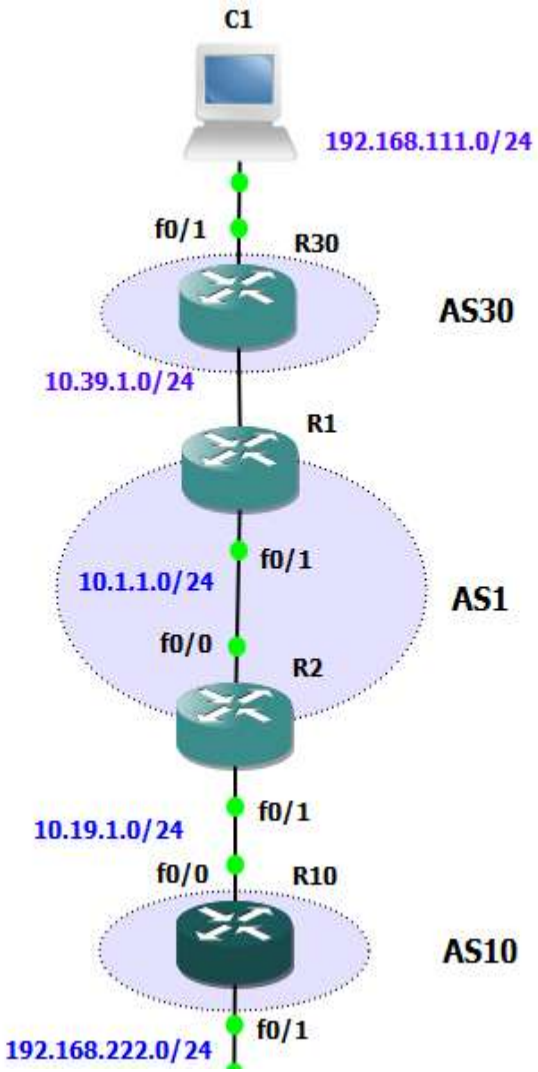
- Rules are intuitive
  - eBGP advertises everything to everyone by default
  - iBGP router does NOT advertise “3rd-party iBGP routes” to other iBGP peers. Why?
  - No way to do loop detection via AS-PATH with iBGP, so this solves it

- eBGP - Should be directly connected, do not run an IGP between eBGP peers in different ASes

- i-BGP - Each iBGP speaker must peer with every other iBGP speaker in the AS (full mesh), not required to be directly connected

eBGP Configuration Example

```
!
router bgp 30
 bgp router-id 10.1.255.30
 neighbor 10.39.1.1 remote-as 1
!

--

R30#sh ip bgp summary
BGP router identifier 10.1.255.30, local AS number 30
BGP table version is 24, main routing table version 24
3 network entries using 303 bytes of memory
3 path entries using 144 bytes of memory
2 BGP path attribute entries using 120 bytes of memory
1 BGP AS-PATH entries using 24 bytes of memory
0 BGP route-map cache entries using 0 bytes of memory
0 BGP filter-list cache entries using 0 bytes of memory
BGP using 591 total bytes of memory
BGP activity 10/7 prefixes, 13/10 paths, scan interval 60 secs

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxR
10.39.1.1       4     1     138     135       24    0    0 01:23:27        2
R30#

--

R30#sh ip bgp ipv4 unicast neighbors 10.39.1.1 routes
BGP table version is 24, local router ID is 10.1.255.30
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal
              r RIB-failure, S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*> 192.168.222.0    10.39.1.1                              0 1 10 ?
*> 192.168.223.0    10.39.1.1                              0 1 10 ?

Total number of prefixes 2
R30#
```



