



BGP in the Internet

Best Current Practices and Scaling Techniques

CISCO SYSTEMS





What is BGP for??

What is an IGP not for?

BGP versus OSPF/ISIS

- **Internal Routing Protocols (IGPs)**

examples are ISIS and OSPF

used for carrying infrastructure addresses

NOT used for carrying Internet prefixes or customer prefixes

BGP versus OSPF/ISIS

- **BGP used internally (iBGP) and externally (eBGP)**
- **iBGP used to carry
some/all Internet prefixes across backbone
customer prefixes**
- **eBGP used to
exchange prefixes with other ASes
implement routing policy**

BGP versus OSPF/ISIS

- **DO NOT:**
 - distribute BGP prefixes into an IGP**
 - distribute IGP routes into BGP**
 - use an IGP to carry customer prefixes**
- **YOUR NETWORK WILL NOT SCALE**



Generating an Aggregate

Aggregation

- **ISPs receive address block from Regional Registry or upstream provider**
- **Aggregation** means announcing the **address block** only, not subprefixes
- **Aggregate should be generated internally**

Configuring Aggregation - Cisco IOS

- **ISP has 221.10.0.0/19 address block**
- **To put into BGP as an aggregate:**

```
router bgp 100
```

```
network 221.10.0.0 mask 255.255.224.0
```

```
ip route 221.10.0.0 255.255.224.0 null0 250
```

- **The static route is a “pull up” route**

**more specific prefixes within this address block
ensure connectivity to ISP’s customers**

“longest match lookup”



Announcing Aggregate

Aggregation

- **Address block should be announced to the Internet as an aggregate**
- **Subprefixes of address block should NOT be announced to Internet unless **special** circumstances (more later)**

Announcing Aggregate - Cisco IOS

- **Configuration Example**

```
router bgp 100
```

```
network 221.10.0.0 mask 255.255.224.0
```

```
neighbor 222.222.10.1 remote-as 101
```

```
neighbor 222.222.10.1 prefix-list out-filter out
```

```
!
```

```
ip route 221.10.0.0 255.255.224.0 null0
```

```
!
```

```
ip prefix-list out-filter permit 221.10.0.0/19
```

```
ip prefix-list out-filter deny 0.0.0.0/0 le 32
```


Announcing an Aggregate

- **ISPs who don't and won't aggregate are held in poor regard by community**
- **Registries' minimum allocation sizes are /19s or /20s now**

no real reason to see anything longer than a /21 or /22 prefix in the Internet

BUT there are currently >46000 /24s!



Receiving Prefixes

Receiving Prefixes from downstream peers

- **ISPs should only accept prefixes which have been assigned or allocated to their downstream peer**
- **For example**
 - downstream has 220.50.0.0/20 block**
 - should only announce this to peers**
 - peers should only accept this from them**

Receiving Prefixes - Cisco IOS

- **Configuration Example on upstream**

```
router bgp 100
```

```
neighbor 222.222.10.1 remote-as 101
```

```
neighbor 222.222.10.1 prefix-list customer in
```

```
!
```

```
ip prefix-list customer permit 220.50.0.0/20
```

```
ip prefix-list customer deny 0.0.0.0/0 le 32
```

Receiving Prefixes from upstream peers

- **Not desirable unless really necessary**
special circumstances
- **Ask upstream to either:**
originate a default-route
announce one prefix you can use as default

Receiving Prefixes from upstream peers

- **Downstream Router Configuration**

```
router bgp 100
  network 221.10.0.0 mask 255.255.224.0
  neighbor 221.5.7.1 remote-as 101
  neighbor 221.5.7.1 prefix-list infilt in
  neighbor 221.5.7.1 prefix-list outfilt out
!
ip prefix-list infilt permit 0.0.0.0/0
ip prefix-list infilt deny 0.0.0.0/0 le 32
!
ip prefix-list outfilt permit 221.10.0.0/19
ip prefix-list outfilt deny 0.0.0.0/0 le 32
```


