

BGP Techniques for ISP

Terutaka Komorizono <teru@ntt.co.th>

Introduction

- Presentation has many configuration examples
- Using Cisco IOS CLI
- Aimed at Service Providers
 - Techniques can be used by many enterprises too
- Feel free to ask questions

BGP Techniques for ISP

- Applying Policy by BGP
- Route Flap Damping
- Using Communities
- Deploying BGP in an ISP network
- Aggregation
- Service Providers Multihoming

Applying Policy by BGP

- Policies are applied to:
 - Influence BGP Path Selection by setting BGP attributes
 - Determine which prefixes are announced or blocked
 - Determine which AS-paths are preferred, permitted, or denied
 - Determine route groupings and their effects
- Decisions are generally based on prefix, AS-path and community

Applying Policy by BGP

- Most implementations have tools to apply Policies to BGP
 - Prefix manipulation/filtering
 - AS-PATH manipulations/filtering
 - Community Attribute setting and matching
- Implementations also have policy language which can do various match/set constructs on the attributes of chosen BGP routes

Route Refresh

- BGP peer reset required after every policy change
 - Because the router does not store prefixes which are rejected by policy
- Hard BGP peer reset:
 - Terminate BGP peering & Consumes CPU
 - Severely disrupts connectivity for all networks
- Soft BGP peer reset (or **Route Refresh**)
 - BGP peering remains active
 - Impacts only those prefixes affected by policy change

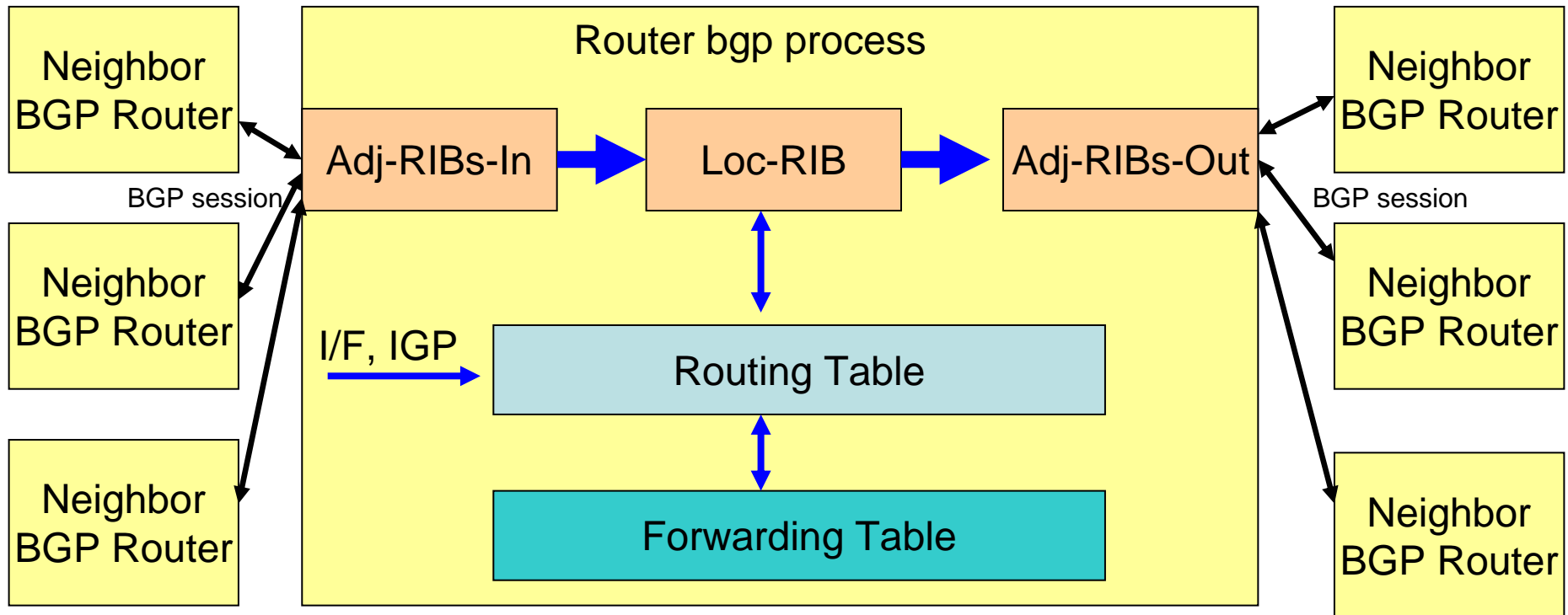
Route Refresh

- Facilitates non-disruptive policy changes
- For most implementations, no configuration is needed
 - Automatically negotiated at peer establishment
- No additional memory is used
- Required peering routers to support “route refresh capability” – RFC2918

Route Refresh

- Use Route Refresh capability if supported
 - Find out from the BGP neighbor status display
 - Non-disruptive, “Good For the internet”
- If not supported, see if implementations has a workaround
- Only hard-reset a BGP peering as a last resort

Router BGP route process



- Router manages RIB (Routing Information Base).
- RIB composes update-time, prefix, and prefix-attribute.
- 3 kinds of RIB.
 - Adj-RIBs-In The unedited routing information sent by neighboring routers.
 - Loc-RIB The actual routing information the router uses, developed from Adj-RIBs-In
 - Adj-RIBs-Out The information the router chooses to send to neighboring routers.

