



Deploying BGP

Philip Smith <pfs@cisco.com>

APRICOT 2005

Kyoto, Japan

Presentation Slides

- **Slides are at:**

**[ftp://ftp-eng.cisco.com
/pfs/seminars/APRICOT2005-Deploying-BGP.pdf](ftp://ftp-eng.cisco.com/pfs/seminars/APRICOT2005-Deploying-BGP.pdf)**

And on the APRICOT2005 website

- **Feel free to ask questions any time**

BGP for Internet Service Providers

- **Scaling BGP**
- **Using Communities**
- **Deploying BGP in an ISP network**

BGP Scaling Techniques

BGP Scaling Techniques

- **How does a service provider:**

Scale the iBGP mesh beyond a few peers?

Implement new policy without causing flaps and route churning?

Keep the network stable, scalable, as well as simple?

- **Three Techniques:**

Route Refresh, Flap Damping, Route Reflector

Route Refresh

Route Refresh

Problem:

- **Hard BGP peer reset required after every policy change because the router does not store prefixes that are rejected by policy**
- **Hard BGP peer reset:**
 - Tears down BGP peering**
 - Consumes CPU**
 - Severely disrupts connectivity for all networks**

Solution:

- **Route Refresh**

Route Refresh Capability

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

Dynamic Reconfiguration

- **Use Route Refresh capability if supported**
find out from the BGP neighbour status display
Non-disruptive, “Good For the Internet”
- **If not supported, see if implementation has a workaround**
- **Only hard-reset a BGP peering as a last resort**

Consider the impact to be equivalent to a router reboot

Route Flap Damping

Stabilising the Network

Route Flap Damping

- **Route flap**

Going up and down of path or change in attribute

BGP WITHDRAW followed by UPDATE = 1 flap

eBGP neighbour peering reset is NOT a flap

Ripples through the entire Internet

Wastes CPU

- **Damping aims to reduce scope of route flap propagation**

Route Flap Damping (continued)

- **Requirements**

Fast convergence for normal route changes

History predicts future behaviour

Suppress oscillating routes

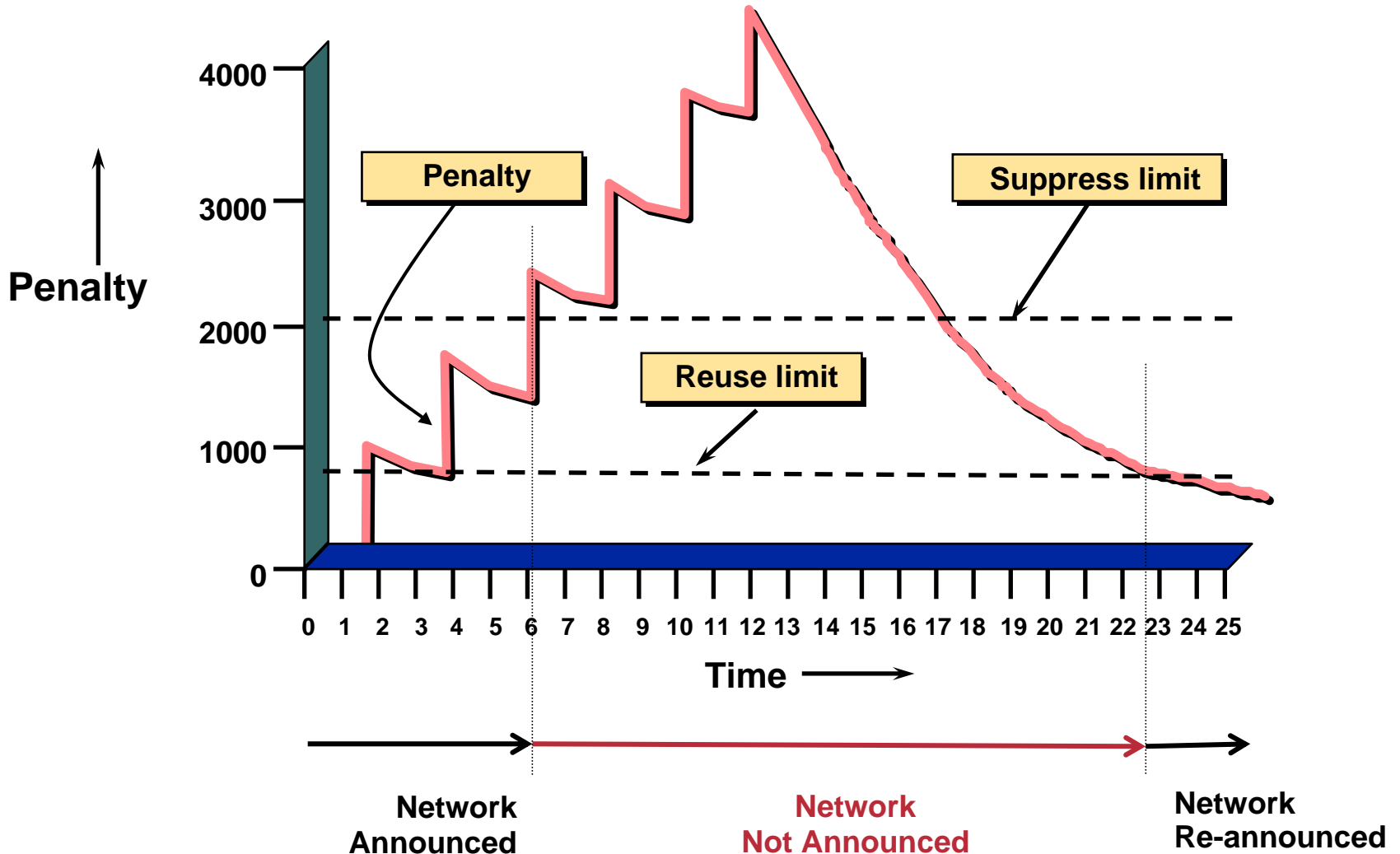
Advertise stable routes

- **Documented in RFC2439**

Operation

- **Add penalty for each flap**
NB: Change in attribute can also be 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

Operation



Operation

- **Only applied to inbound announcements from eBGP peers**
- **Alternate paths still usable**
- **Controllable by at least:**
 - Half-life**
 - reuse-limit**
 - suppress-limit**
 - maximum suppress time**

Configuration

- **Implementations allow various policy control with flap damping**

Fixed damping, same rate applied to all prefixes

Variable damping, different rates applied to different ranges of prefixes

- **Recommendations for ISPs**

<http://www.ripe.net/docs/ripe-229.html>

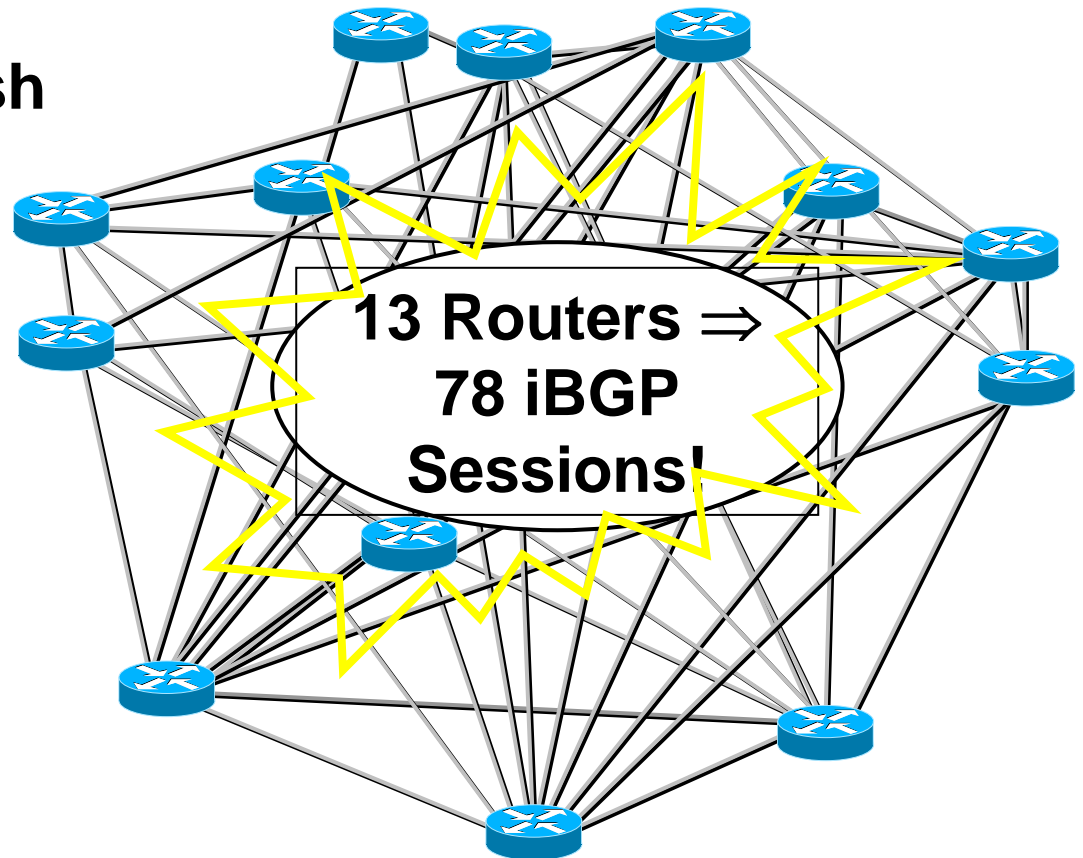
(work by European and US ISPs a few years ago as vendor defaults were considered to be too aggressive)