Receiving Prefixes from upstream peers

- **Upstream Router Configuration**

```
router bgp 101
```

```
neighbor 221.5.7.2 remote-as 100
```

```
neighbor 221.5.7.2 default-originate
```

```
neighbor 221.5.7.2 prefix-list cust-in in
```

```
neighbor 221.5.7.2 prefix-list cust-out out
```

```
!
```

```
ip prefix-list cust-in permit 221.10.0.0/19
```

```
ip prefix-list cust-in deny 0.0.0.0/0 le 32
```

```
!
```

```
ip prefix-list cust-out permit 0.0.0.0/0
```

```
ip prefix-list cust-out deny 0.0.0.0/0 le 32
```

Receiving Prefixes from upstream peers

- **If necessary to receive prefixes from upstream provider, care is required**
 - don't accept RFC1918 etc prefixes**
 - don't accept your own prefix**
 - don't accept default (unless you need it)**
 - don't accept prefixes longer than /24**

Receiving Prefixes

```
router bgp 100
  network 221.10.0.0 mask 255.255.224.0
  neighbor 221.5.7.1 remote-as 101
  neighbor 221.5.7.1 prefix-list in-filter in
!
ip prefix-list in-filter deny 0.0.0.0/0                ! Block default
ip prefix-list in-filter deny 0.0.0.0/8 le 32
ip prefix-list in-filter deny 10.0.0.0/8 le 32
ip prefix-list in-filter deny 127.0.0.0/8 le 32
ip prefix-list in-filter deny 169.254.0.0/16 le 32
ip prefix-list in-filter deny 172.16.0.0/12 le 32
ip prefix-list in-filter deny 192.0.2.0/24 le 32
ip prefix-list in-filter deny 192.168.0.0/16 le 32
ip prefix-list in-filter deny 221.10.0.0/19 le 32      ! Block local prefix
ip prefix-list in-filter deny 224.0.0.0/3 le 32       ! Block multicast
ip prefix-list in-filter deny 0.0.0.0/0 ge 25         ! Block prefixes >/24
ip prefix-list in-filter permit 0.0.0.0/0 le 32
```


“Documenting Special Use Addresses” - DSUA

- **This prefix-list MUST be applied to all external BGP peerings, in and out!**

<http://www.ietf.org/internet-drafts/draft-manning-dsua-03.txt>

```
ip prefix-list rfc1918-dsua deny 0.0.0.0/8 le 32
ip prefix-list rfc1918-dsua deny 10.0.0.0/8 le 32
ip prefix-list rfc1918-dsua deny 127.0.0.0/8 le 32
ip prefix-list rfc1918-dsua deny 169.254.0.0/16 le 32
ip prefix-list rfc1918-dsua deny 172.16.0.0/12 le 32
ip prefix-list rfc1918-dsua deny 192.0.2.0/24 le 32
ip prefix-list rfc1918-dsua deny 192.168.0.0/16 le 32
ip prefix-list rfc1918-dsua deny 224.0.0.0/3 le 32
ip prefix-list rfc1918-dsua deny 0.0.0.0/0 ge 25
ip prefix-list rfc1918-dsua permit 0.0.0.0/0 le 32
```



Prefixes into iBGP

Injecting prefixes into iBGP

- **Use iBGP to carry customer prefixes
don't use IGP**
- **Point static route to customer interface**
- **Use BGP network statement**
- **As long as static route exists (interface active), prefix will be in BGP**

Router Configuration network statement

- **Example:**

```
interface loopback 0
  ip address 215.17.3.1 255.255.255.255
!
interface Serial 5/0
  ip unnumbered loopback 0
  ip verify unicast reverse-path
!
ip route 215.34.10.0 255.255.252.0 Serial 5/0
!
router bgp 100
  network 215.34.10.0 mask 255.255.252.0
```