Adj-RIBs-In and Loc-RIB

ISP-B0#**show ip bgp neighbor 172.16.11.2 received-routes**

BGP table version is 38, local router ID is 172.16.10.1

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 |
|----------------------|--------------------|----------|----------|-------------|----------|
| * 192.168.0.0 | 172.16.11.2 | 0 | 0 | 1000 | i |
| * 192.168.1.0 | 172.16.11.2 | 0 | 0 | 1000 | i |
| * 192.168.2.0 | 172.16.11.2 | 0 | 0 | 1000 | i |

Total number of prefixes 3

ISP-B0#**show ip bgp neighbor 172.16.11.2 routes**

BGP table version is 38, local router ID is 172.16.10.1

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 |
|----------------|-------------|--------|--------|--------|--------|
| *> 192.168.0.0 | 172.16.11.2 | 0 | 110 | 0 | 1000 i |
| *> 192.168.1.0 | 172.16.11.2 | 0 | 110 | 0 | 1000 i |

Total number of prefixes 2

Adj-RIBs-In



RIB

RIB

- If you do not send all BGP routes from the other party (to other party) again (Do not receive all it again), It is because own RIB and internal BGP-Speaker's RIB becomes a disagreement, then RIB becomes a disagreement, and there is a possibility that the state that cannot be communicated. *****
- `clear ip bgp x.x.x.x [soft] in/out` (cisco)
- Automatically refresh (juniper)

Route Flap Damping

Stabilising the network

Route Flap Damping

- Route flap
 - Going up and down of path or change is attribute
 - BGP WITHDRAW followed by UPDATE = 1 flap
 - eBGP neighbor peering reset is NOT a flap
 - Ripples through the entire Internet
 - Causes instability, wastes CPU
- Damping aims to reduce scope of route flap propagation

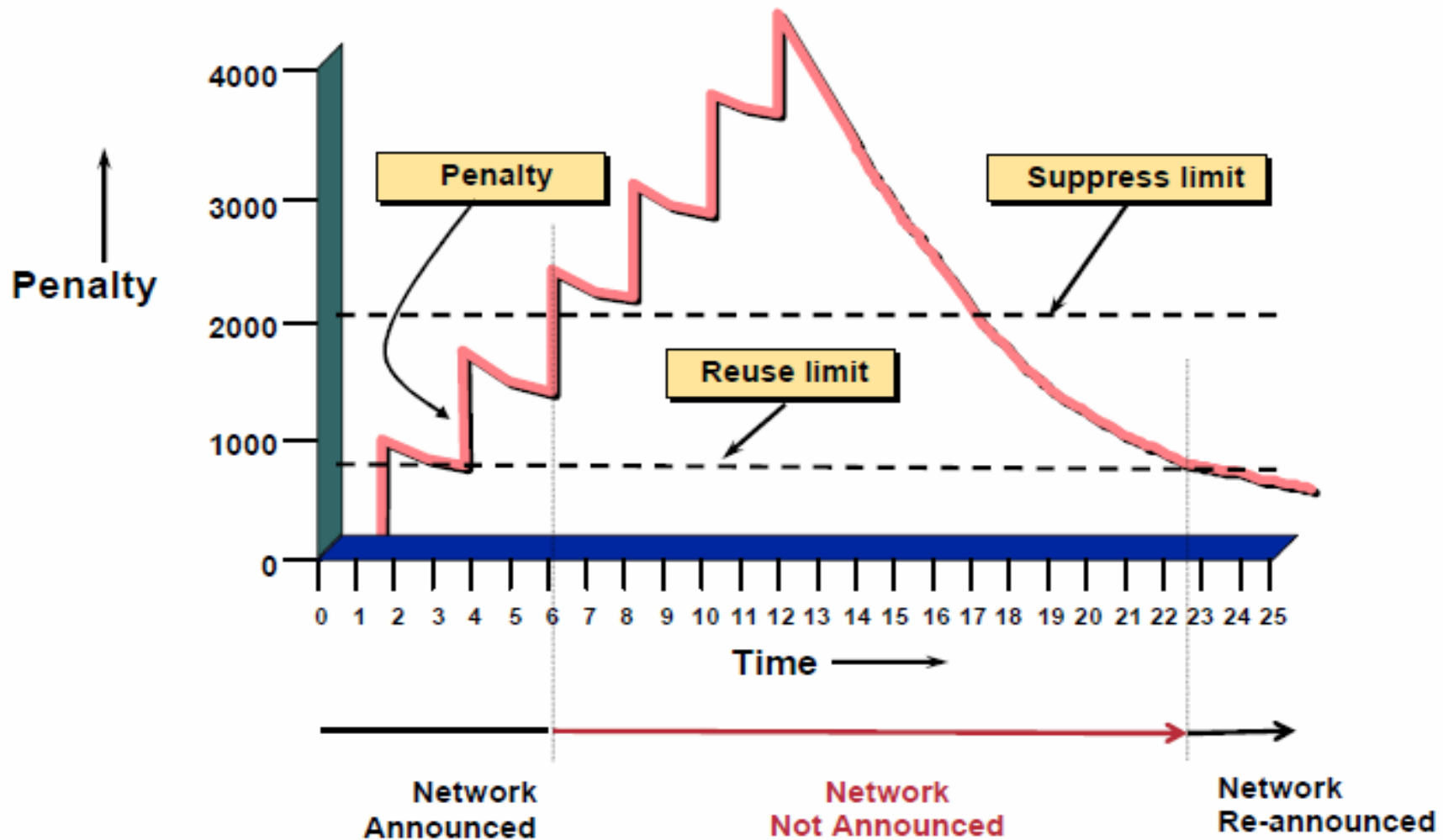
Route Flap Damping

- Requirements
 - Fast convergence for normal route changes
 - History predicts future behavior
 - Suppress oscillating routes
 - Advertise stable routes
- Documented in RFC2439