Route Reflectors

Scaling iBGP mesh

Avoid $\frac{1}{2}n(n-1)$ iBGP mesh

**$n=1000 \Rightarrow$ nearly
half a million
ibgp sessions!**

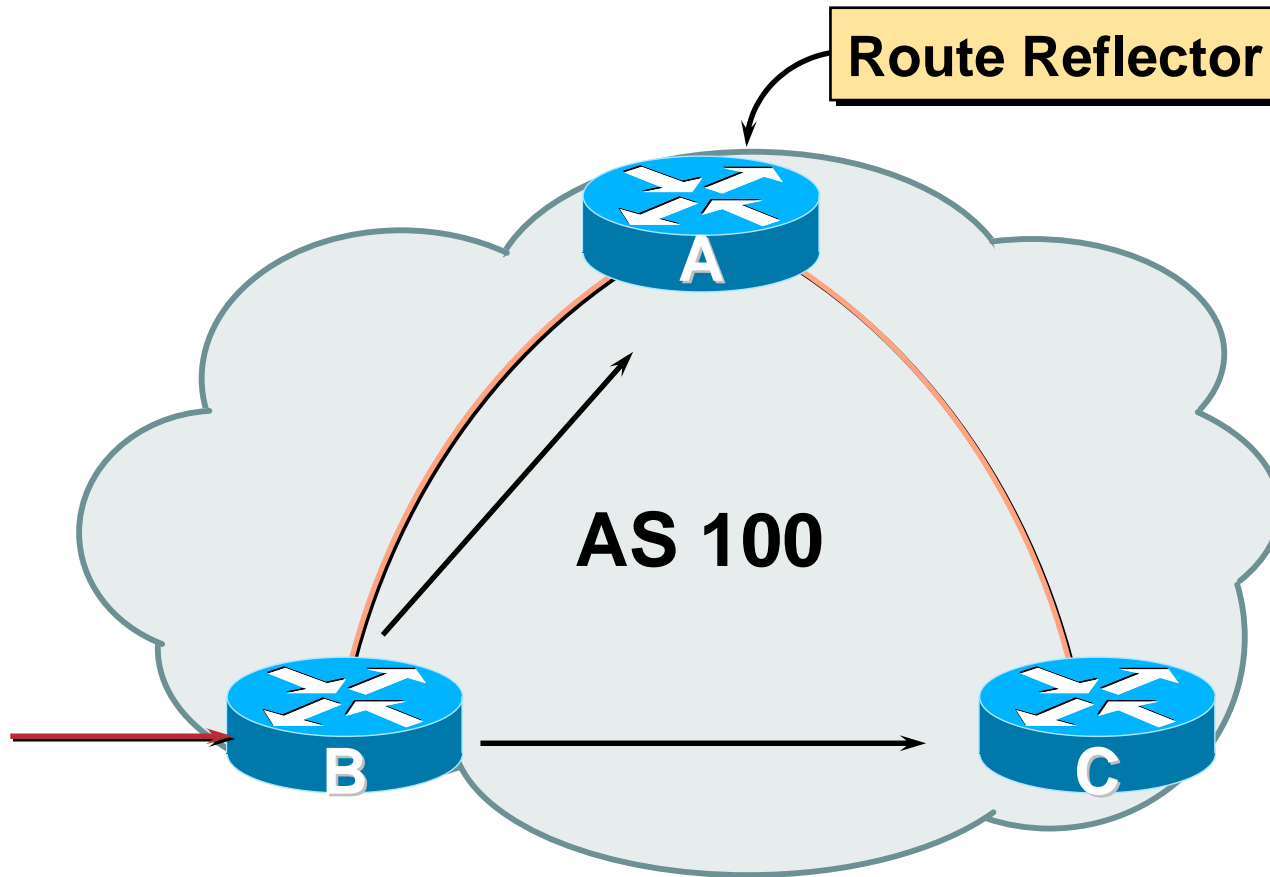


Two solutions

Route reflector – simpler to deploy and run

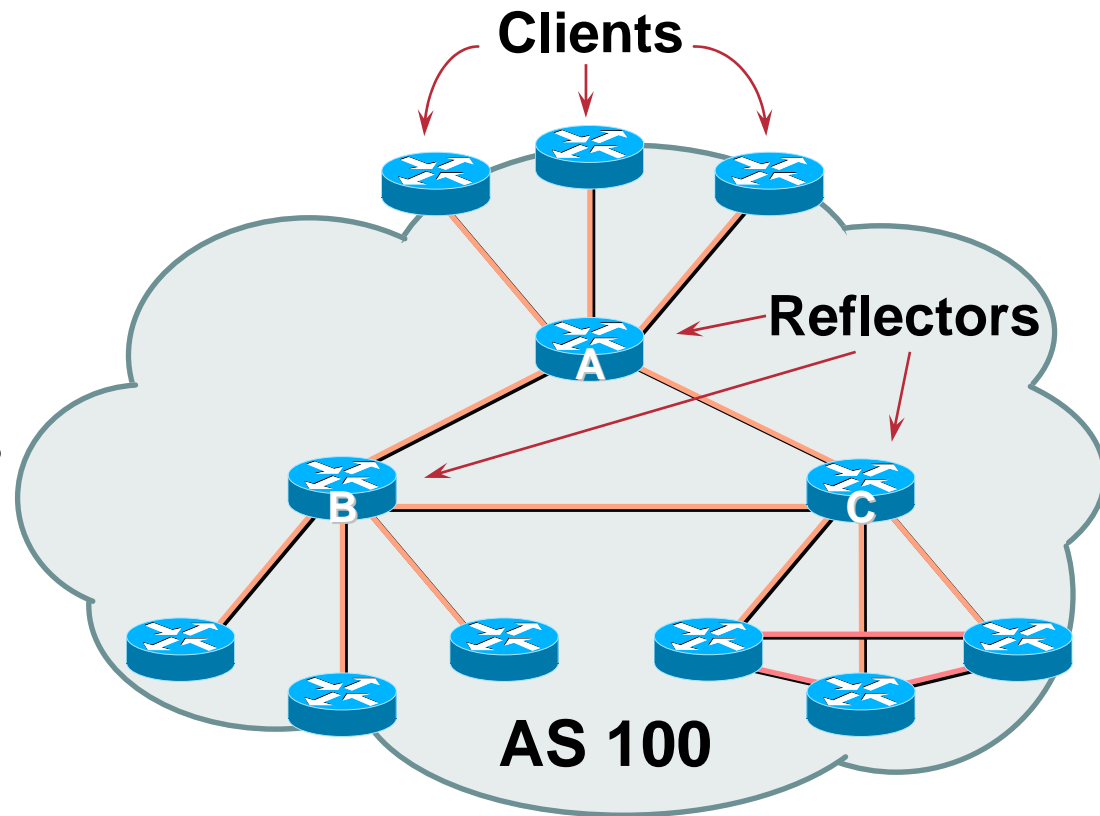
Confederation – more complex, corner case benefits

Route Reflector: Principle



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 by the RR

Best to set cluster-id is from router-id (address of loopback)

(Some ISPs use their own cluster-id assignment strategy – but needs to be well documented!)