Scaling BGP

Scaling BGP

- **Dynamic Reconfiguration**
- **Peer Groups**
- **Route Flap Dampening**



Dynamic Reconfiguration

Soft Reconfiguration and Route Refresh

Soft Reconfiguration

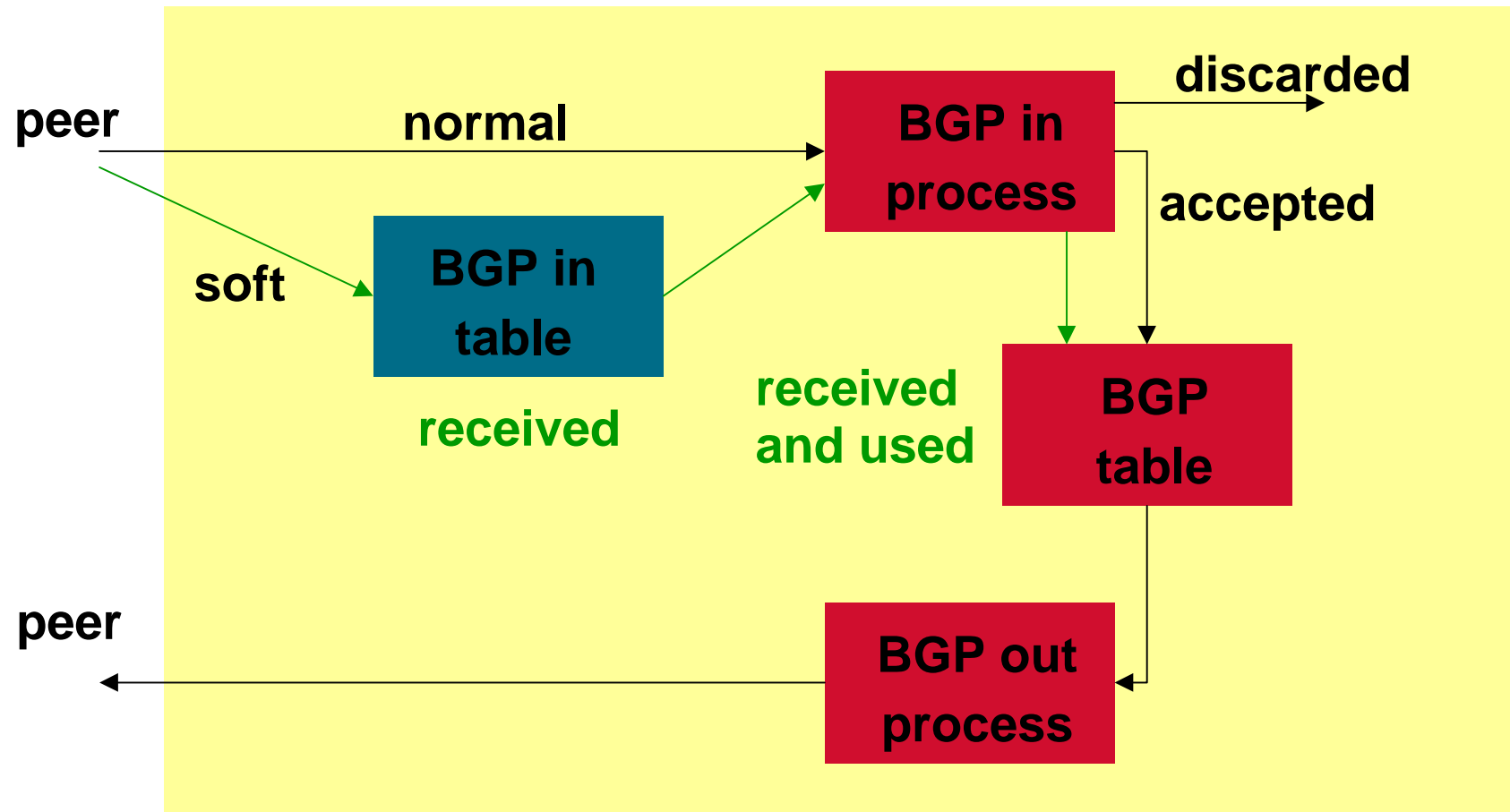
Problem:

- **Hard BGP peer clear required after every policy change because the router does not store prefixes that are denied by a filter**
- **Hard BGP peer clearing consumes CPU and affects connectivity for all networks**

Solution:

- **Soft-reconfiguration**

Soft Reconfiguration



Soft Reconfiguration

- **New policy is activated without tearing down and restarting the peering session**
- **Per-neighbour basis**
- **Use more memory to keep prefixes whose attributes have been changed or have not been accepted**

Configuring Soft reconfiguration

```
router bgp 100
```

```
neighbor 1.1.1.1 remote-as 101
```

```
neighbor 1.1.1.1 route-map infilter in
```

```
neighbor 1.1.1.1 soft-reconfiguration inbound
```

! Outbound does not need to be configured !

Then when we change the policy, we issue an exec command

```
clear ip bgp 1.1.1.1 soft [in | out]
```


Route Refresh Capability

- **Facilitates non-disruptive policy changes**
- **No configuration is needed**
- **No additional memory is used**
- **Requires peering routers to support “route refresh capability” - RFC2842**
- **clear ip bgp x.x.x.x in** tells peer to resend full BGP announcement

Soft Reconfiguration vs Route Refresh

- **Use Route Refresh capability if supported**
 - find out from “show ip bgp neighbor”
 - uses much less memory
- **Otherwise use Soft Reconfiguration**



Peer Groups

**Speeding up the building of the
iBGP mesh**

Peer Groups

Without peer groups

- **iBGP neighbours receive same update**
- **Large iBGP mesh slow to build**
- **Router CPU wasted on repeat calculations**

Solution - peer groups!

- **Group peers with same outbound policy**
- **Updates are generated once per group**

Peer Groups - Advantages

- **Makes configuration easier**
- **Makes configuration less prone to error**
- **Makes configuration more readable**
- **Lower router CPU load**
- **iBGP mesh builds more quickly**
- **Members can have different inbound policy**
- **Can be used for eBGP neighbours too!**

Configuring Peer Group iBGP

```
router bgp 100
  neighbor ibgp-peer peer-group
  neighbor ibgp-peer remote-as 100
  neighbor ibgp-peer update-source loopback 0
  neighbor ibgp-peer send-community
  neighbor ibgp-peer route-map outfilter out
  neighbor 1.1.1.1 peer-group ibgp-peer
  neighbor 2.2.2.2 peer-group ibgp-peer
  neighbor 2.2.2.2 route-map infilter in
  neighbor 3.3.3.3 peer-group ibgp-peer
```

! note how 2.2.2.2 has different inbound filter from peer-group !

Configuring Peer Group eBGP

```
router bgp 109
  neighbor external-peer peer-group
  neighbor external-peer send-community
  neighbor external-peer route-map set-metric out
  neighbor 160.89.1.2 remote-as 200
  neighbor 160.89.1.2 peer-group external-peer
  neighbor 160.89.1.4 remote-as 300
  neighbor 160.89.1.4 peer-group external-peer
  neighbor 160.89.1.6 remote-as 400
  neighbor 160.89.1.6 peer-group external-peer
  neighbor 160.89.1.6 filter-list infiltrer in
```



