



BGP Techniques for Internet Service Providers

Philip Smith <pfs@cisco.com>

APRICOT 2009

18th-27th February 2009

Manila, Philippines

Presentation Slides

- Will be available on
[ftp://ftp-eng.cisco.com
/pfs/seminars/APRICOT2009-BGP-Techniques.pdf](ftp://ftp-eng.cisco.com/pfs/seminars/APRICOT2009-BGP-Techniques.pdf)
And on the APRICOT 2009 website
- Feel free to ask questions any time

BGP Techniques for Internet Service Providers

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



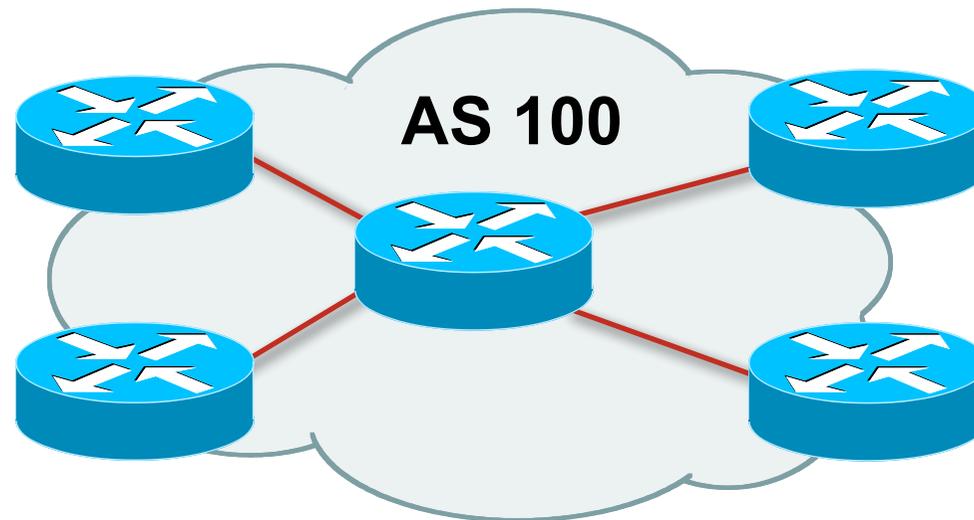
BGP Basics

What is BGP?

Border Gateway Protocol

- A Routing Protocol used to exchange routing information between different networks
 - Exterior gateway protocol
- Described in RFC4271
 - RFC4276 gives an implementation report on BGP
 - RFC4277 describes operational experiences using BGP
- The Autonomous System is the cornerstone of BGP
 - It is used to uniquely identify networks with a common routing policy

Autonomous System (AS)



- Collection of networks with same routing policy
- Single routing protocol
- Usually under single ownership, trust and administrative control
- Identified by a unique 32-bit integer (ASN)