Route Reflectors: Redundancy

- **Multiple RRs can be configured in the same cluster – not advised!**

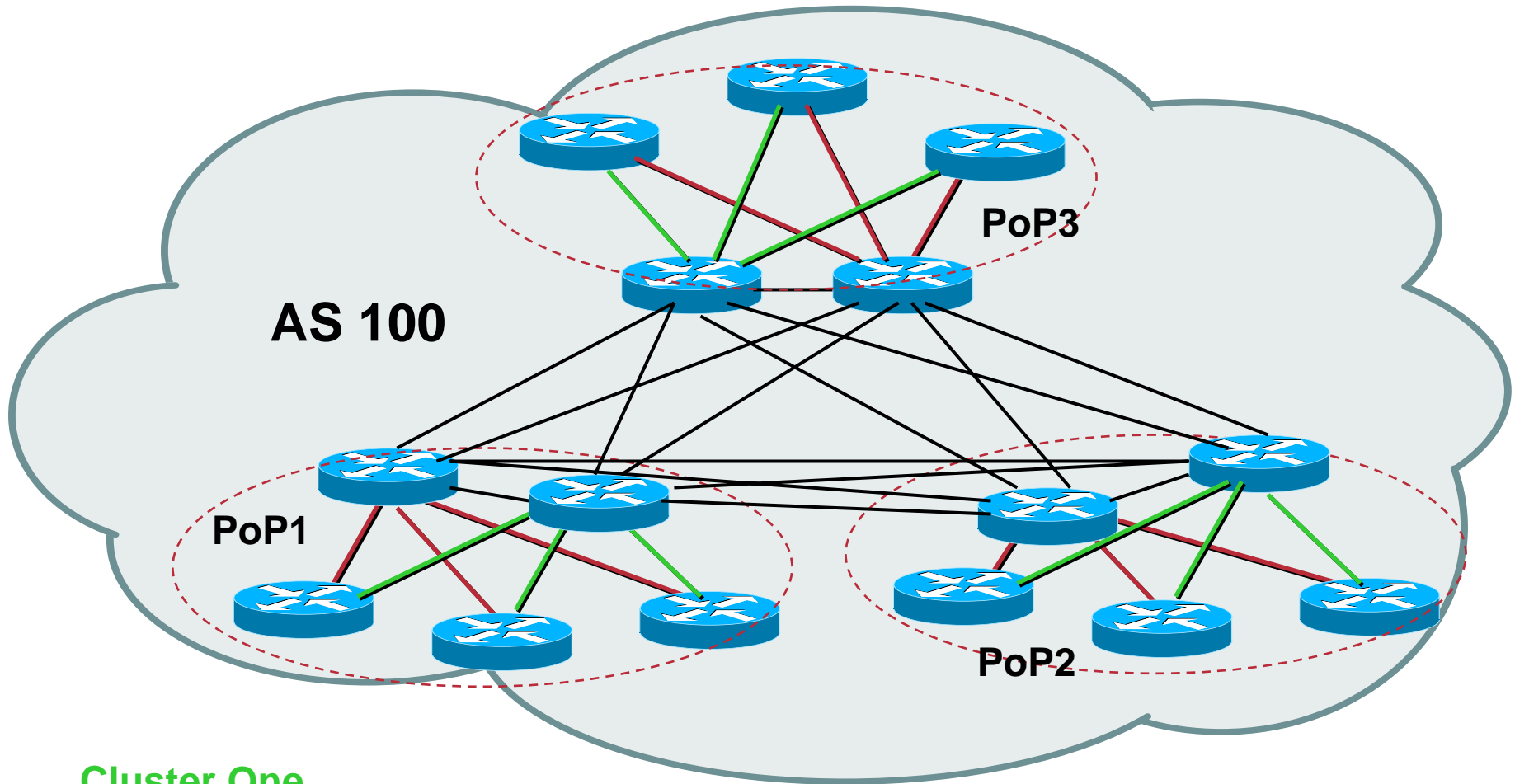
All RRs in the cluster **must** have the same cluster-id (otherwise it is a different cluster)

- **A router may be a client of RRs in different clusters**

Common today in ISP networks to overlay two clusters – redundancy achieved that way

→ Each client has two RRs = redundancy

Route Reflectors: Redundancy



Cluster One

Cluster Two

Route Reflectors: Migration

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

Always follow the physical topology!

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

- **Typical ISP network:**

PoP has two core routers

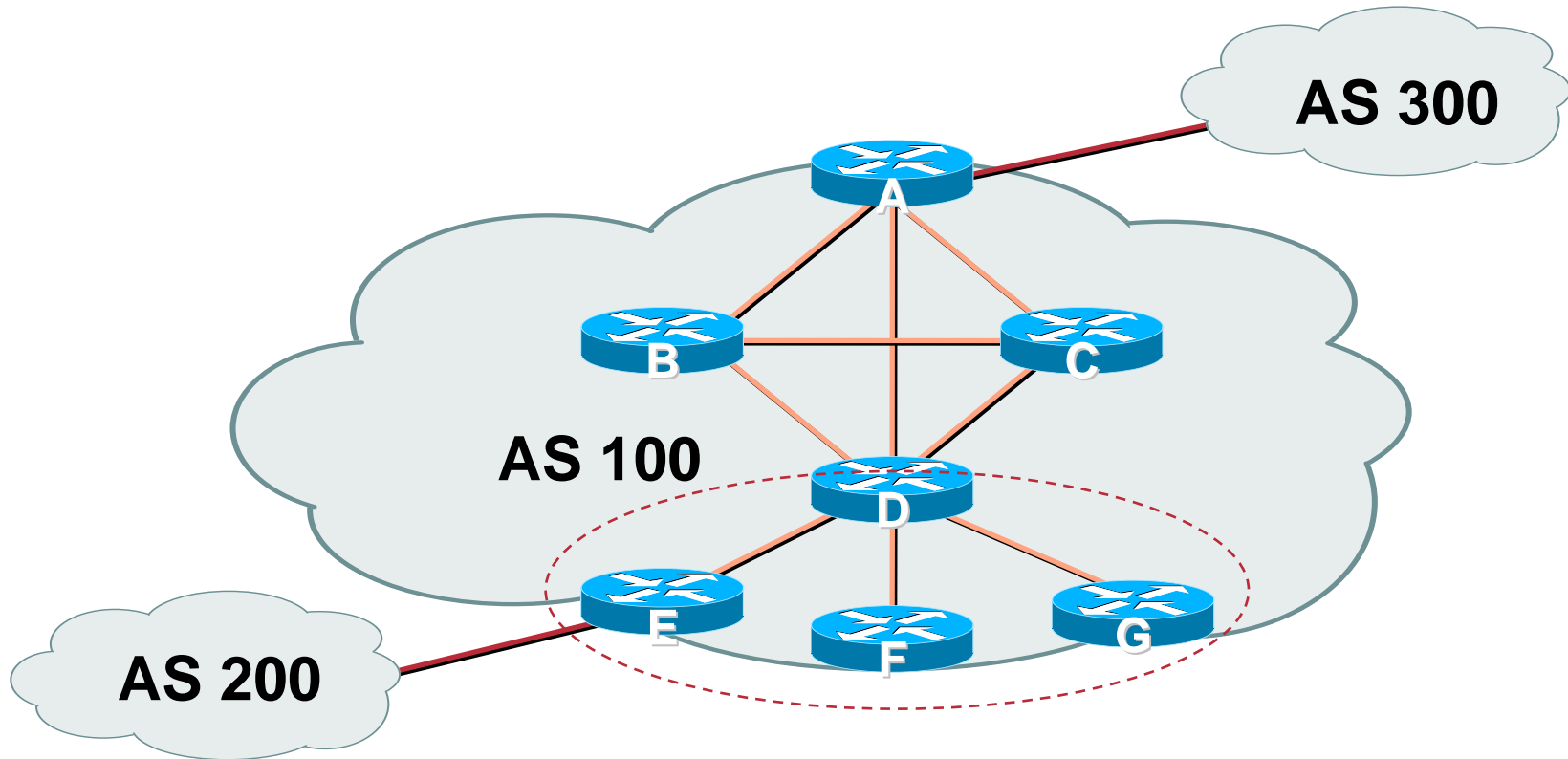
Core routers are RR for the PoP

Two overlaid clusters

Route Reflectors: Migration

- **Typical ISP network:**
 - Core routers have fully meshed iBGP**
 - Create further hierarchy if core mesh too big**
 - Split backbone into regions**
- **Configure one cluster pair at a time**
 - Eliminate redundant iBGP sessions**
 - Place maximum one RR per cluster**
 - Easy migration, multiple levels**