Route Flap Dampening

Stabilising the Network

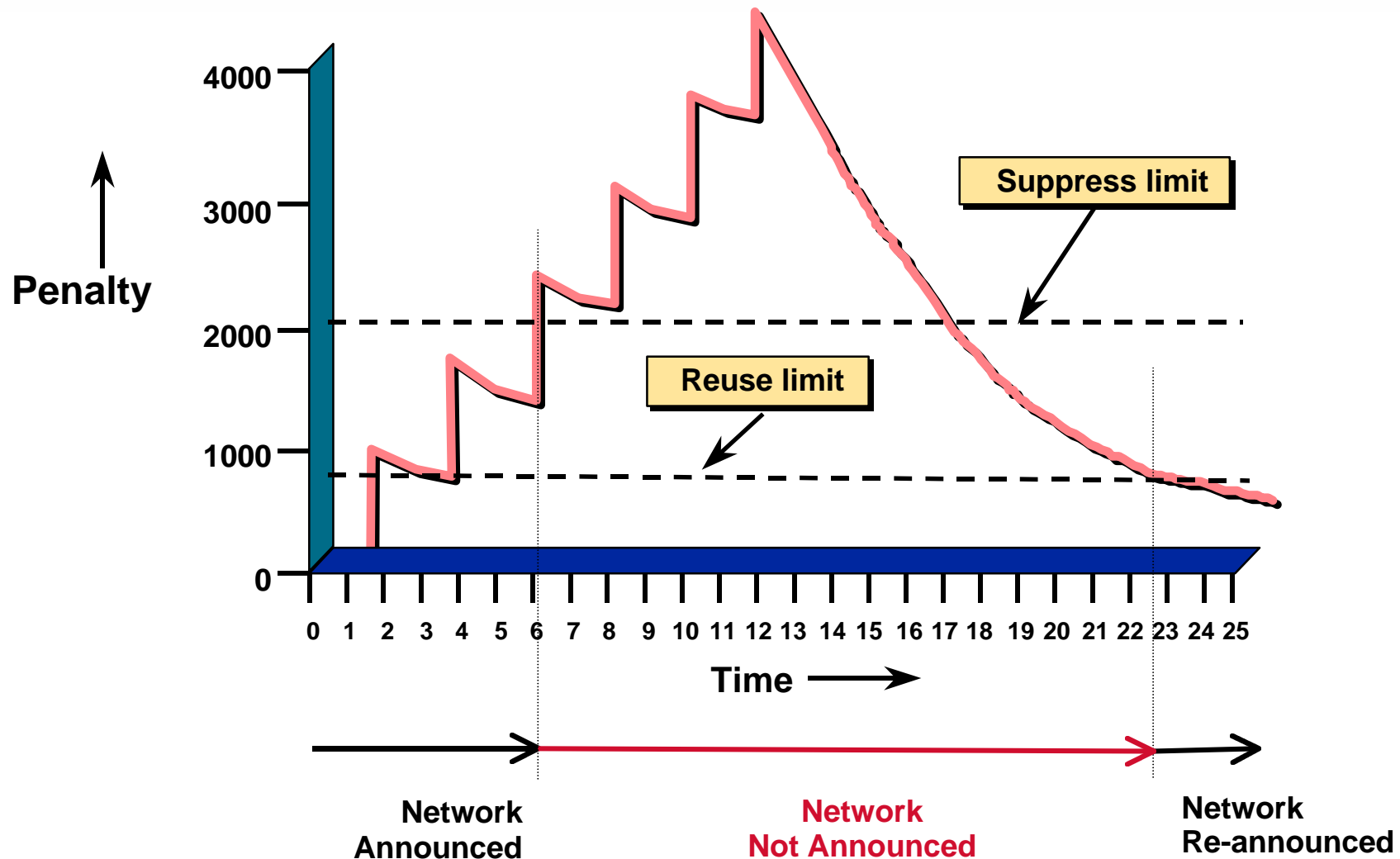
Route Flap Dampening

- **Route flap**
 - Going up and down of path/change in attribute**
 - Ripples through the entire Internet, wastes CPU**
- **Dampening aims to reduce flap propagation**
 - Fast convergence for normal route changes**
 - History predicts future behaviour**
 - Suppress oscillating routes, advertise stable routes**
- **Described in RFC2439**

Route Flap Dampening - Operation

- **Add penalty (1000) for each flap**
- **Exponentially decay penalty**
half life determines decay rate
- **Penalty above suppress-limit**
do not advertise route to BGP peers
- **Penalty decayed below reuse-limit**
re-advertise route to BGP peers