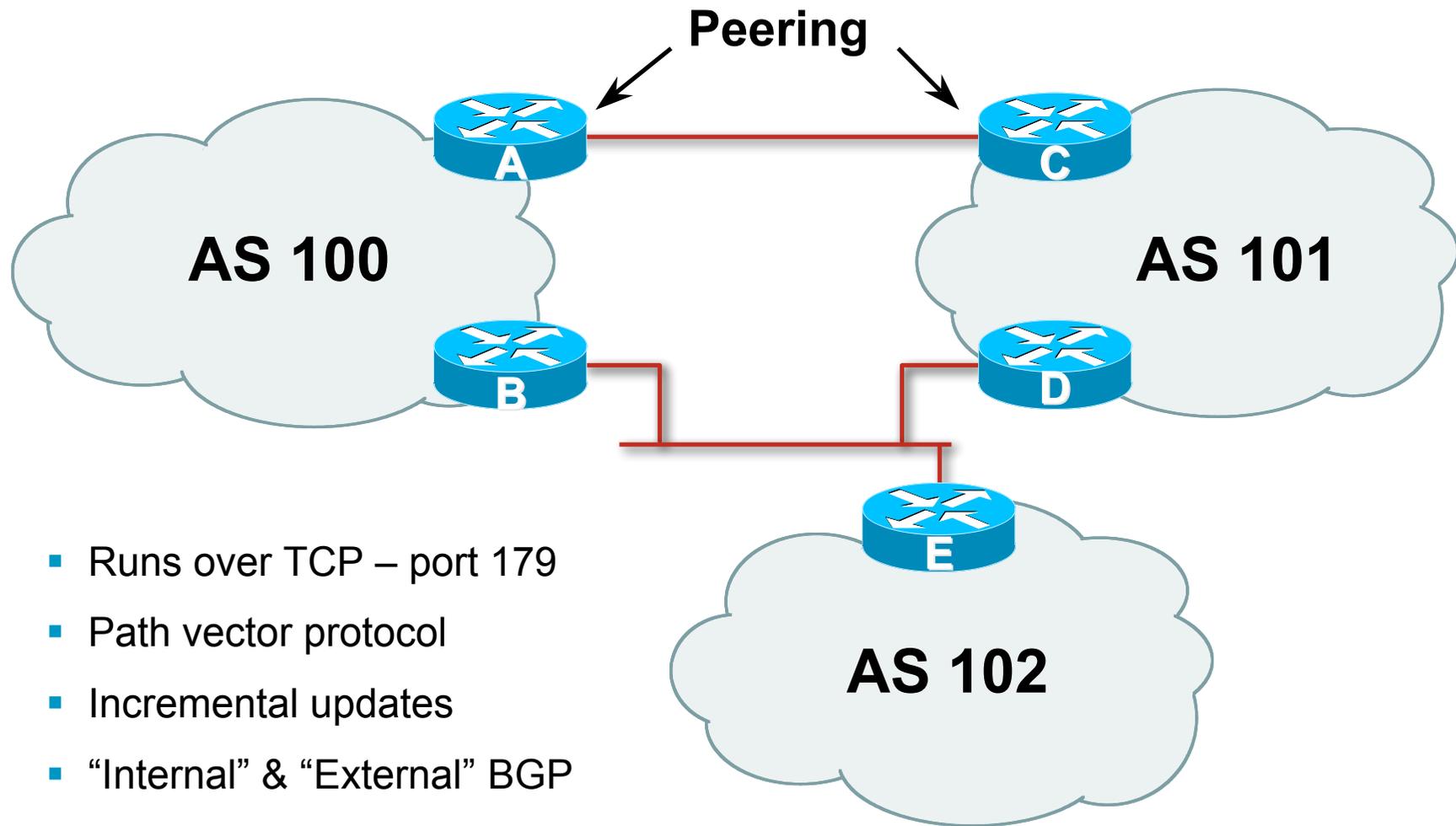
Autonomous System Number (ASN)

- Two ranges
 - 0-65535 (original 16-bit range)
 - 65536-4294967295 (32-bit range - RFC4893)
- Usage:
 - 0 and 65535 (reserved)
 - 1-64495 (public Internet)
 - 64496-64511 (documentation - RFC5398)
 - 64512-65534 (private use only)
 - 23456 (represent 32-bit range in 16-bit world)
 - 65536-65551 (documentation - RFC5398)
 - 65552-4294967295 (public Internet)
- 32-bit range representation specified in RFC5396
 - Defines “asplain” (traditional format) as standard notation