Route Reflector: Migration



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

BGP for Internet Service Providers

- **Scaling BGP**
- **Using Communities**
- **Deploying BGP in an ISP network**

Service Providers use of Communities

Some examples 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**

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 customers might be

- **Two simple examples follow to explain the concept**

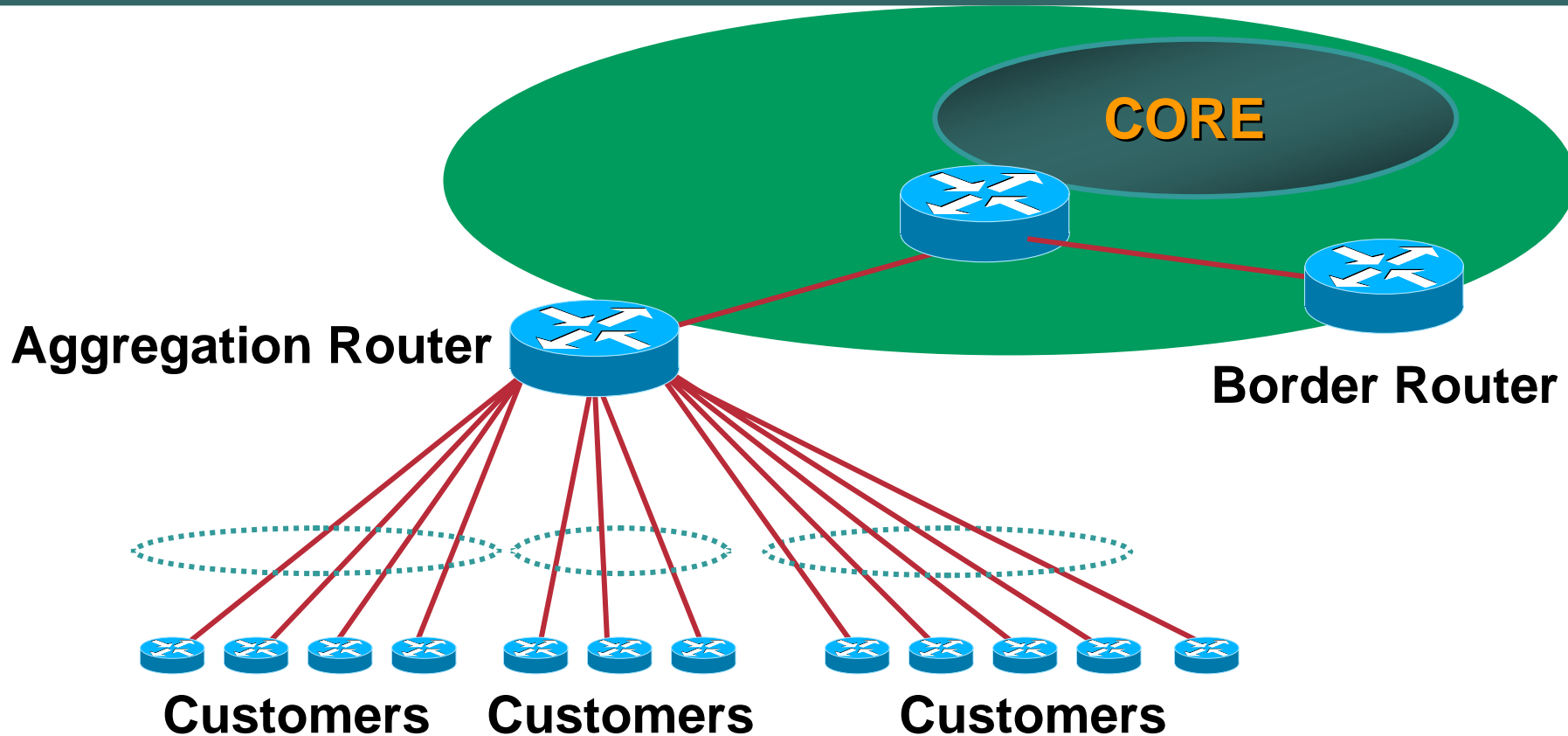
Community Example – Customer Edge

- **This demonstrates how communities might be used at the customer edge of an ISP network**
- **ISP has three connections to the Internet:**
 - IXP connection, for local peers**
 - Private peering with a competing ISP in the region**
 - Transit provider, who provides visibility to the entire Internet**
- **Customers have the option of purchasing combinations of the above connections**

Community Example – Customer Edge

- **Community assignments:**
 - IXP connection: community 100:2100**
 - Private peer: community 100:2200**
- **Customer who buys local connectivity (via IXP) is put in community 100:2100**
- **Customer who buys peer connectivity is put in community 100:2200**
- **Customer who wants both IXP and peer connectivity is put in 100:2100 and 100:2200**
- **Customer who wants “the Internet” has no community set**
 - We are going to announce his prefix everywhere**

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!

Community Example – Internet Edge

- **This demonstrates how communities might be used at the peering edge of an ISP network**
- **ISP has four types of BGP peers:**
 - Customer**
 - IXP peer**
 - Private peer**
 - Transit provider**
- **The prefixes received from each can be classified using communities**
- **Customers can opt to receive any or all of the above**

Community Example – Internet Edge

- **Community assignments:**

Customer prefix:	community 100:3000
IXP prefix:	community 100:3100
Private peer prefix:	community 100:3200
- **BGP customer who buys local connectivity gets 100:3000**
- **BGP customer who buys local and IXP connectivity receives community 100:3000 and 100:3100**
- **BGP customer who buys full peer connectivity receives community 100:3000, 100:3100, and 100:3200**
- **Customer who wants “the Internet” gets everything**
 - Gets default route originated by aggregation router**
 - Or pays money to get all 135k prefixes**

Community Example – Internet Edge

- **No need to create customised filters when adding customers**

Border router already sets communities

Installation engineers pick the appropriate community set when establishing the customer BGP session

⇒ Ease of operation!

Community Example – Summary

- **Two examples of customer edge and internet edge can be combined to form a simple community solution for ISP prefix policy control**
- **More experienced operators tend to have more sophisticated options available**

Advice is to start with the easy examples given, and then proceed onwards as experience is gained

Some ISP Examples

- **ISPs also create communities to give customers bigger routing policy control**
- **Public policy is usually listed in the IRR**

Following examples are all in the IRR

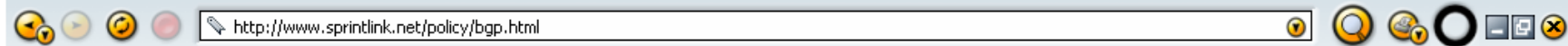
Examples build on the configuration concepts from the introductory example

- **Consider creating communities to give policy control to customers**

Reduces technical support burden

Reduces the amount of router reconfiguration, and the chance of mistakes

Some ISP Examples: Sprintlink



WHAT YOU CAN CONTROL

AS-PATH PREPENDS

Sprint allows customers to use AS-path prepending to adjust route preference on the network. Such prepending will be received and passed on properly without notifying Sprint of your change in announcements.

Additionally, Sprint will prepend AS1239 to eBGP sessions with certain autonomous systems depending on a received community. Currently, the following ASes are supported: 1668, 209, 2914, 3300, 3356, 3549, 3561, 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 America
65010:XXX	Do not advertise to ASXXX
65011:XXX	1239 (default) ...
65012:XXX	1239 1239 ...
65013:XXX	1239 1239 1239 ...
65014:XXX	1239 1239 1239 1239 ...
String	Resulting AS Path to all supported ASes
65000:0	Do not advertise
65001:0	1239 (default) ...
65002:0	1239 1239 ...
65003:0	1239 1239 1239 ...
65004:0	1239 1239 1239 1239 ...

More info at
www.sprintlink.net/policy/bgp.html

Some ISP Examples

Connect.com.au

```
aut-num:      AS2764
descr:        connect.com.au Pty Ltd
remarks:      Community Definition
remarks:      -----
remarks:      2764:1 Announce to "domestic" rate ASes only
remarks:      2764:2 Don't announce outside local POP
remarks:      2764:3 Lower local preference by 25
remarks:      2764:4 Lower local preference by 15
remarks:      2764:5 Lower local preference by 5
remarks:      2764:6 Announce to non customers with "no-export"
remarks:      2764:7 Only announce route to customers
remarks:      2764:8 Announce route over satellite link
notify:       routing@connect.com.au
mnt-by:       CONNECT-AU
changed:      mrp@connect.com.au 19990506
source:       CCAIR
```

More at <http://info.connect.com.au/docs/routing/general/multi-faq.shtml#q13>

Some ISP Examples


MCI Europe

```
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.

<snip>

Additional details of the MCI communities are located at:
http://global.mci.com/uk/customer/bgp/


mnt-by: WCOM-EMEA-RICE-MNT
changed: rice@lists.mci.com 20041006
source: RIPE
```



Some ISP Examples

BT Ignite

```
aut-num:      AS5400
descr:        BT Ignite European Backbone
remarks:
remarks:      Community to
remarks:      Not announce      To peer:      Community to
remarks:                                             AS prepend 5400
remarks:      5400:1000 All peers & Transits      5400:2000
remarks:
remarks:      5400:1500 All Transits      5400:2500
remarks:      5400:1501 Sprint Transit (AS1239)      5400:2501
remarks:      5400:1502 SAVVIS Transit (AS3561)      5400:2502
remarks:      5400:1503 Level 3 Transit (AS3356)      5400:2503
remarks:      5400:1504 AT&T Transit (AS7018)      5400:2504
remarks:      5400:1505 UUnet Transit (AS701)      5400:2505
remarks:
remarks:      5400:1001 Nexica (AS24592)      5400:2001
remarks:      5400:1002 Fujitsu (AS3324)      5400:2002
remarks:      5400:1003 Unisource (AS3300)      5400:2003
<snip>
notify:       notify@eu.bt.net
mnt-by:       CIP-MNT
source:       RIPE
```




