BGP 3

Autonomous System Number (ASN)

- A unique ASN is allocated to each AS for use in BGP routing
- Until 2007, AS numbers were defined as 16-bit integer RFC1771
  - 65536 assignments (IANA)
  - AS numbers 64512 – 65534 for private purposes
  - ASNs 0 and 65535 are reserved
- RFC4893 introduced 32-bit AS numbers
  - IANA has begun to allocate
  - Written as simple integers
  - or in the form **X.Y**
  - Numbers of the form **0.Y** are exactly the old 16-bit AS numbers
  - **1.Y** numbers and **65535.65535** are reserved
- The number of ASes – 5.000 in 1999, 30k in late 2008, 35k in mid 2010 and 42k in late 2012

Causes of BGP Routing Changes

- Topology changes
  - Equipment going up or down
  - Deployment of new routers or sessions
- BGP session failures
  - Due to equipment failures, maintenance, etc.
  - Due to congestion on the physical path
- Changes in routing policy
  - Reconfiguration of preferences
  - Reconfiguration of route filters
- Persistent protocol oscillation
  - Conflicts between policies in different ASes

BGP Session Failure

- BGP runs over TCP
  - BGP only sends updates when changes occur
  - TCP doesn’t detect lost connectivity on its own

- Detecting a failure
  - Keep-alive: 60 seconds
  - Hold timer: 180 seconds

- Reacting to a failure
  - Discard all routes learned from the neighbor
  - Send new updates for any routes that change

BGP Converges Slowly

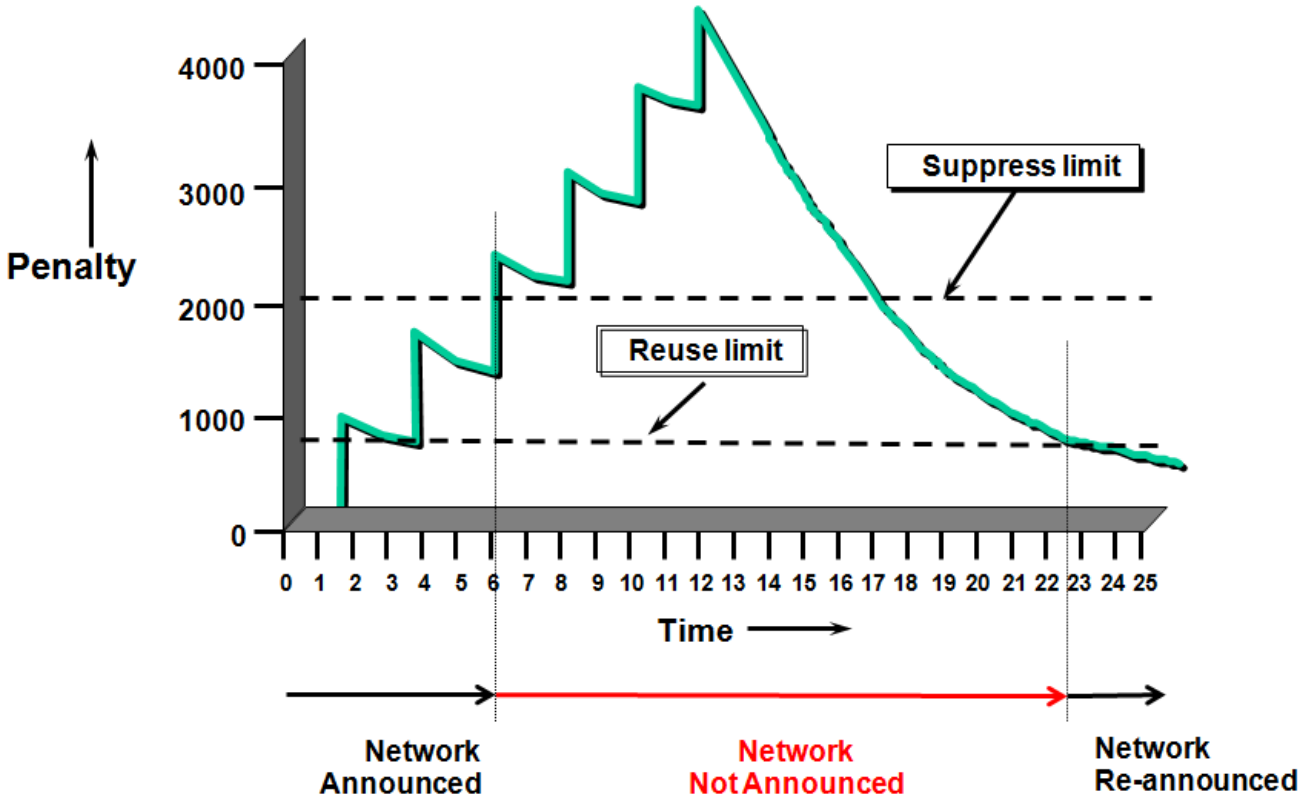
- Can be tens of seconds to tens of minutes
  - Important for interactive applications