Next-Hop-Self

The next-hop-self command will allow us to force BGP to use a specified IP address as the next hop

```
R1#
router bgp 1
 bgp router-id 10.1.255.1
 neighbor 10.1.255.2 remote-as 1
 neighbor 10.1.255.2 update-source Loopback0
 neighbor 10.1.255.2 next-hop-self
 neighbor 10.39.1.2 remote-as 30
!
--
R2#sh ip bgp ipv4 uni nei 10.1.255.1 rout
BGP table version is 21, local router ID is 10.1.255.2
   Network          Next Hop            Metric LocPrf Weight Path
*>i192.168.111.0    10.1.255.1               0    100      0 30 ?
--
R2#sh ip route
B    192.168.111.0/24 [200/0] via 10.1.255.1, 00:00:47
     10.0.0.0/8 is variably subnetted, 4 subnets, 2 masks
C       10.1.1.0/24 is directly connected, FastEthernet0/0
C       10.19.1.0/24 is directly connected, FastEthernet0/1
O       10.1.255.1/32 [110/2] via 10.1.1.1, 02:13:25, FastEthernet0/0
C       10.1.255.2/32 is directly connected, Loopback0
B    192.168.223.0/24 [20/0] via 10.19.1.2, 01:47:51
B    192.168.222.0/24 [20/0] via 10.19.1.2, 01:47:51
R2#
```

```
R1#
router bgp 1
 bgp router-id 10.1.255.1
 neighbor 10.1.255.2 remote-as 1
 neighbor 10.1.255.2 update-source Loopback0
 neighbor 10.39.1.2 remote-as 30
!
--
R2#sh ip bgp ipv4 uni nei 10.1.255.1 route
   Network          Next Hop            Metric LocPrf Weight Path
* i192.168.111.0    10.39.1.2                0    100      0 30 ?
--
R2#sh ip route
     10.0.0.0/8 is variably subnetted, 4 subnets, 2 masks
C       10.1.1.0/24 is directly connected, FastEthernet0/0
C       10.19.1.0/24 is directly connected, FastEthernet0/1
O       10.1.255.1/32 [110/2] via 10.1.1.1, 02:09:21, FastEthernet0/0
C       10.1.255.2/32 is directly connected, Loopback0
B    192.168.223.0/24 [20/0] via 10.19.1.2, 01:43:47
B    192.168.222.0/24 [20/0] via 10.19.1.2, 01:43:47
R2#
--
```

Improves internet reacheability in the network

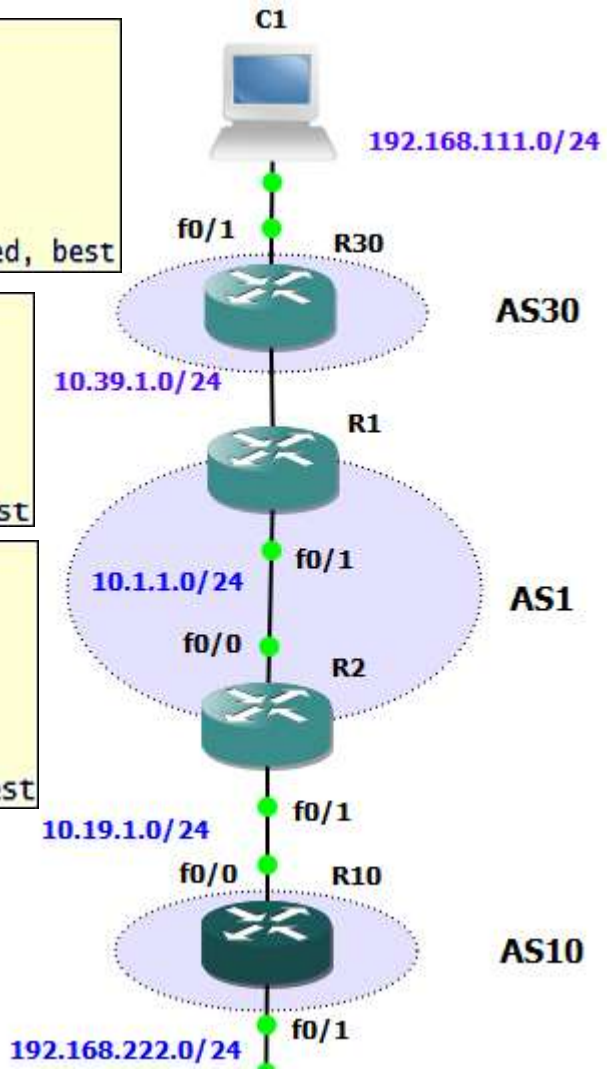
AS-PATH and Next Hop Example

```
R30#sh ip bgp 192.168.111.0
BGP routing table entry for 192.168.111.0/24, version 2
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  Advertised to non peer-group peers:
  10.39.1.1
  Local
    0.0.0.0 from 0.0.0.0 (10.1.255.30)
      Origin incomplete, metric 0, localpref 100, weight 32768, valid, sourced, best
```

```
R1#sh ip bgp 192.168.111.0
BGP routing table entry for 192.168.111.0/24, version 18
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  Advertised to non peer-group peers:
  10.1.255.2
  30
    10.39.1.2 from 10.39.1.2 (10.1.255.30)
      Origin incomplete, metric 0, localpref 100, valid, external, best
```

```
R2#sh ip bgp 192.168.111.0
BGP routing table entry for 192.168.111.0/24, version 19
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  Advertised to non peer-group peers:
  10.19.1.2
  30
    10.1.255.1 (metric 2) from 10.1.255.1 (10.1.255.1)
      Origin incomplete, metric 0, localpref 100, valid, internal, best
```

```
R10#sh ip bgp 192.168.111.0
BGP routing table entry for 192.168.111.0/24, version 18
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  Not advertised to any peer
  1 30
    10.19.1.1 from 10.19.1.1 (10.1.255.2)
      Origin incomplete, localpref 100, valid, external, best
```