**And many
many more!**

Some ISP Examples

Carrier1

```
aut-num:      AS8918
descr:        Carrier1 Autonomous System
<snip>
remarks:      Community   Definition
remarks:      *
remarks:      8918:2000    Do not announce to C1 customers
remarks:      8918:2010    Do not announce to C1 peers, peers+ and transit
remarks:      8918:2015    Do not announce to C1 transit providers
remarks:      *
remarks:      8918:2020    Do not announce to Teleglobe (AS 6453)
remarks:      8918:2035    Do not announce to UUNet      (AS 702)
remarks:      8918:2040    Do not announce to Cogent      (AS 174)
remarks:      8918:2050    Do not announce to T-Systems (AS 3320)
remarks:      8918:2060    Do not announce to Sprint      (AS 1239)
remarks:      *
remarks:      8918:2070    Do not announce to AMS-IX peers
remarks:      8918:2080    Do not announce to NL-IX peers
remarks:      8918:2090    Do not announce to Packet Exchange Peers
<snip>
notify:        inoc@carrier1.net
mnt-by:        CARRIER1-MNT
source:        RIPE
```




**And many
many more!**

Some ISP Examples

Level 3

```
aut-num:      AS3356
descr:        Level 3 Communications
<snip>
remarks:      -----
remarks:      customer traffic engineering communities - Suppression
remarks:      -----
remarks:      64960:XXX - announce to AS XXX if 65000:0
remarks:      65000:0   - announce to customers but not to peers
remarks:      65000:XXX - do not announce at peerings to AS XXX
remarks:      -----
remarks:      customer traffic engineering communities - Prepending
remarks:      -----
remarks:      65001:0   - prepend once   to all peers
remarks:      65001:XXX - prepend once   at peerings to AS XXX
remarks:      65002:0   - prepend twice  to all peers
remarks:      65002:XXX - prepend twice  at peerings to AS XXX
remarks:      65003:0   - prepend 3x     to all peers
remarks:      65003:XXX - prepend 3x     at peerings to AS XXX
remarks:      65004:0   - prepend 4x     to all peers
remarks:      65004:XXX - prepend 4x     at peerings to AS XXX
<snip>
mnt-by:        LEVEL3-MNT
source:        RIPE
```



And many
many more!

BGP for Internet Service Providers

- **Scaling BGP**
- **Using Communities**
- **Deploying BGP in an ISP network**

Deploying BGP in an ISP Network

Okay, so we've learned all about BGP now; how do we use it on our network??

Deploying BGP

- **The role of IGPs and iBGP**
- **Aggregation**
- **Receiving Prefixes**
- **Configuration Tips**

The role of IGP and iBGP

Ships in the night?

Or

Good foundations?

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

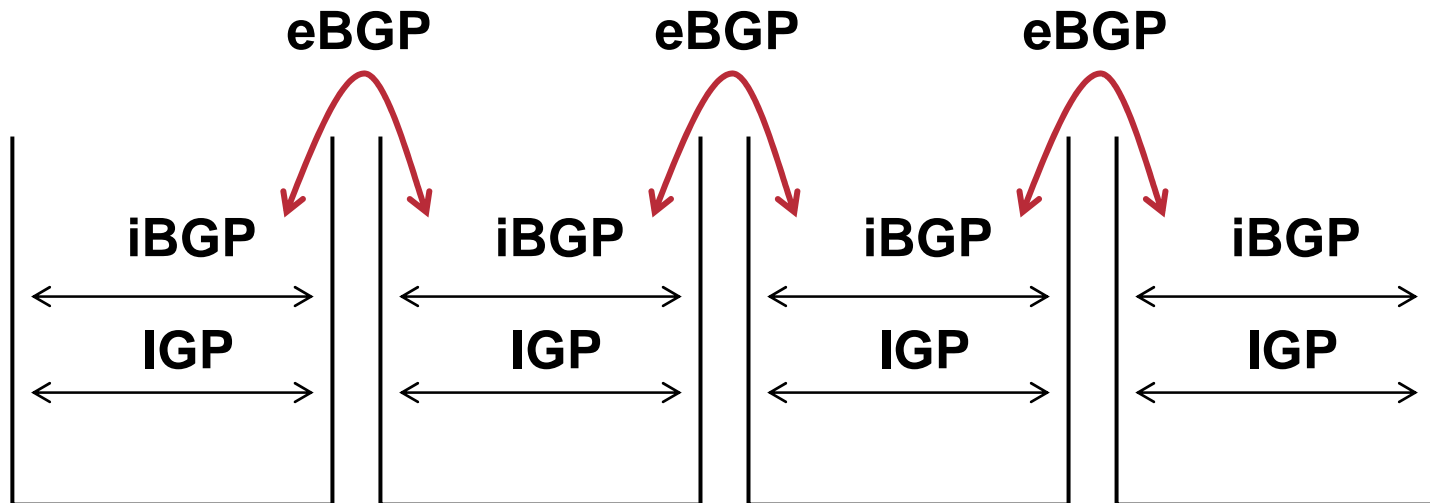
design goal is to **minimise** number of prefixes in IGP to aid scalability and rapid convergence

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/IGP model used in ISP networks

- **Model representation**



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**

Injecting 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 flaps

Aggregation

Quality or Quantity?

Aggregation

- **Aggregation means announcing the address block received from the RIR to the other ASes connected to 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**