- Fortunately in practice:
  - Most popular destinations have very stable BGP routes
  - Most instability lies in a few unpopular destinations

- Minimum Route Advertisement Interval (MRAI)
  - Minimum spacing between announcements/updates
  - For a particular (prefix, peer) pair
  - Provides a rate limit on BGP updates
  - But adds additional delay to the convergence process

Route Flap Damping

- Motivation
  - An equipment goes up and down repeatedly
  - Leading to excessive BGP update messages
- Route Flap Damping
  - Accumulate a penalty for each (prefix, peer) pair with each routing change



iBGP Scaling Issues in Transit or ISP AS

- iBGP requires full mesh between all BGP speakers
  - Large number of TCP sessions
  - Unnecessary duplicate routing control traffic
  - n*(n - 1) / *2*

- Solution

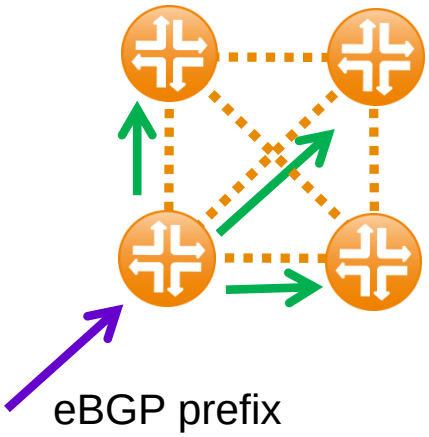
  - **Route Reflectors**
    - Modify iBGP split horizon rules

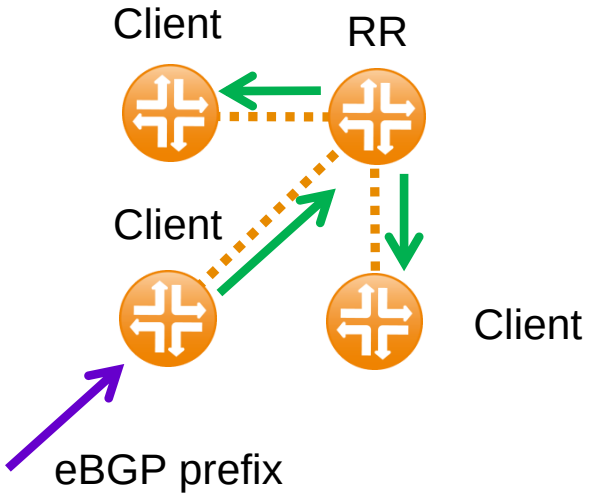
  - **BGP Confederations**
    - Modify iBGP AS path processing

Route Reflector Split Horizon Rules

- Classic BGP
  - iBGP routes are not propagated to the other iBGP peers
  - Full mesh of iBGP peers is required

- RR can propagate iBGP routes to the other iBGP peers (typically to clients)
  - Full mesh of iBGP peers is no longer required
  - RR does not change any attribute, also next hop remains – the client still uses optimum route learned via IGP