Route Flap Damping - Operation

- Add penalty for each flap
 - NB: **Change in attribute is also penalized**
- 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 Damping - Operation



Implement Flap Damping

- Flap Damping should only be implemented to address a specific network stability problem
- Flap Damping can and does make stability worse
 - “Flap Amplification” from AS path attribute changes caused by BGP exploring alternate paths being unnecessarily penalized

“Router Flap Damping Exacerbates Internet Routing Convergence”

Zhouqing Morley Mao, Ramesh Govindan, George Verghese
& Randy H. Katz, August 2002

Route Flap Damping - Operations

- Only applied to inbound announcements from BGP peers
- Alternate paths are still usable
- Controllable by at least:
 - half life
 - reuse-limit
 - suppress-limit
 - maximum suppress time

Implementing Flap Damping

- If you have to implement flap damping, understand the impact on the network
 - Vendors default are very severe
 - Variable flap damping can bring benefits
 - Transit provider flap damping impact pass ASes more harshly due to flap amplification
- Recommended for ISPs
 - <http://www.ripe.net/docs/ripe-229.html>

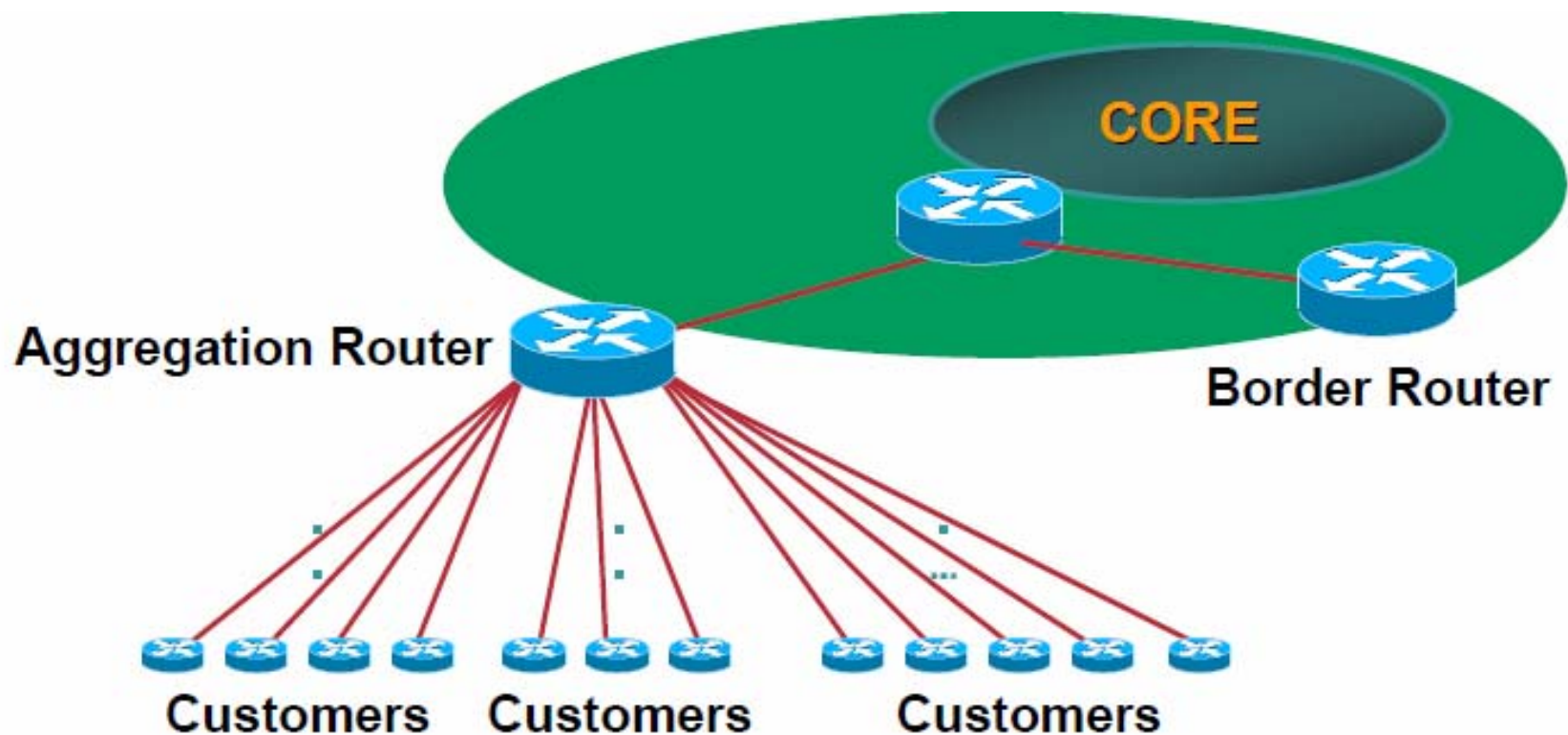
Using Communities

Some samples of how ISPs make life easier for themselves

BGP communities

- Another ISP “scaling technique”
- Prefixes are grouped into different “classes” or communities within the ISP network
- Each community means a different thing, has a different result in the ISP network

Community Example – Customer Edge



Communities set at the aggregation router where the prefix is injected into the ISP's iBGP

Community Example – Customer Edge

- No need to alter filters at the network border when adding a new customer
- New customer simply is added to the appropriate community
 - Border filters already in place take care of announcements
 - > Ease of operation!
- More experienced operators tend to have more sophisticated options available