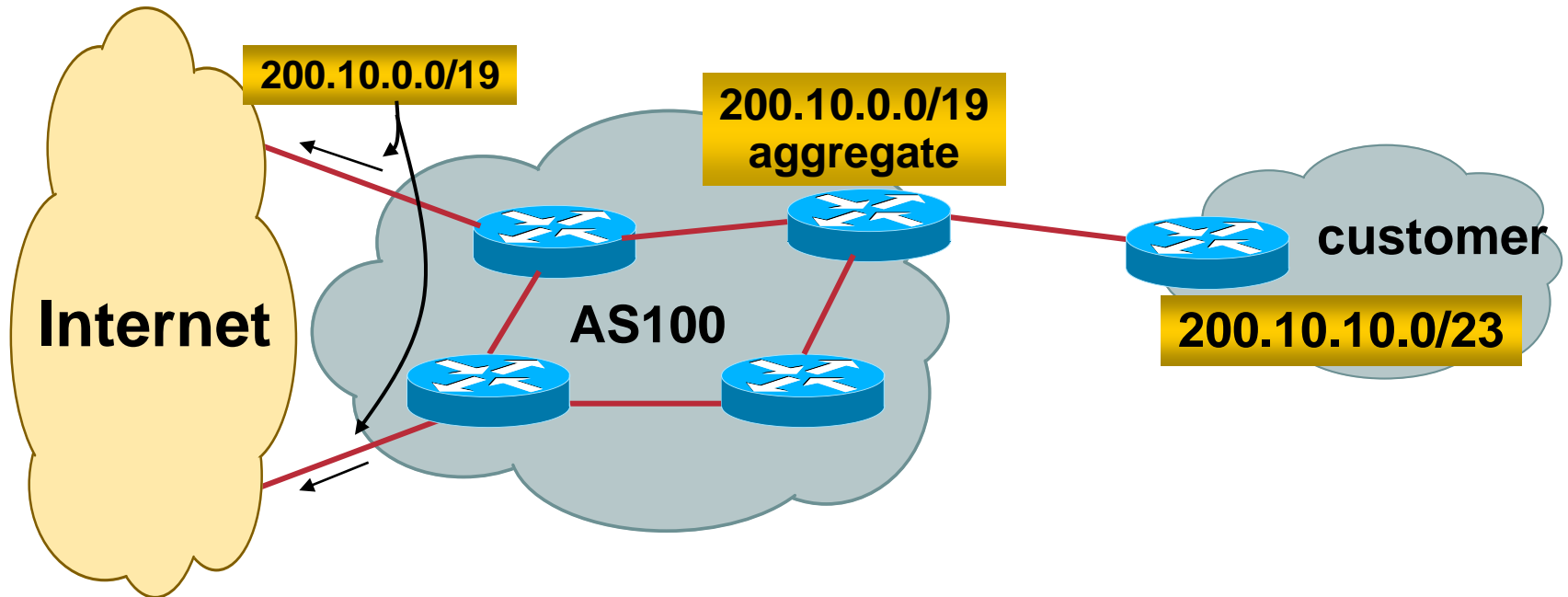
Aggregation

- 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 Aggregate

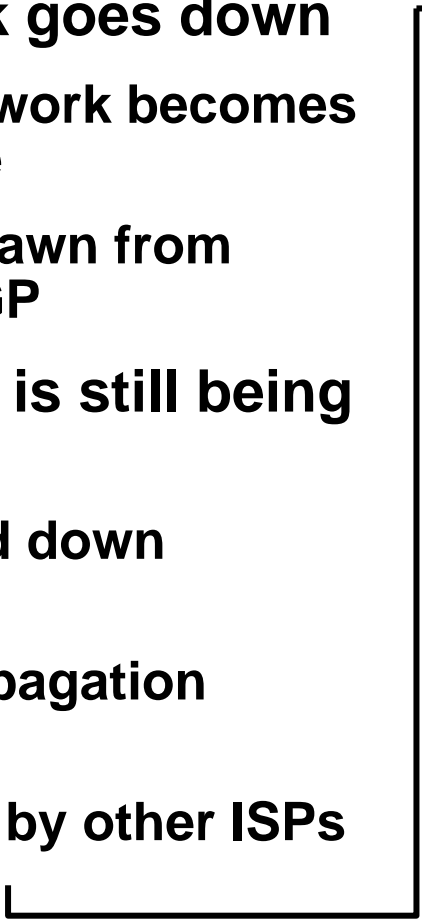
- **ISPs who don't and won't aggregate are held in poor regard by community**
- **Registries publish their minimum allocation size**
 - Anything from a /20 to a /22 depending on RIR**
 - Different sizes for different address blocks**
- **No real reason to see anything longer than a /22 prefix in the Internet**
 - BUT there are currently >84000 /24s!**

Aggregation – Example

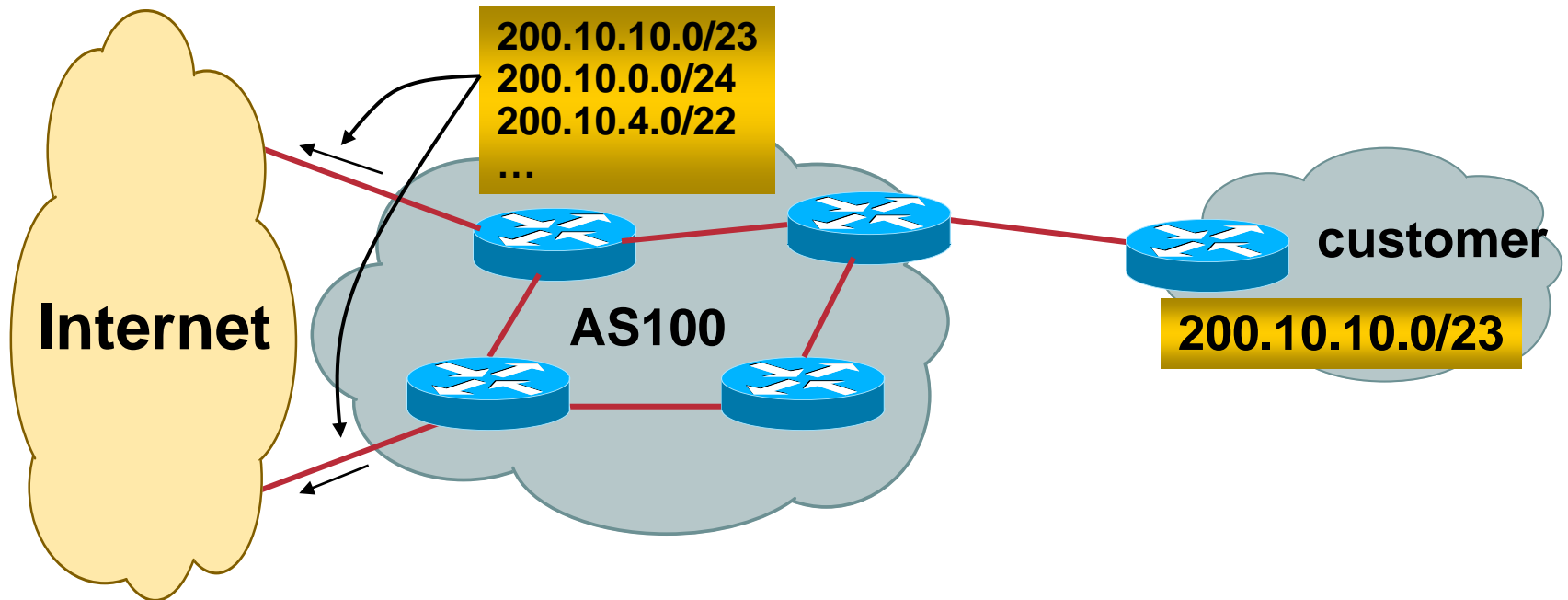


- Customer has /23 network assigned from AS100's /19 address block
- AS100 announced /19 aggregate to the Internet

Aggregation – Good Example

- 
- **Customer link goes down**
their /23 network becomes unreachable
/23 is withdrawn from AS100's iBGP
 - **/19 aggregate is still being 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 – Example



- Customer has /23 network assigned from AS100's /19 address block
- AS100 announces customers' individual networks to the Internet

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 withdrawal announced to peers**
 - starts 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**
 - Starts rippling through Internet**
 - Load on Internet backbone routers as network is reinserted into routing table**
 - Some ISP's suppress the flaps**
 - 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!**
 - Adds 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 (February 2005)

- **Current Internet Routing Table Statistics**

BGP Routing Table Entries	154984
----------------------------------	---------------

Prefixes after maximum aggregation	90381
---	--------------

Unique prefixes in Internet	74096
------------------------------------	--------------

Prefixes smaller than registry alloc	72278
---	--------------

/24s announced	84524
-----------------------	--------------

only 5646 /24s are from 192.0.0.0/8

ASes in use	18880
--------------------	--------------

Efforts to improve aggregation

- **The CIDR Report**

Initiated and operated for many years by Tony Bates

Now combined with Geoff Huston's routing analysis

www.cidr-report.org

Results e-mailed on a weekly basis to most operations lists around the world

Lists the top 30 service providers who could do better at aggregating

Efforts to improve aggregation

The CIDR Report

- Also computes the size of the routing table assuming ISPs performed optimal aggregation
- Website allows searches and computations of aggregation to be made on a per AS basis

flexible and powerful tool to aid ISPs

Intended to show how greater efficiency in terms of BGP table size can be obtained without loss of routing and policy information

Shows what forms of origin AS aggregation could be performed and the potential benefit of such actions to the total table size

Very effectively challenges the traffic engineering excuse

Aggregation Summary

- Aggregation on the Internet could be **MUCH** better

35% saving on Internet routing table size is quite feasible

Tools **are** available

Commands on the router are not hard

CIDR-Report webpage

Receiving Prefixes

Receiving Prefixes

- **There are three scenarios for receiving prefixes from other ASNs**
 - Customer talking BGP**
 - Peer talking BGP**
 - Upstream/Transit talking BGP**
- **Each has different filtering requirements and need to be considered separately**

Receiving Prefixes: From Customers

- ISPs should only accept prefixes which have been assigned or allocated to their downstream customer
- If ISP has assigned address space to its customer, then the customer **IS** entitled to announce it back to his ISP
- If the ISP has **NOT** assigned address space to its customer, then:

Check in the four RIR databases to see if this address space really has been assigned to the customer

The tool: **whois** -h whois.apnic.net x.x.x.0/24

Receiving Prefixes: From Customers

- Example use of whois to check if customer is entitled to announce address space:

```
pfs-pc$ whois -h whois.apnic.net 202.12.29.0
inetnum:      202.12.29.0 - 202.12.29.255
netname:      APNIC-AP-AU-BNE
descr:        APNIC Pty Ltd - Brisbane Offices + Servers
descr:        Level 1, 33 Park Rd
descr:        PO Box 2131, Milton
descr:        Brisbane, QLD.
country:      AU
admin-c:      HM20-AP
tech-c:       NO4-AP
mnt-by:       APNIC-HM
changed:      hm-changed@apnic.net 20030108
status:       ASSIGNED PORTABLE
source:       APNIC
```

Portable – means its an assignment to the customer, the customer can announce it to you