Route Reflector Split Horizon Rules

Type of router	Incoming update from	Is forwarded to
Classic	EBGP peer	All peers (IBGP and EBGP)
	IBGP peer	EBGP peers
Route reflector	EBGP peer	All peers (IBGP and EBGP)
	Nonclient IBGP peer	EBGP peers and clients
	Client IBGP peer	All peers but the sender
Client	EBGP peer	All peers (IBGP and EBGP)
	IGBP peer	EBGP peers

Redundant Route Reflectors

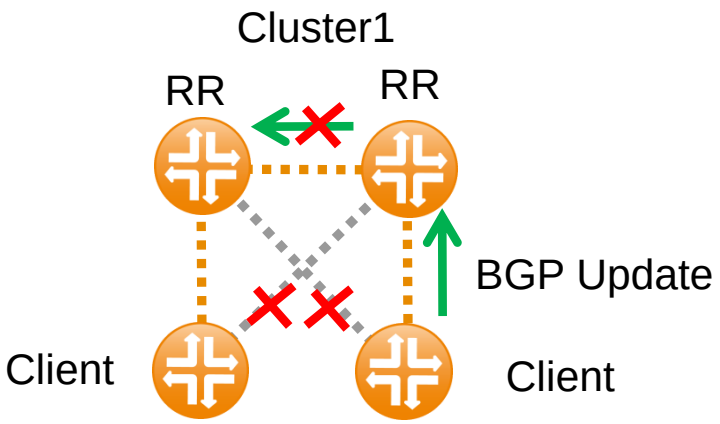
- Client may have any number of eBGP peers but may have iBGP sessions only with RR(s)
- Therefore RR functionality must be redundant in the network
- Each client will receive the same prefix from all RRs
  - When using Weight attribute (Cisco) improperly, the routing loop can occurs
  - Therefore additional BGP attributes are necessary to prevent it

Route Reflectors Cluster

- A group of redundant RRs and their clients form a cluster
- Each cluster must have a unique **cluster-ID** value
- The cluster-ID is configured on RRs only
- Each time a route is reflected, the cluster-ID is added to the **cluster-list** BGP attribute
- If the route, for any reason, is reflected back, the reflector recognize its cluster-id and is not reflected
- The first RR also sets an additional BGP attribute **originator-ID** using client's router-ID
- Cluster-list and Originator-ID are non-transitive optional attributes
- The BGP path selection mechanism is modified
  - Nonreflected routes (no originator-ID) are preferred
  - Shorter cluster-lists are preferred

Redundant Route Reflectors with different Cluster-ID

- Clients do not have sessions with all RRs in a cluster
  - Issue: Clients might not receive all iBGP routes
- Clients have sessions with RRs in several clusters
  - Clients will always receive duplicate copies of the same route
  - Larger BGP table but better convergence
- Often used design in ISP network



Route Reflector Quiz

- Issue
  - RRs not in full BGP mesh
  - Clients do not have sessions with all RRs in a cluster
  - Clients have sessions with RRs in several clusters
  - Clients have iBGP session with other clients

- Results to
  - Clients could receive duplicate updates of the same route
  - Some clusters will not receive all iBGP routes
  - Clients will always receive duplicate copies of the same route
  - Clients will not receive all iBGP routes

Route Reflector Configuration

```
router bgp 65000
 bgp cluster-id 10.1.255.100
 neighbor 10.1.255.1 remote-as 65000
 neighbor 10.1.255.1 update-source Loopback0
 neighbor 10.1.255.1 route-reflector-client
 neighbor 10.1.255.1 send-community
 neighbor 10.1.255.2 remote-as 65000
 neighbor 10.1.255.2 update-source Loopback0
 neighbor 10.1.255.2 route-reflector-client
 neighbor 10.1.255.2 send-community
 neighbor 10.1.255.3 remote-as 65000
 neighbor 10.1.255.3 update-source Loopback0
 neighbor 10.1.255.3 route-reflector-client
 neighbor 10.1.255.3 send-community
```

```
Client3#show ip bgp 10.33.1.0/24
BGP routing table entry for 10.33.1.0/24, version 4
Paths: (2 available, best #2, table Default-IP-Routing-Table)
  Not advertised to any peer
  Local
    10.1.255.1 (metric 156160) from 10.1.255.20 (10.1.255.20)
      Origin IGP, metric 0, localpref 100, valid, internal
      Originator: 10.1.255.1, Cluster list: 10.1.255.100
  Local
    10.1.255.1 (metric 156160) from 10.1.255.10 (10.1.255.10)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      Originator: 10.1.255.1, Cluster list: 10.1.255.100
```