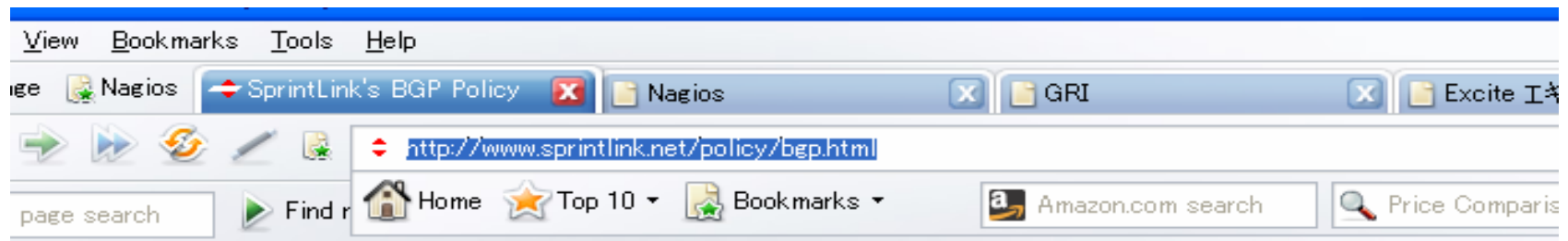
BGP communities

- Communities are generally set at the edge of the ISP network

Customer edge: customer prefixes belong to different communities depending on the services they have purchased

Internet edge: transit provider prefixes belong to different communities, depending on the loadsharing or traffic engineering requirements of the local ISP, or what the demands from its BGP customer might be

Some ISP Examples: Sprintlink



received community. Currently, the following ASes are supported: 1668, 209, 2914, 3300, 3356, 3544, 4635, 701, 7018, 702 and 8220.

| String | Resulting AS Path to ASXXX |
|-----------|--------------------------------------|
| 65000:XXX | Do not advertise to ASXXX |
| 65001:XXX | 1239 (default) ... |
| 65002:XXX | 1239 1239 ... |
| 65003:XXX | 1239 1239 1239 ... |
| 65004:XXX | 1239 1239 1239 1239 ... |
| String | Resulting AS Path to ASXXX in Asia |
| 65070:XXX | Do not advertise to ASXXX |
| 65071:XXX | 1239 (default) ... |
| 65072:XXX | 1239 1239 ... |
| 65073:XXX | 1239 1239 1239 ... |
| 65074:XXX | 1239 1239 1239 1239 ... |
| String | Resulting AS Path to ASXXX in Europe |
| 65050:XXX | Do not advertise to ASXXX |
| 65051:XXX | 1239 (default) ... |
| 65052:XXX | 1239 1239 ... |
| 65053:XXX | 1239 1239 1239 ... |
| 65054:XXX | 1239 1239 1239 1239 ... |
| String | Resulting AS Path to ASXXX in North |

Some ISP Examples: MCI

```
aut-num: AS702
descr: MCI EMEA - Commercial IP service provider in Europe
remarks: MCI uses the following communities with its customers:
702:80 Set Local Pref 80 within AS702
702:120 Set Local Pref 120 within AS702
702:20 Announce only to MCI AS'es and MCI customers
702:30 Keep within Europe, don't announce to other MCI AS's
702:1 Prepend AS702 once at edges of MCI to Peers
702:2 Prepend AS702 twice at edges of MCI to Peers
702:3 Prepend AS702 thrice at edges of MCI to Peers
Advanced communities for customers
702:7020 Do not announce to AS702 peers with a scope of
National but advertise to Global Peers, European
Peers and MCI customers.
702:7001 Prepend AS702 once at edges of MCI to AS702
peers with a scope of National.
702:7002 Prepend AS702 twice at edges of MCI to AS702
peers with a scope of National.
```

(more)

Some ISP Examples: BT

- One of the most comprehensive community lists around

`whois -h whois.ra.net AS5400`

- Extensive community definitions allow sophisticated traffic engineering by customers

Deploying BGP in an ISP network

Deploying BGP

The role of IGPs and iBGP

Aggregation

Receiving Prefixes

Configuration Tips

The role of IGPs and iBGP

BGP versus IGP (OSPF/ISIS)

- Internal Routing Protocols (IGPs)
 - used for carrying **infrastructure** addresses
 - **NOT** used for carrying internet prefixes or customer prefixes
 - design goal is to **minimize** number of prefixes in IGP to aid scalability and rapid convergence

BGP versus IGP (OSPF/ISIS)

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

BGP versus IGP (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 BE SCALE**

Injection prefixes into iBGP

- Use iBGP to carry customer prefixes
don't ever use IGP
- Point static route to customer interface
- Enter network into BGP process
Ensure that implementation options are used so that the prefix always remains in iBGP, regardless of state of interface
i.e. avoid iBGP flaps caused by interface flap

Aggregation

Aggregation

- Aggregation means announcing the address block received from the RIR to the other ASes connected your network
- Subprefixes of this aggregate may be:
 - Used internally in the ISP network
 - Announced to other ASes to aid with multihoming
- Unfortunately too many people are still thinking about class Cs, resulting in a proliferation of /24s in the internet routing table