Stable iBGP peering

- Unlink iBGP peering from physical topology
- Carry loopback address in IGP

  router ospf <ID>

  passive-interface loopback0
- Unlink peering from physical topology

  router bgp <AS1>

  neighbor <a.b.c.d> remote-as <AS1>

  neighbor <a.b.c.d> **update-source** loopback0

Inserting prefixes into BGP
Originating routes manually

- **Using BGP network command**

  network 192.168.222.0 mask 255.255.255.0

- Interface configuration or static route

  ip route 192.168.222.0 255.255.255.0 fast0/0

- matching route must exist in the routing table before network is announced
- Route-maps allows to redistribute just selected networks to BGP process
- Route-map can modify attributes
- Origin IGP (i)

Inserting prefixes into BGP
Route redistribution

- Redistribution from IGP/EGP process or redistribute static / redistribute connected
- Easier than listing networks manually
- Route-maps allows to redistribute just selected connected networks to BGP process or modify attributes
- ACL or Prefix-List can be used for matching

```
ip prefix-list pl-con-100 seq 5 permit 192.168.222.0/24
ip prefix-list pl-con-100 seq 10 permit 192.168.223.0/24
!
route-map con-100 permit 10
 match ip address prefix-list pl-con-100
!
route-map con-100 deny 20
!
router bgp 10
        redistribute connected route-map con-100
```

- http://www.cisco.com/c/en/us/td/docs/ios/iproute_bgp/command/reference/irg_book/irg_bgp4.html#wp1145787
- Origin: incomplete (?)

Inserting prefixes into BGP Summarization

- Summarization is called aggregation in BGP
- Aggregation creates summary routes (Aggregates) from networks which are already in BGP table
- Individual networks can be announced but typically supressed
- If any route in BGP table is within the range then the summary route is injected

aggregate-address <address> <mask> [summary-only]
ip route <address> <mask> null 0

- Smaller BGP table
- Less route flapping
- Might be a problem with multihomed customers

IGP and BGP Synchronization

- By default BGP synchronizes state with the IGP to avoid black holing
- When router receives an UPDATE tries to check routing table
- BUT injecting all BGP routes (customer ones) inside an IGP is costly and not necessary