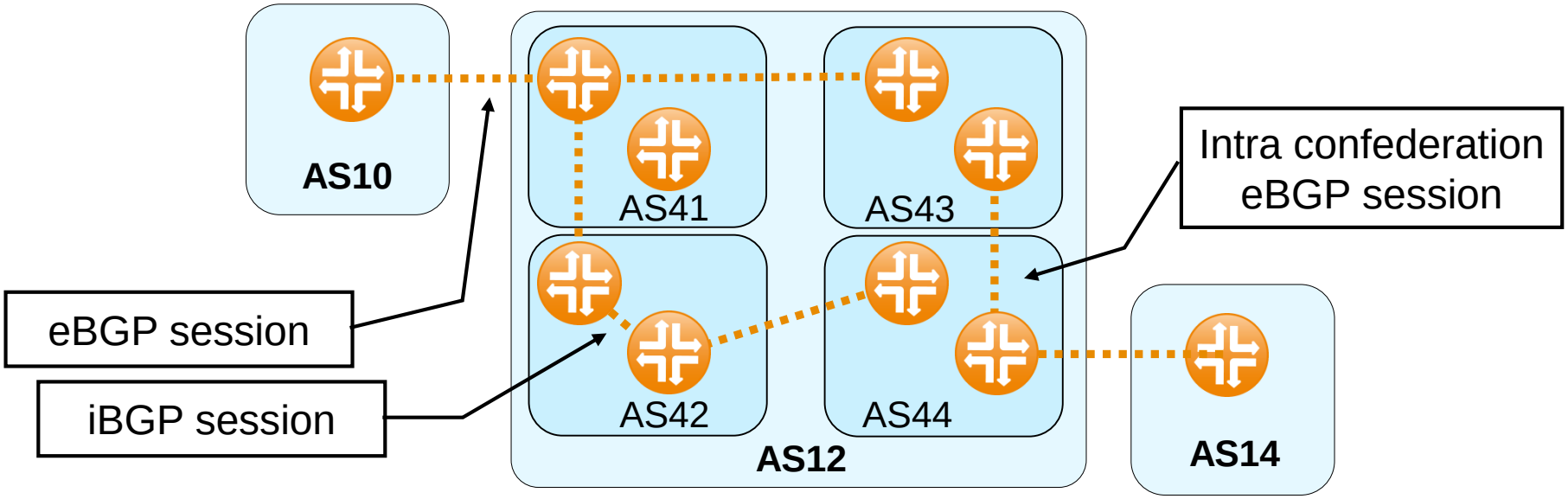
Peer Groups

- Reduction of resource requirements (CPU load and memory) when formatting and propagating the update
- Peer groups reduce the BGP configuration

```
router bgp 65000

  neighbor CLIENT-DEF peer-group
  neighbor CLIENT-DEF remote-as 65000
  neighbor CLIENT-DEF update-source Loopback0
  neighbor CLIENT-DEF route-reflector-client
  neighbor CLIENT-DEF route-map DEFAULT-OUT out
  neighbor CLIENT-DEF send-community

  neighbor CLIENT-FULL peer-group
  neighbor CLIENT-FULL remote-as 65000
  neighbor CLIENT-FULL update-source Loopback0
  neighbor CLIENT-FULL route-reflector-client
  neighbor CLIENT-FULL route-map FULL-OUT out
  neighbor CLIENT-FULL send-community

  neighbor 10.1.255.3 peer-group CLIENT-DEF
  neighbor 10.1.255.4 peer-group CLIENT-FULL
!
```