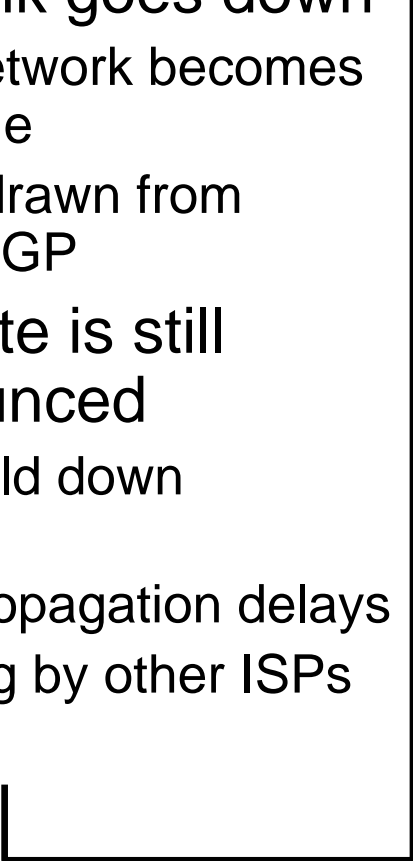
Aggregation

- 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)
- Aggregate should be generated internally
not on the network borders!

Announcing an Aggregation

- ISPs who don't and won't aggregate are held in poor regard by community
- Registries publish their minimum allocation size
 - Anything from a /20 to a /22 depending on RIR
 - Different size for different address block
- No real reason to see anything longer than a /22 prefix in the Internet
 - But there are currently >93000 /24s!

Aggregation – Good Example

- Customer link goes down
 - their /23 network becomes unreachable
 - /23 is withdrawn from AS100's iBGP
 - /19 aggregate is still begin announced
 - no BGP hold down problems
 - no BGP propagation delays
 - no damping by other ISPs
 - Customer link returns
 - Their /23 network is visible again
 - The /23 is re-injected into AS100's iBGP
 - The whole Internet becomes visible immediately
 - Customer has Quality of Service perception
- 

Aggregation – Bad Example

- Customer link goes down
 - their /23 network becomes unreachable
 - /23 is withdrawn from AS100's iBGP
 - Their ISP doesn't aggregate its /19 network block
 - /23 network withdraw announced to peers
 - start rippling through the Internet
 - added load on all Internet backbone routers as network is removed from routing table
 - Customer link returns
 - Their /23 network is now visible to their ISP
 - Their /23 network is re-advertised to peers
 - Start rippling through the Internet
 - Load on Internet backbone routers as network is reinserted into routing table
 - Internet may take 10-20 min or longer to be visible
 - Where is the Quality of Service???
-

Aggregation - Summary

- Good example is what everyone should do!
 - Added to Internet stability
 - Reduces size of routing table
 - Reduces routing churn
 - Improves Internet QoS for **everyone**
- Bad example is what too many still do!
 - Why? Lack of knowledge? Laziness?

The Internet today (Aug 2005)

- Current Internet Routing Table Statistics

| | |
|--------------------------------------|--------|
| BGP Routing Table Entries | 168367 |
| Prefixes after maximum aggregation | 96812 |
| Unique Prefixes in Internet | 81588 |
| Prefixes smaller than registry alloc | 79815 |
| /24s announced | 91392 |
| ASes in use | 20329 |

Service Providers Multihoming

Why Multihome?

- Redundancy

One connection to internet means the network is dependent on:

Local router (configuration, software, hardware)

WAN media (physical failure, carrier failure)

Upstream Service Provider (configuration, software, hardware)

Why Multihome?