Route Flap Dampening



Route Flap Dampening - Operation

- **Only applied to inbound announcements from eBGP peers**
- **Alternate paths still usable**
- **Controlled by:**
 - Half-life (default 15 minutes)**
 - reuse-limit (default 750)**
 - suppress-limit (default 2000)**
 - maximum suppress time (default 30 minutes)**

Flap Dampening: Enhancements

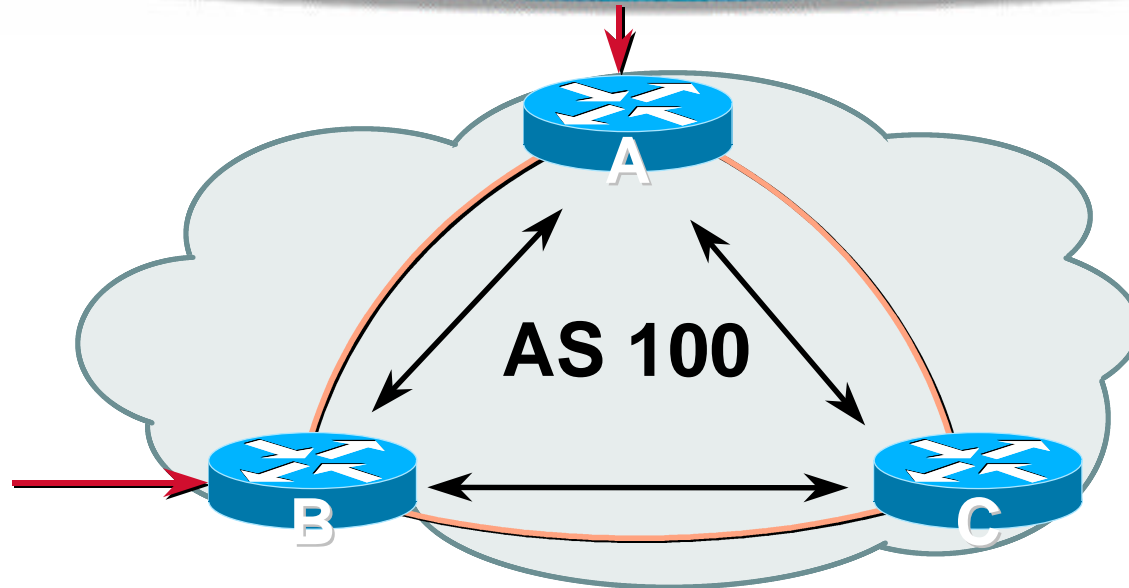
- **Selective dampening based on AS-path, Community, Prefix**
- **Variable dampening recommendations for ISPs**
<http://www.ripe.net/docs/ripe-210.html>