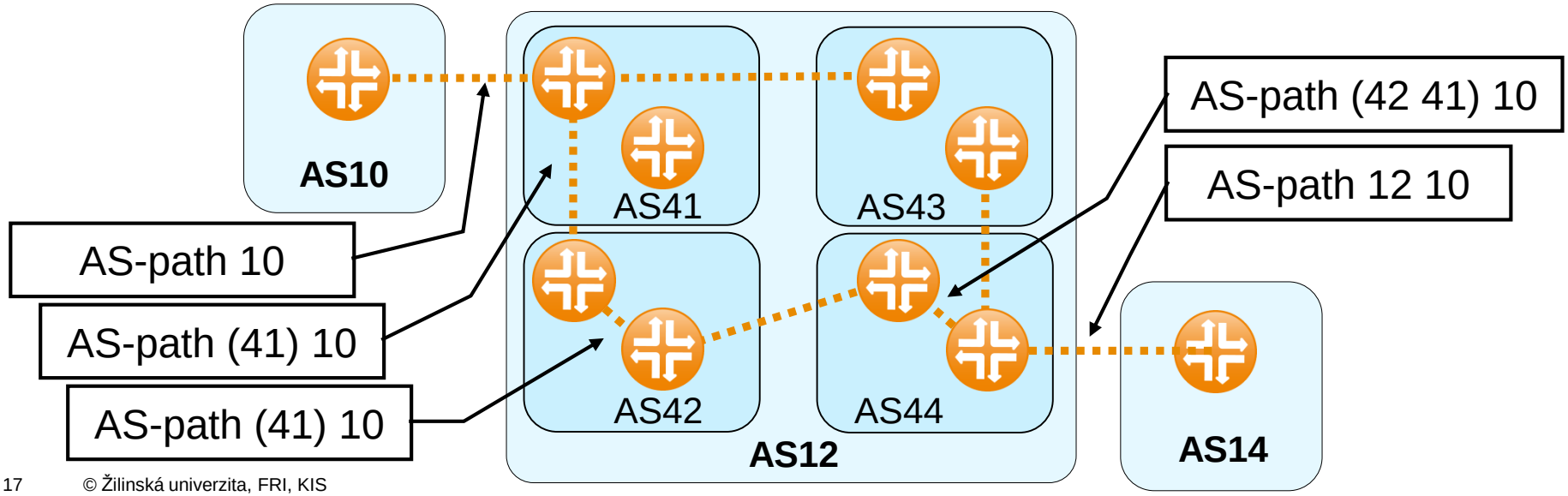
BGP Confederations

- BGP confederations introduce the concept of a number of smaller autonomous systems within original AS
- The small autonomous systems exchange BGP updates between them using intra confederation eBGP sessions
- Internal ASNs are hidden and only external AS is announced to eBGP neighbors



AS-Path Propagation within the BGP Confederation

- iBGP session
  - AS path is not changed
- Intraconfederation eBGP session
  - Intraconfederation ASN is prepended to AS path
- eBGP session with external neighbor
  - Intraconfederation ASNs are removed from AS path
  - External AS number is prepended to the AS path



AS-Path Processing in BGP Confederations

- Intraconfederation AS path is encoded as a separate segment of the AS path
  - Displayed in parentheses
- All routers within the BGP confederation have to support BGP confederations
  - A router not supporting it will reject AS path with unknown segment type
- Intraconfederation eBGP:
  - Behaves like eBGP session
    - directly connected neighbor
  - Behaves like iBGP when propagating routing updates
    - Local Pref, MED and Next Hop attributes are not changed
    - The whole confederation can run one IGP

Confederation Design Rules and Configuration

- iBGP full mesh within each member-AS
  - RRs might be used within each AS
- The member-ASes and intraconfederation eBGP sessions typically follows the physical topology of the network
- Use ASNs from private range

```
router bgp 41
 bgp confederation identifier 12
 bgp confederation peers 42 43 44
! iBGP
 neighbor 10.1.255.42 remote-as 41
 neighbor 10.1.255.42 update-source Loopback0
 neighbor 10.1.255.42 next-hop-self
! intraconfederation peers
 neighbor 10.1.255.42 remote-as 42
 neighbor 10.1.255.42 update-source Loopback0
 neighbor 10.1.255.42 next-hop-self
! external peers
 neighbor 20.20.12.1 remote-as 10
```