- Reliability

Business critical applications demand continuous availability

Lack of redundancy implies lack of reliability implies loss of revenue

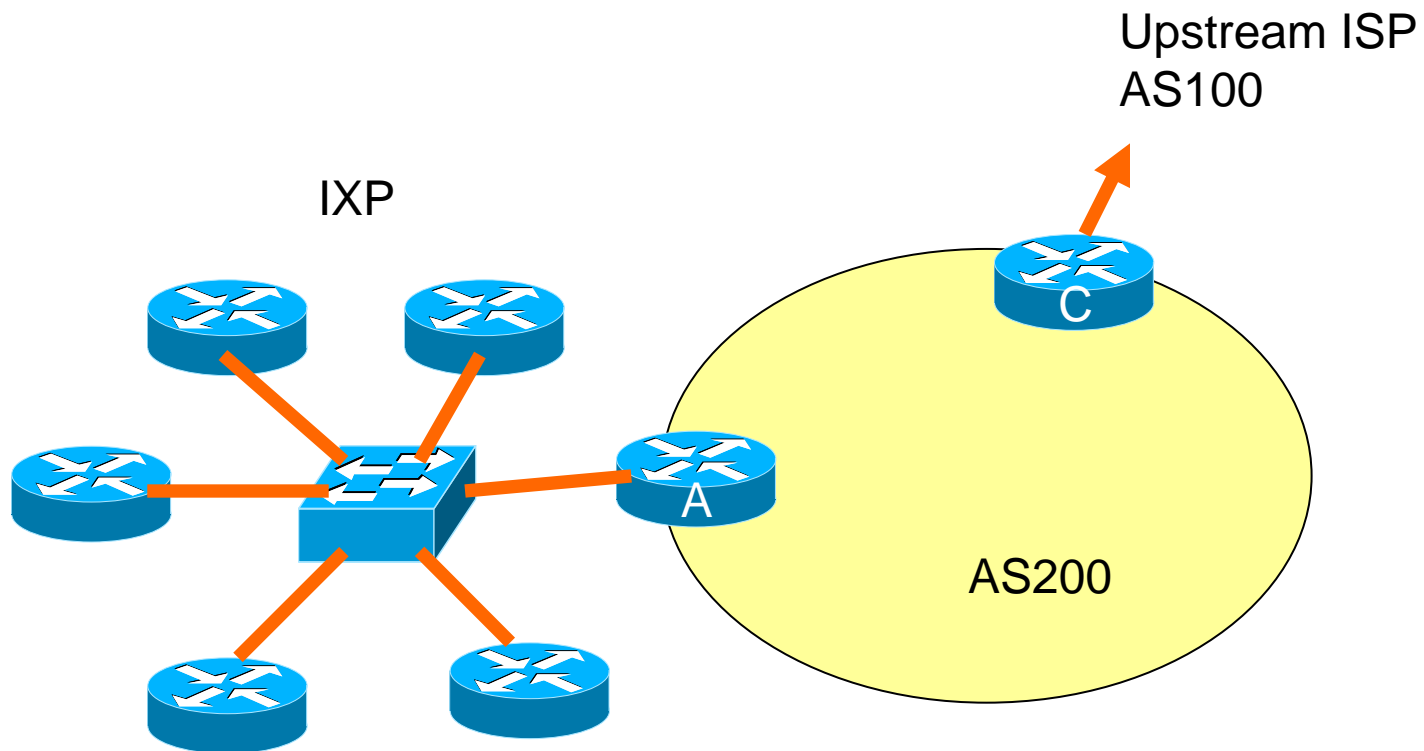
Service Providers Multihoming

One upstream, Local Exchange Point

One upstream, Local Exchange Point

- Announce /19 aggregate to every neighboring AS
- Accept default route only one from upstream
 - Either 0.0.0.0/0 or a network which can be used as default
- Accept all routes from IXP peers

One Upstream, Local Exchange Point



One Upstream, Local Exchange Point

- **Router A Configuration**

```
!  
interface fastethernet 0/0  
  description Exchange Point LAN  
  ip address 220.5.10.1 mask 255.255.255.224  
  ip verify unicast reverse-path  
  no ip directed-broadcast  
  no ip proxy-arp  
  no ip redirect  
!  
router bgp 200  
  network 221.10.0.0 mask 255.255.224.0  
  neighbor IXP-PEERS peer-group  
  neighbor IXP-PEERS soft-reconfiguration inbound  
  neighbor IXP-PEERS prefix-list AS200-CIDR out  
  
..next slide
```

One Upstream, Local Exchange Point

```
neighbor 220.5.10.2 remote-as 101
neighbor 220.5.10.2 peer-group IXP-PEERS
neighbor 220.5.10.2 prefix-list PEER-AS101 in
neighbor 220.5.10.3 remote-as 102
neighbor 220.5.10.3 peer-group IXP-PEERS
neighbor 220.5.10.3 prefix-list PEER-AS102 in
neighbor 220.5.10.4 remote-as 103
neighbor 220.5.10.4 peer-group IXP-PEERS
neighbor 220.5.10.4 prefix-list PEER-AS103 in
neighbor 220.5.10.5 remote-as 104
neighbor 220.5.10.5 peer-group IXP-PEERS
neighbor 220.5.10.5 prefix-list PEER-AS104 in
```

One Upstream, Local Exchange Point

```
ip route 221.10.0.0 255.255.224.0 null0
!
ip prefix-list AS200-CIDR permit 221.10.0.0/19
ip prefix-list PEER-AS101 permit 222.0.0.0/19
ip prefix-list PEER-AS102 permit 222.30.0.0/19
ip prefix-list PEER-AS103 permit 222.12.0.0/19
ip prefix-list PEER-AS104 permit 222.18.128.0/19
!
```

One Upstream, Local Exchange Point

- **Router C Configuration**

```
!  
router bgp 200  
  network 221.10.0.0 mask 255.255.224.0  
  neighbor 222.222.10.1 remote-as 100  
  neighbor 222.222.10.1 prefix-list DEFAULT in  
  neighbor 222.222.10.1 prefix-list AS200-CIDR out  
!  
ip prefix-list AS200-CIDR permit 221.10.0.0/19  
ip prefix-list DEFAULT permit 0.0.0.0/0  
!  
ip route 221.10.0.0 255.255.224.0 null0  
!
```

One Upstream, Local Exchange Point

- Note Router A configuration
 - prefix-list higher maintenance, but safer uRPF on the FastEthernet interface
- IXP traffic goes to and from local IXP, everything else goes to upstream