Route Reflectors

Scaling the iBGP mesh

Scaling iBGP mesh



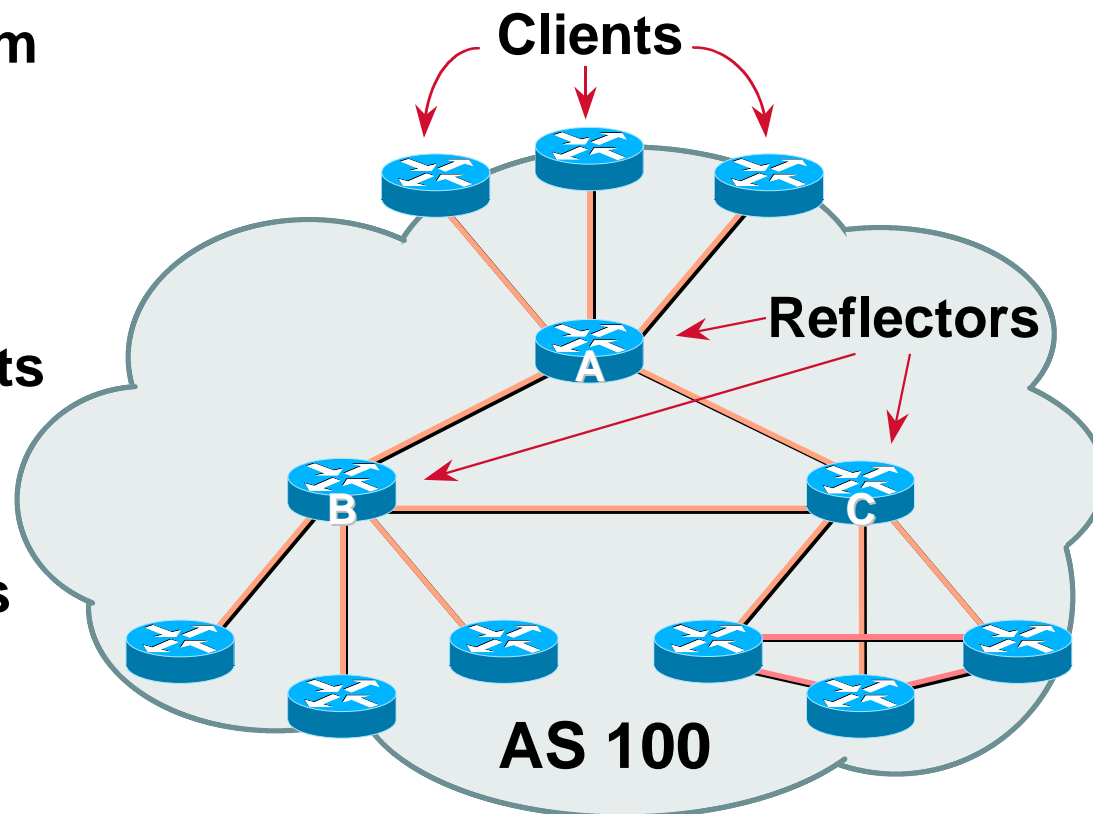
- Need to avoid routing information loop
- Solution should not change the current behaviour
- Two solutions

Route reflector - simpler to deploy and run

Confederation - more complex to manage, corner case benefits

Route Reflector

- Reflector receives path from clients and non-clients
- Selects best path
- If best path is from client, reflect to other clients and non-clients
- If best path is from non-client, reflect to clients only
- Non-meshed clients
- Described in RFC2796



Route Reflector Topology

- **Divide the backbone into multiple clusters**
- **At least one route reflector and few clients per cluster**
- **Route reflectors are fully meshed**
- **Clients in a cluster could be fully meshed**
- **Single IGP to carry next hop and local routes**

Route Reflectors: Loop Avoidance

- **Originator_ID attribute**

Carries the RID of the originator of the route in the local AS (created by the RR)

- **Cluster_list attribute**

The local cluster-id is added when the update is sent to (added by the RR)

Route Reflector: Benefits

- **Solves iBGP mesh problem**
- **Packet forwarding is not affected**
- **Normal BGP speakers co-exist**
- **Multiple reflectors for redundancy**
- **Easy migration**
- **Multiple levels of route reflectors**

Route Reflectors: Migration

- **Where to place the route reflectors?**

Follow the physical topology!