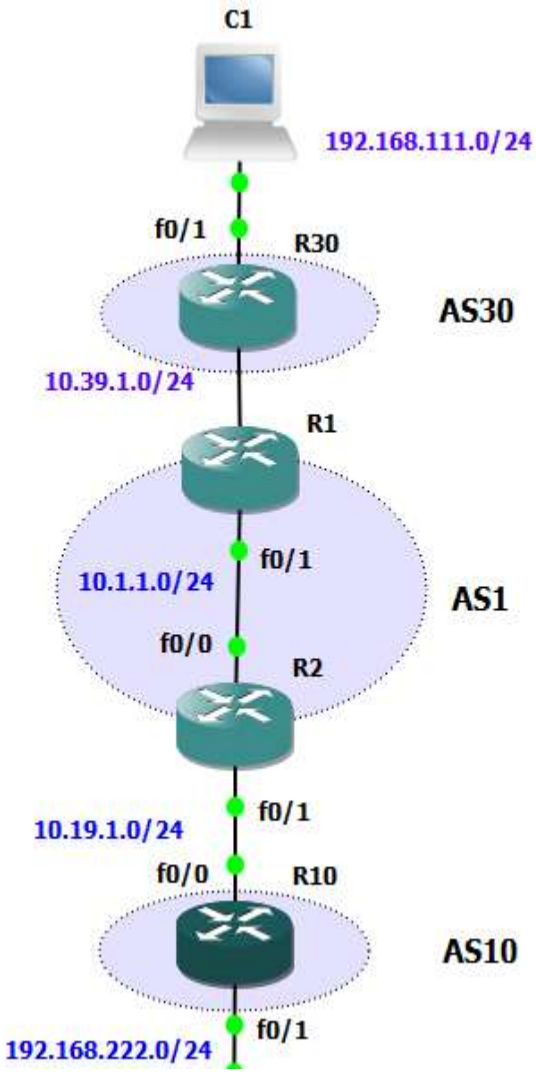
- Typically none or Default route in IGP is enough

```
R2#sh ip bgp
BGP table version is 7, local router ID is 10.1.255.2
Status codes: s suppressed, d damped, h history, * valid, > best, i
              r RIB-failure, S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
* i192.168.111.0    10.1.255.3               0    100      0 30 ?
* i                 10.1.255.1               0    100      0 30 ?
*> 192.168.222.0    10.19.1.2                0             0 10 i
* i                 10.1.255.4               0    100      0 10 i
*> 192.168.223.0    10.19.1.2                0             0 10 i
* i                 10.1.255.4               0    100      0 10 i
R2#sh ip route 102.168.111.0
% Network not in table
R2#sh ip bgp ipv4 uni nei 10.19.1.2 advertised-routes

R2#

router bgp 1
 no synchronization
!


R2#sh ip bgp ipv4 uni nei 10.19.1.2 advertised-routes
   Network          Next Hop            Metric LocPrf Weight Path
*>i192.168.111.0    10.1.255.1               0    100      0 30 ?
R2#
```

Secure Design

- The customer cannot see internal core IP addresses but still can connect to the remote customer/Internet site
- The customer interfaces are not part of core IGP

```
R10#sh ip route
B    192.168.111.0/24 [20/0] via 10.19.1.1, 00:28:48
     10.0.0.0/8 is variably subnetted, 2 subnets, 2 masks
C       10.19.1.0/24 is directly connected, FastEthernet0/0
C       10.1.255.10/32 is directly connected, Loopback0
C    192.168.223.0/24 is directly connected, FastEthernet1/0
C    192.168.222.0/24 is directly connected, FastEthernet0/1
R10#

R10#ping 192.168.111.111 source 192.168.222.222

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 192.168.111.111, timeout is 2 seconds:
Packet sent with a source address of 192.168.222.222
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 56/69/84 ms
R10#
```



Ďakujem za pozornosť

roman dot kaloc at gmail dot com