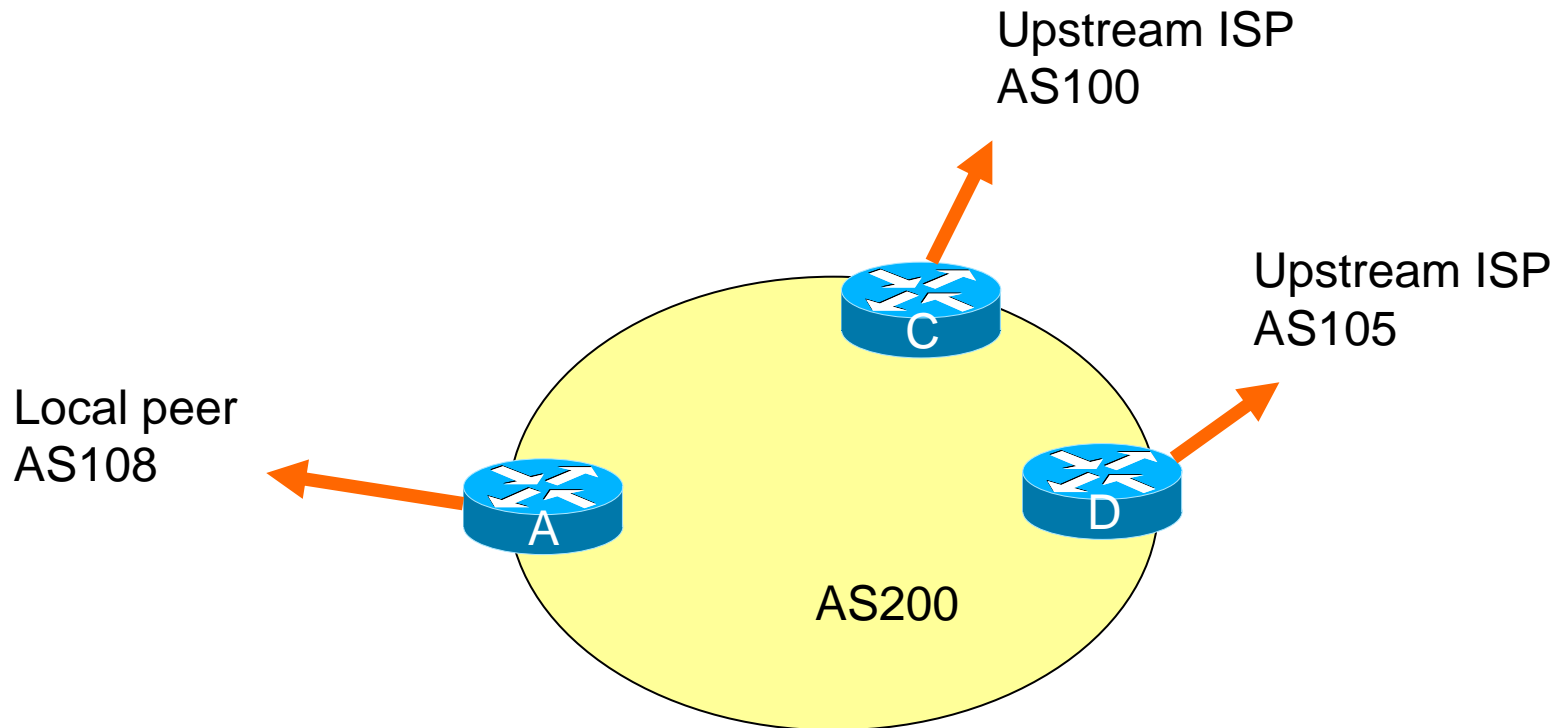
Service Providers Multihoming

Two upstream, One local peer

Two Upstream, One local peer

- Announce /19 aggregate on each link
- Accept default route only from upstreams
 - Either 0.0.0.0/0 or a network which can be used as default
- Accept all routers from local peer

One Upstream, Local Exchange Point



One Upstream, Local Exchange Point

- Router A

Same routing configuration as in example which one upstream and one local peer

Same hardware configuration

Two Upstream, One local peer

- **Router C Configuration**

```
!  
router bgp 200  
network 221.10.0.0 mask 255.255.224.0  
neighbor 222.222.10.1 remote-as 100  
neighbor 222.222.10.1 prefix-list DEFAULT in  
neighbor 222.222.10.1 prefix-list AS200-CIDR out  
!  
ip prefix-list AS200-CIDR permit 222.10.0.0/19  
ip prefix-list DEFAULT permit 0.0.0.0/0  
!  
ip route 221.10.0.0 255.255.224.0 null0  
!
```

Two Upstream, One local peer

- **Router D Configuration**

```
!  
router bgp 109  
network 221.10.0.0 mask 255.255.224.0  
neighbor 222.222.10.5 remote-as 105  
neighbor 222.222.10.5 prefix-list DEFAULT in  
neighbor 222.222.10.5 prefix-list AS200-CIDR out  
!  
ip prefix-list AS200-CIDR permit 222.10.0.0/19  
ip prefix-list DEFAULT permit 0.0.0.0/0  
!  
ip route 221.10.0.0 255.255.224.0 null0  
!
```

Two Upstream, One local peer

- This is the simple configuration for Router C and D
- Traffic out to the two upstreams will take nearest exit

Inexpensive routers required

This is not useful in practice especially for international links

Loadsharing needs to be better

Two Upstream, One local peer

- Better configuration options:

- Accept full routing from both upstreams

- Expensive & unnecessary!

- Accept default from one upstream and some routes from the other upstream

- The way to go!

Two Upstream, One local peer – Full Routes

- **Router C Configuration**

```
!  
router bgp 200  
network 221.10.0.0 mask 255.255.224.0  
neighbor 222.222.10.1 remote-as 100  
=neighbor 222.222.10.1 prefix-list RFC1918-DENY in  
neighbor 222.222.10.1 prefix-list AS200-CIDR out  
neighbor 222.222.10.1 route-map AS100-LOADSHARE in  
!  
ip prefix-list AS200-CIDR permit 222.10.0.0/19  
!ADD RFC1918 deny  
!  
  
..next slide
```

Two Upstream, One local peer – Full Routes

```
!  
ip route 221.10.0.0 255.255.224.0 null0  
!  
ip as-path access-list 10 permit ^(100_)+$  
ip as-path access-list 10 permit ^(100_)+_[0-9]+$  
!  
route-map AS100-LOADSHARE permit 10  
  match ip as-path 10  
  set local-preference 120  
route-map AS100-LOADSHARE permit 10  
  set local-preference 80  
!
```

Two Upstream, One local peer – Full Routes

Router D Configuration

```
!  
router bgp 200  
  network 221.10.0.0 mask 255.255.224.0  
  neighbor 222.222.10.5 remote-as 105  
  neighbor 222.222.10.5 prefix-list DEFAULT in  
  neighbor 222.222.10.5 prefix-list AS200-CIDR out  
!  
ip prefix-list AS200-CIDR permit 221.10.0.0/19  
ip prefix-list DEFAULT permit 0.0.0.0/0  
!  
ip route 221.10.0.0 255.255.224.0 null0  
!
```