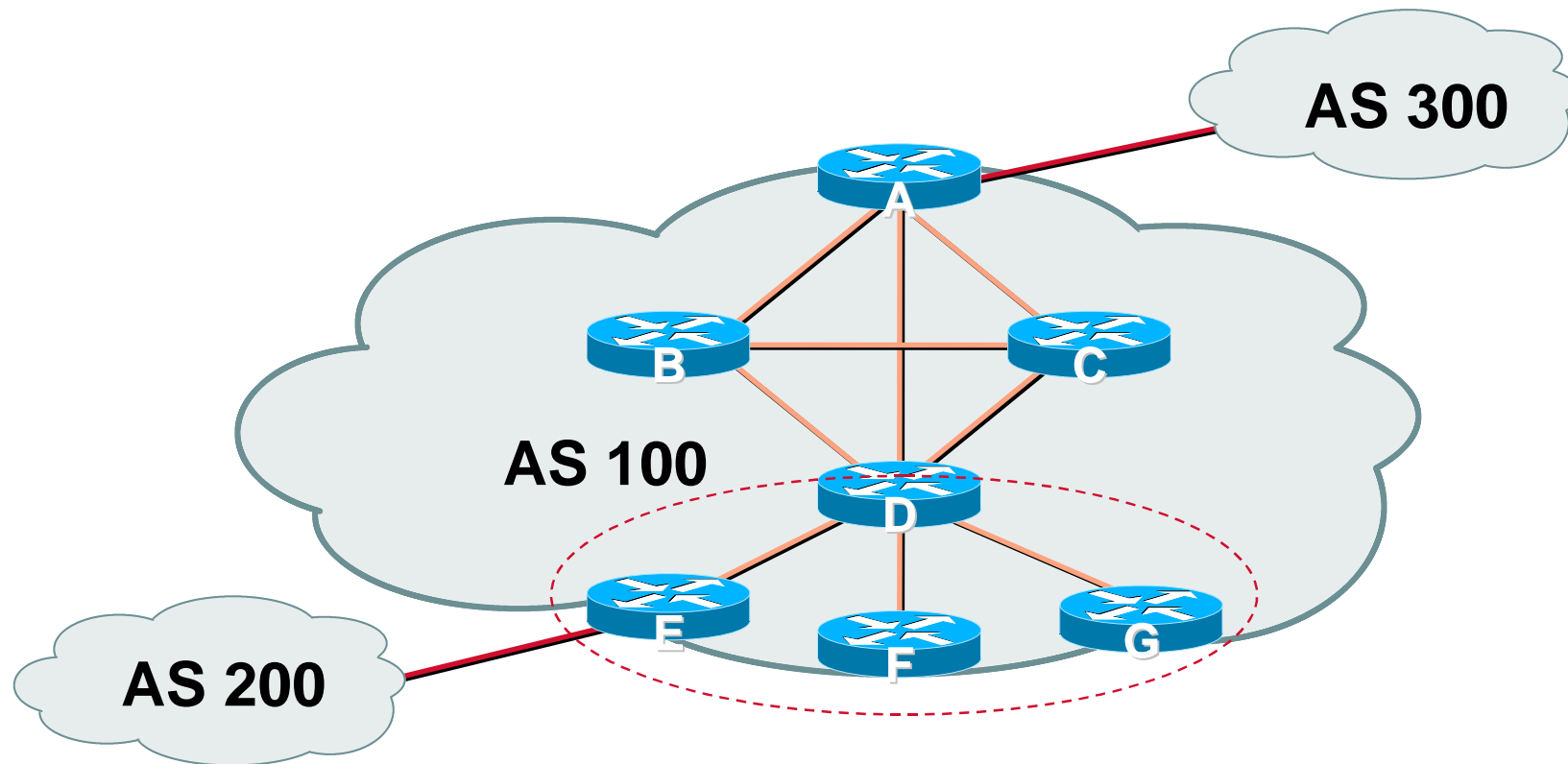
This will guarantee that the packet forwarding won't be affected

- **Configure one RR at a time**

Eliminate redundant iBGP sessions

Place one RR per cluster

Route Reflector: Migration



- **Migrate small parts of the network, one part at a time.**

Configuring a Route Reflector

```
router bgp 100
  neighbor 1.1.1.1 remote-as 100
  neighbor 1.1.1.1 route-reflector-client
  neighbor 2.2.2.2 remote-as 100
  neighbor 2.2.2.2 route-reflector-client
  neighbor 3.3.3.3 remote-as 100
  neighbor 3.3.3.3 route-reflector-client
```

BGP Scaling Techniques

- **These 4 techniques should be core requirements on all ISP networks**

Soft reconfiguration/Route Refresh

Peer groups

Route flap dampening

Route reflectors

Summary

- **BGP versus IGP**
 - **ALWAYS** announce aggregate
 - **Receiving & originating prefixes**
 - **The 4 BGP scaling techniques**
-
- **Any questions?**