Customer-to-ISP connectivity

- Customer connectivity types:
  - **Internet connectivity**
  - **VPN connectivity**

- Internet customers have a wide range of requirements

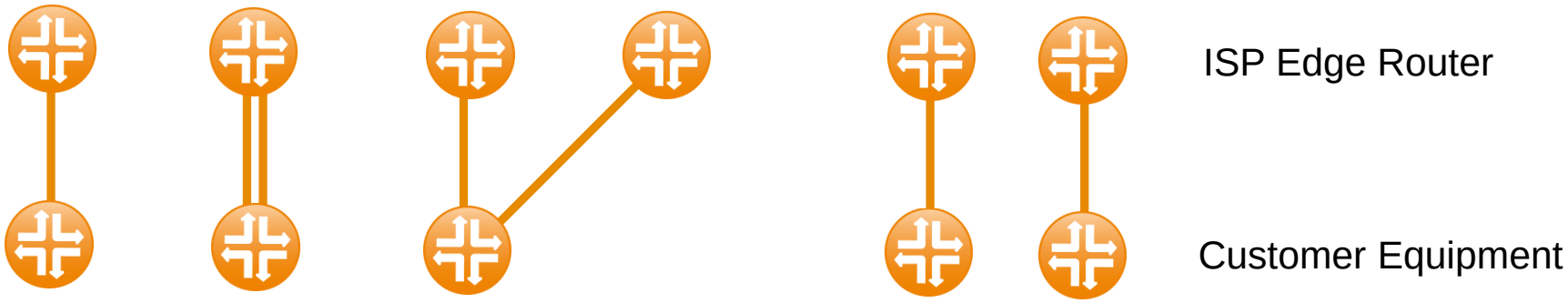
1. **Single permanent connection to the Internet**
2. **Multiple permanent connections to a single ISP**
   - Primary and backup configurations
   - Traffic load sharing
3. **Multiple permanent connections to a different ISPs**
   - Maximum redundancy
4. **Dial up residential customers**

Customer-to-ISP Addressing and Routing

- Public IP addresses for interconnection numbering
  - The customer and ISP private address blocks may overlap

- Routing
  - Static Routes
  - eBGP
  - OSPF, RIPv2 – possible but not so appropriate (no loop avoidance in case of multihoming, limited control of prefix redistribution, RIP updates)

Typical customer-to-ISP topologies:



Customer-to-ISP Static Routing

- Simple, stable and do not require a huge part of router resources
- Default routes often configured on the customer side
- To improve the routing stability, the keyword “permanent” can be appended to the static route in order to avoid flapping

- Drawbacks
  - Static configuration, always needs to be reconfigured if the customer routing requirement changes
  - Might lead to the administrative overhead on Provider Edge (PE) routers especially if default route not possible (pointing elsewhere)

```
router bgp 1
  redistribute connected <redistribution options>
  redistribute static <redistribution options>
!
ip route <address> <mask> <destination>
```

Customer-to-ISP eBGP

- The best routing protocol for redundant PE-CE links
  - Extensive routing policy features
  - Prevents routing loops, good for multi-homing
  - Private or public (multihomed to different ISPs) ASN

- Drawback: CE image requires expensive eBGP feature

- Internet Gateway router (IGW) should not propagate the AS_Path with the private ASN of the customer to the Internet
  - **neighbor a.b.c.d remove-private-as**
  - Note: If the AS_Path includes both private and public AS numbers, BGP doesn't remove the private AS numbers. This situation is considered a configuration error

```
router bgp 1
  redistribute connected <redistribution options>
  neighbor a.b.c.d remote-as 65001
  neighbor a.b.c.d send-community
  neighbor a.b.c.d maximum-prefix 100
  neighbor a.b.c.d route-map customer1 in
```