Two Upstream, One local peer – Full Routes

- Router C configuration:
 - Accept full route from AS100
 - Tag prefixes originated by AS100 and AS100's neighboring ASes with local preference 120
 - Traffic to those ASes will go over AS100 link
 - Remaining prefixes tagged with local preference 80
 - Traffic to other all other ASes will go over the link to AS105
- Router D configuration same as Router C without the route-map

Two Upstream, One local peer – Full Routes

- Full routes from upstreams
 - Expensive – needs 256Mbytes RAM today
 - Need to play preference games
 - Previous example is only an example – real life will need improve fine-tuning!
 - Previous example doesn't consider inbound

Configuration Tips

iBGP and IGP

- Make sure loopback is configured on router
iBGP between loopbacks, NOT real interface
- Make sure IGP carries loopback /32 address
- Make sure IGP carries DMZ nets
Use ip-unnumbered where possible
Or use next-hop-self on iBGP neighbors
`neighbor x.x.x.x next-hop-self`

Next-hop-self

- Used by many ISPs on edge routers

Preferable to carrying DMZ /30 addresses in the IGP

Reduces size of IGP to just core infrastructure

Alternative to using `ip unnumbered`

helps scale network

BGP speaker announces external network using local address (loopback) as next-hop

Templates

- Good practice to configure templates for everything

Vendor defaults tend not to be optimal or even very useful for ISPs

ISPs create their own defaults by using configuration templates

Sample iBGP and eBGP templates follow for Cisco IOS

iBGP template



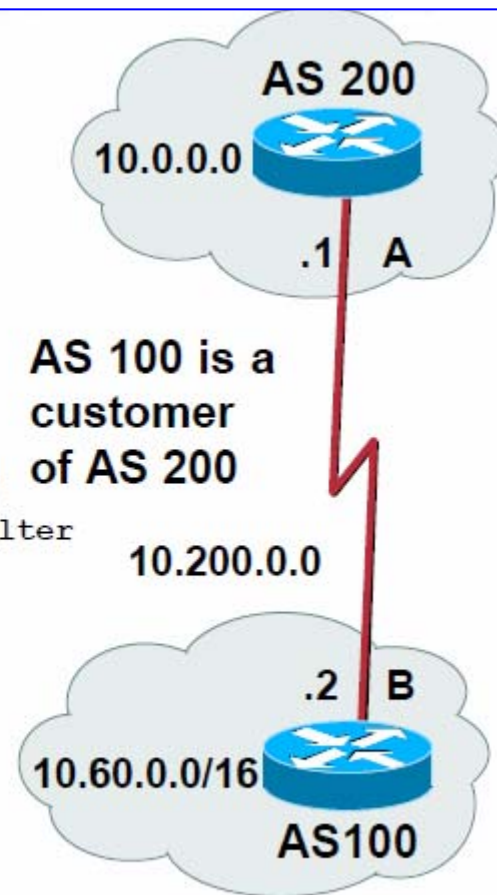
```
router bgp 100
neighbor internal peer-group
neighbor internal description ibgp peers
neighbor internal remote-as 100
neighbor internal update-source Loopback0
neighbor internal next-hop-self
neighbor internal send-community
neighbor internal version 4
neighbor internal password 7 03085A09
neighbor 1.0.0.1 peer-group internal
neighbor 1.0.0.2 peer-group internal
```

iBGP template

- Use peer-groups
- iBGP between loopbacks!
- Next-hop-self
 - Keep DMZ and point-to-point out of IGP
- Always send communities in iBGP
 - Otherwise accidents will happen
- Hardware BGP to version 4
 - Yes, this is things paranoid!
- Use passwords on iBGP session
 - Not being paranoid, **VERY** necessary

eBGP template

```
Router B:
router bgp 100
  bgp dampening route-map RIPE229-flap
  network 10.60.0.0 mask 255.255.0.0
  neighbor external peer-group
  neighbor external remote-as 200
  neighbor external description ISP connection
  neighbor external remove-private-AS
  neighbor external version 4
  neighbor external prefix-list ispout out ! "real" filter
  neighbor external filter-list 1 out      ! "accident" filter
  neighbor external route-map ispout out
  neighbor external prefix-list ispin in
  neighbor external filter-list 2 in
  neighbor external route-map ispin in
  neighbor external password 7 020A0559
  neighbor external maximum-prefix 120000 [warning-only]
  neighbor 10.200.0.1 peer-group external
  !
ip route 10.60.0.0 255.255.0.0 null0 254
```



eBGP template

- BGP damping – use RIPE-229 parameters
- Remove private ASes from announcements
 - Common omission today
- Use extensive filters, with “backup”
 - Use as-path filters to back-up prefix-lists
 - Use route-map for policy
- Use password agreed between you and peer on a eBGP session
- Use maximum-prefix tracking
 - Router will warn you if there are sudden changes in BGP table size, bringing down eBGP if desired

More BGP “defaults”

- Log neighbor changes
`bgp log-neighbor-changes`
- Enable deterministic MED
`bgp deterministic-med`
Otherwise bestpath could be different every time BGP session is reset
- Make BGP admin distance higher than any IGP
`distance bgp 200 200 200`

BGP Techniques for ISP

End of Tutorial

Many Thanks to Philip Smith pfs@cisco.com