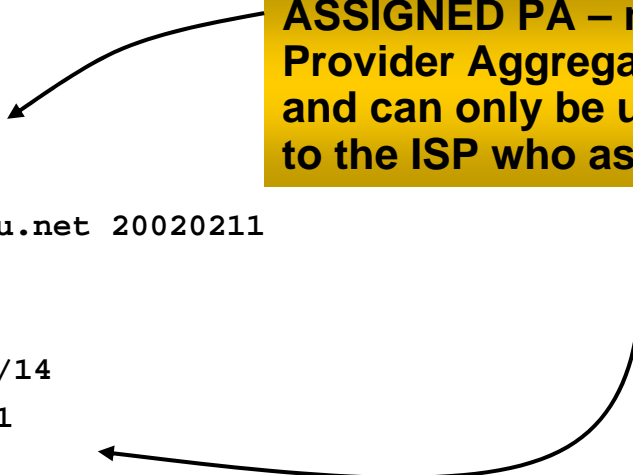
Receiving Prefixes: From Customers

- Example use of whois to check if customer is entitled to announce address space:

```
$ whois -h whois.ripe.net 193.128.2.0
inetnum:      193.128.2.0 - 193.128.2.15
descr:        Wood Mackenzie
country:      GB
admin-c:      DB635-RIPE
tech-c:       DB635-RIPE
status:       ASSIGNED PA
mnt-by:       AS1849-MNT
changed:      dauids@uk.uu.net 20020211
source:       RIPE

route:         193.128.0.0/14
descr:         PIPEX-BLOCK1
origin:        AS1849
notify:        routing@uk.uu.net
mnt-by:        AS1849-MNT
changed:       beny@uk.uu.net 20020321
source:        RIPE
```

**ASSIGNED PA – means that it is
Provider Aggregatable address space
and can only be used for connecting
to the ISP who assigned it**



Receiving Prefixes: From Peers

- **A peer is an ISP with whom you agree to exchange prefixes you originate into the Internet routing table**

Prefixes you accept from a peer are only those they have indicated they will announce

Prefixes you announce to your peer are only those you have indicated you will announce

Receiving Prefixes: From Peers

- **Agreeing what each will announce to the other:**

Exchange of e-mail documentation as part of the peering agreement, and then ongoing updates

OR

Use of the Internet Routing Registry and configuration tools such as the IRRToolSet

<http://www.isc.org/sw/IRRToolSet/>

Receiving Prefixes: From Upstream/Transit Provider

- Upstream/Transit Provider is an ISP who you pay to give you transit to the **WHOLE** Internet
- Receiving prefixes from them is not desirable unless really necessary
 - special circumstances – see Multihoming Tutorial
- Ask upstream/transit provider to either:
 - originate a default-route
 - OR*
 - announce one prefix you can use as default

Receiving Prefixes: From Upstream/Transit Provider

- If necessary to receive prefixes from any provider, care is required

don't accept RFC1918 *etc* prefixes

<http://ftp.rfc-editor.org/in-notes/rfc3330.txt>

don't accept your own prefixes

don't accept default (unless you need it)

don't accept prefixes longer than /24

- Check Rob Thomas' list of “bogons”

<http://www.cymru.com/Documents/bogon-list.html>

Receiving Prefixes

- **Paying attention to prefixes received from customers, peers and transit providers assists with:**

The integrity of the local network

The integrity of the Internet

- **Responsibility of all ISPs to be good Internet citizens**

Preparing the Network

Preparing the Network

- **We want to deploy BGP now...**
- **BGP will be used therefore an ASN is required**
- **If multihoming to different ISPs is intended in the near future, a public ASN should be obtained:**

Either go to upstream ISP who is a registry member, or

Apply to the RIR yourself for a one off assignment, or

Ask an ISP who is a registry member, or

**Join the RIR and get your own IP address allocation too
(this option strongly recommended)!**

Preparing the Network

- **The network is not running any BGP at the moment**
single statically routed connection to upstream ISP
- **The network is not running any IGP at all**
Static default and routes through the network to do “routing”

Preparing the Network IGP

- **Decide on IGP: OSPF or ISIS 😊**
- **Assign loopback interfaces and /32 addresses to each router which will run the IGP**

Loopback is used for OSPF and BGP router id anchor

Used for iBGP and route origination

- **Deploy IGP (e.g. OSPF)**

IGP can be deployed with **NO IMPACT** on the existing static routing

e.g. OSPF distance might be 110, static distance is 1

Smallest distance wins

Preparing the Network

IGP (cont)

- **Be prudent deploying IGP – keep the Link State Database Lean!**

Router loopbacks go in IGP

Backbone WAN point to point links go in IGP

(In fact, any link where IGP dynamic routing will be run should go into IGP)

Summarise on area/level boundaries (if possible) – i.e. think about your IGP address plan

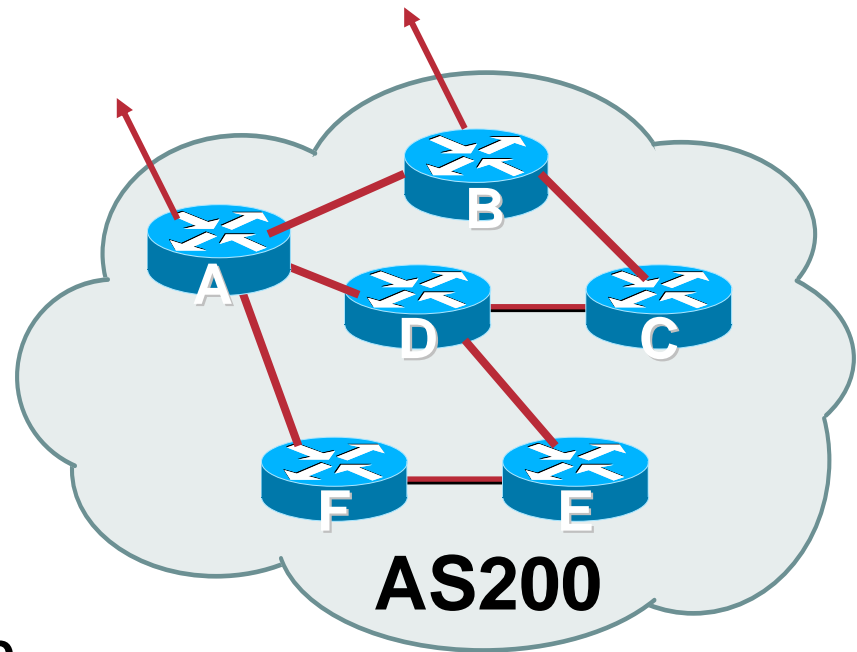
Preparing the Network

IGP (cont)

- **Routes which don't go into the IGP include:**
 - Dynamic assignment pools (DSL/Cable/Dial)**
 - Customer point to point link addressing**
 - (using next-hop-self in iBGP ensures that these do NOT need to be in IGP)**
 - Static/Hosting LANs**
 - Customer assigned address space**
 - Anything else not listed in the previous slide**

Preparing the Network iBGP

- Second step is to configure the local network to use iBGP
- iBGP can run on
 - all routers, or
 - a subset of routers, or
 - just on the upstream edge
- *iBGP must run on all routers which are in the transit path between external connections*



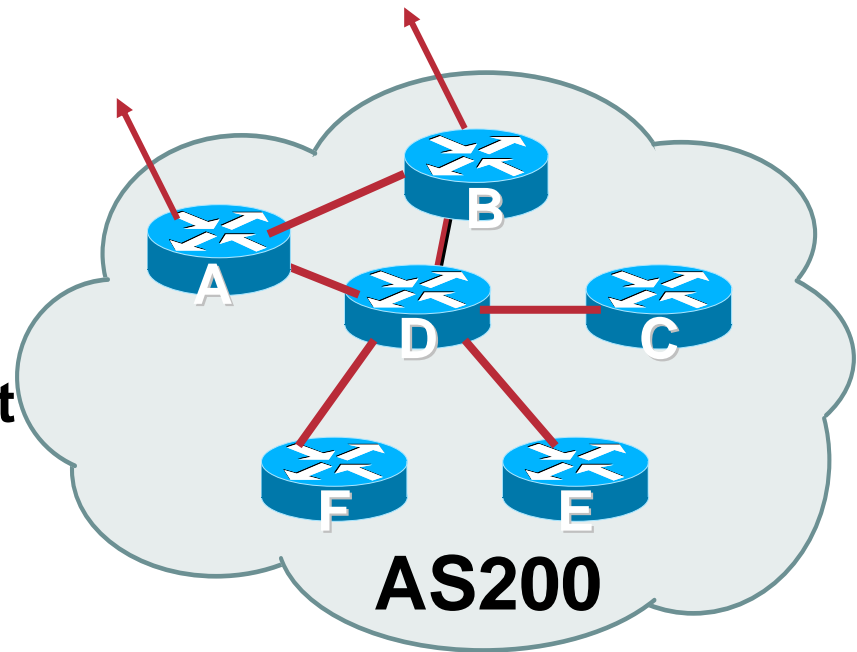
Preparing the Network iBGP (Transit Path)

- *iBGP must run on all routers which are in the transit path between external connections*
- Routers C, E and F are not in the transit path

Static routes or IGP will suffice

- Router D is in the transit path

Will need to be in iBGP mesh, otherwise routing loops will result



Preparing the Network Layers

- **Typical SP networks have three layers:**
 - Core – the backbone, usually the transit path**
 - Distribution – the middle, PoP aggregation layer**
 - Aggregation – the edge, the devices connecting customers**

Preparing the Network Aggregation Layer

- **iBGP is optional**

Many ISPs run iBGP here, either partial routing (more common) or full routing (less common)

Full routing is not needed unless customers want full table

Partial routing is cheaper/easier, might usually consist of internal prefixes and, optionally, external prefixes to aid external load balancing

Communities make this administratively easy

- **Many aggregation devices can't run iBGP**

Static routes from distribution devices for address pools

IGP for best exit

Preparing the Network Distribution Layer

- **Usually runs iBGP**
Partial or full routing (as with aggregation layer)
- **But does not have to run iBGP**
IGP is then used to carry customer prefixes (does not scale)
IGP is used to determine nearest exit
- **Networks which plan to grow large should deploy iBGP from day one**
Migration at a later date is extra work
No extra overhead in deploying iBGP; indeed, the IGP benefits

Preparing the Network Core Layer

- **Core of network is usually the transit path**
- **iBGP necessary between core devices**

Full routes or partial routes:

Transit ISPs carry full routes in core

Edge ISPs carry partial routes only

- **Core layer includes AS border routers**

Preparing the Network

iBGP Implementation

Decide on:

- **Best iBGP policy**

Will it be full routes everywhere, or partial, or some mix?