Customer-to-ISP OSPF

- Customers on PE-CE links often request OSPF because they use it as their IGP
- However, routers were able to support limited number routing processes (typically up to 32) in total (or 30 per each VRF), this includes connected routes, static routes, BGP and IGP
  - IOS release 12.3(4)T when the limitation was removed, resulting in availability of up to 30 routing processes per VRF.
  - Limitation has never been the case of RIPv2
- not recommended to implement more than few OSPF routing processes per router, therefore not so scalable solution

```
router ospf 2
  passive interface loopback2
  redistribute bgp 10 subnets <redistributing options>
  network <PE_CE_subnet> area 0
!
router bgp 10
  redistribute ospf 2

---

CE#
router ospf 2
  network <PE_CE_subnet> area 0
  network <LAN network> 0.0.0.255 area 1
```

BGP load balancing on parallel PE-CE links

- Load balancing can be achieved between two routers sharing multiple paths without having routing updates being duplicated over the two paths
- Dedicated loopback interfaces for single peering session
- Static routes used to point to the loopback interfaces via both of the physical interfaces
- IP routing table will have two paths to reach the Next Hop and will load balance

Lo1 10.1.255.100

Lo1 10.1.255.65

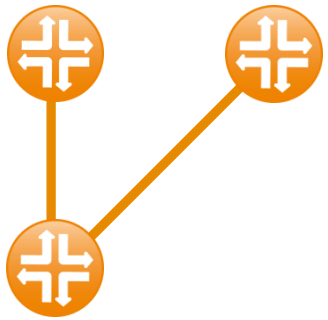
```
interface Loopback1
ip address 10.1.255.100 255.255.255.255

router bgp 10
  neighbor 10.1.255.65 remote-as 65001
  neighbor 10.1.255.65 ebgp-multihop 2
  neighbor 10.1.255.65 update-source Loopback1

ip route 10.1.255.65 255.255.255.255 <remote_IP_address_serial1>
ip route 10.1.255.65 255.255.255.255 <remote_IP_address_serial2>
```

Multihomed Site with a Single CE

- If possible, the configuration of routing policy is completely offloaded to CE in order to simplify PE configuration
- eBGP used for route redistribution

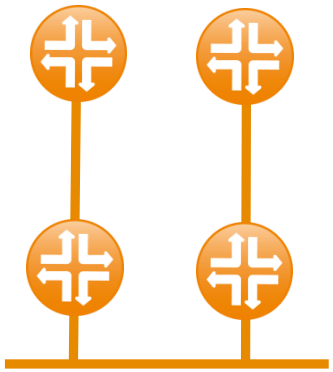
```
router bgp 65001

  ! Primary link
  neighbor <link-on-PE1> remote-as 10
  neighbor <link-on-PE1> route-map MED-primary out
  neighbor <link-on-PE1> route-map LP-primary in

  ! Backup Link
  neighbor <link-on-PE2> remote-as 10
  neighbor <link-on-PE2> route-map MED-secondary out
  neighbor <link-on-PE2> route-map LP-secondary in
!
route-map MED-primary permit 10
  set metric 90
!
route-map LP-primary permit 10
set local-preference 110
!
route-map MED-backup permit 10
set metric 110
!
route-map LP-backup permit 10
set local-preference 90
```

- Primary and Backup links can be implemented using:
  - MED for traffic from PE to CE
  - Local Pref for traffic from CE to PE



Multihomed Site with Multiple CEs

- eBGP is recommended as the routing protocol
- For traffic going from PE to CE, MED can be configured on the CEs
- For traffic from CE to PE, Primary and Backup CE is implemented:
  - VRRP for the traffic originating from the LAN interface of the CE (LAN nodes cannot run IGP)
  - Redistributing routes learnt from the PE (via BGP) into the IGP with different metrics



LAN with VRRP (Virtual Router Redundancy Protocol)

Ďakujem za pozornosť

roman dot kaloc at gmail dot com