- **iBGP scaling technique**

Community policy?

Route-reflectors?

Techniques such as peer templates?

Preparing the Network

iBGP Implementation

- **Then deploy iBGP:**

Step 1: Introduce iBGP mesh on chosen routers

make sure that iBGP distance is greater than IGP distance (it usually is)

Step 2: Install “customer” prefixes into iBGP

Check! Does the network still work?

Step 3: Carefully remove the static routing for the prefixes now in IGP and iBGP

Check! Does the network still work?

Step 4: Deployment of eBGP follows

Preparing the Network iBGP Implementation

Install “customer” prefixes into iBGP?

- **Customer assigned address space**
Network statement/static route combination
Use unique community to identify customer assignments
- **Customer facing point-to-point links**
Redistribute connected routes through filters which only permit point-to-point link addresses to enter iBGP
Use a unique community to identify point-to-point link addresses (these are only required for your monitoring system)
- **Dynamic assignment pools & local LANs**
Simple network statement will do this
Use unique community to identify these networks

Preparing the Network iBGP Implementation

Carefully remove static routes?

- **Work on one router at a time:**

Check that static route for a particular destination is also learned either by IGP or by iBGP

If so, remove it

If not, establish why and fix the problem

(Remember to look in the RIB, not the FIB!)

- **Then the next router, until the whole PoP is done**
- **Then the next PoP, and so on until the network is now dependent on the IGP and iBGP you have deployed**

Preparing the Network Completion

- **Previous steps are NOT flag day steps**

Each can be carried out during different maintenance periods, for example:

Step One on Week One

Step Two on Week Two

Step Three on Week Three

And so on

And with proper planning will have NO customer visible impact at all

Preparing the Network Configuration Summary

- **IGP essential networks are in IGP**
- **Customer networks are now in iBGP**
iBGP deployed over the backbone
Full or Partial or Upstream Edge only
- **BGP distance is greater than any IGP**
- **Now ready to deploy eBGP**

Configuration Tips

Of templates, passwords, tricks, and more templates

iBGP and IGP

Reminder!

- **Make sure loopback is configured on router**
iBGP between loopbacks, **NOT** real interfaces
- **Make sure IGP carries loopback /32 address**
- **Consider the DMZ nets:**
 - Use unnumbered interfaces?
 - Use next-hop-self on iBGP neighbours
 - Or carry the DMZ /30s in the iBGP
 - Basically keep the DMZ nets out of the IGP!

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 unnumbered interfaces

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

- **eBGP and iBGP examples follow**

Also see Project Cymru's BGP templates

www.cymru.com/Documents

iBGP Template

Example

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

eBGP Template

Example

- **BGP damping**
 - Use RIPE-229 parameters, or something even weaker**
 - Don't use the vendor defaults without thinking**
- **Remove private ASes from announcements**
 - Common omission today**
- **Use extensive filters, with “backup”**
 - Use as-path filters to backup prefix filters**
 - Keep policy language for implementing policy, rather than basic filtering**
- **Use password agreed between you and peer on eBGP session**

eBGP Template

Example continued

- **Use maximum-prefix tracking**

Router will warn you if there are sudden increases in BGP table size, bringing down eBGP if desired

- **Log changes of neighbour state**

...and monitor those logs!

- **Make BGP admin distance higher than that of any IGP**

Otherwise prefixes heard from outside your network could override your IGP!!

Limiting AS Path Length

- **Some BGP implementations have problems with long AS_PATHS**

Memory corruption

Memory fragmentation

- **Even using AS_PATH prepends, it is not normal to see more than 20 ASes in a typical AS_PATH in the Internet today**

The Internet is around 5 ASes deep on average

Largest AS_PATH is usually 16-20 ASNs

Limiting AS Path Length

- **Some announcements have ridiculous lengths of AS-paths:**

```
*> 3FFE:1600::/24    3FFE:C00:8023:5::2    22 11537 145 12199
10318 10566 13193 1930 2200 3425 293 5609 5430 13285 6939
14277 1849 33 15589 25336 6830 8002 2042 7610 i
```

This example is an error in one IPv6 implementation

- **If your implementation supports it, consider limiting the maximum AS-path length you will accept**

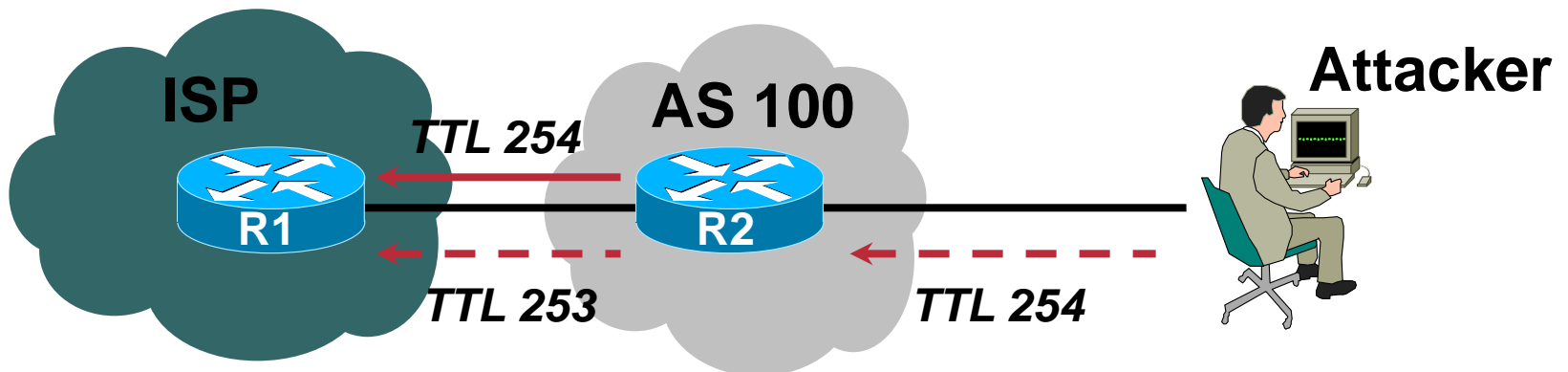
BGP TTL “hack”

- Implement RFC3682 on BGP peerings

Neighbour sets TTL to 255

Local router expects TTL of incoming BGP packets to be 254

No one apart from directly attached devices can send BGP packets which arrive with TTL of 254, so any possible attack by a remote miscreant is dropped due to TTL mismatch



BGP TTL “hack”

- **TTL Hack:**

Both neighbours must agree to use the feature
TTL check is much easier to perform than MD5
(Called BTSH – **BGP TTL Security Hack**)

- **Provides “security” for BGP sessions**

In addition to packet filters of course

MD5 should still be used for messages which slip through the TTL hack

See www.nanog.org/mtg-0302/hack.html for more details

Passwords on BGP sessions

- *Yes, I am mentioning passwords again*

- **Put password on the BGP session**

It's a secret shared between you and your peer

If arriving packets don't have the correct MD5 hash, they are ignored

Helps defeat miscreants who wish to attack BGP sessions

- **Powerful preventative tool, especially when combined with filters and the TTL "hack"**

Summary

- **Use configuration templates**
- **Standardise the configuration**
- **Be aware of standard “tricks” to avoid compromise of the BGP session**
- **Anything to make your life easier, network less prone to errors, network more likely to scale**
- **It's all about scaling – if your network won't scale, then it won't be successful**

BGP for Internet Service Providers

- **Scaling BGP**
- **Using Communities**
- **Deploying BGP in an ISP network**



Deploying BGP

Philip Smith <pfs@cisco.com>

APRICOT 2005

Kyoto, Japan