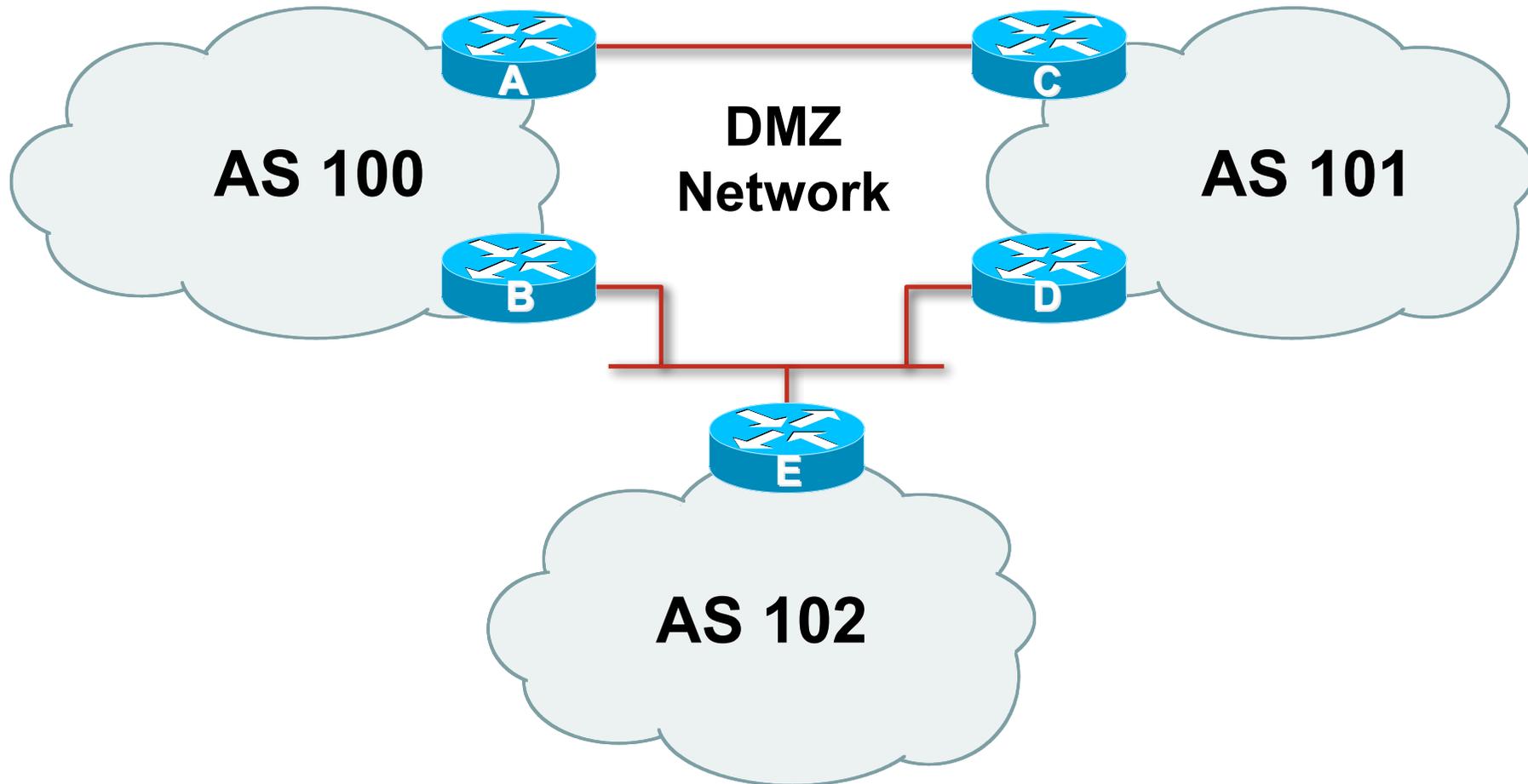
Autonomous System Number (ASN)

- ASNs are distributed by the Regional Internet Registries
 - They are also available from upstream ISPs who are members of one of the RIRs
- Current 16-bit ASN allocations up to 49151 have been made to the RIRs
 - Around 30600 are visible on the Internet
- The RIRs also have received 1024 32-bit ASNs each
 - 18 are visible on the Internet (early adopters)
- See www.iana.org/assignments/as-numbers

BGP Basics



Demarcation Zone (DMZ)



- Shared network between ASes

BGP General Operation

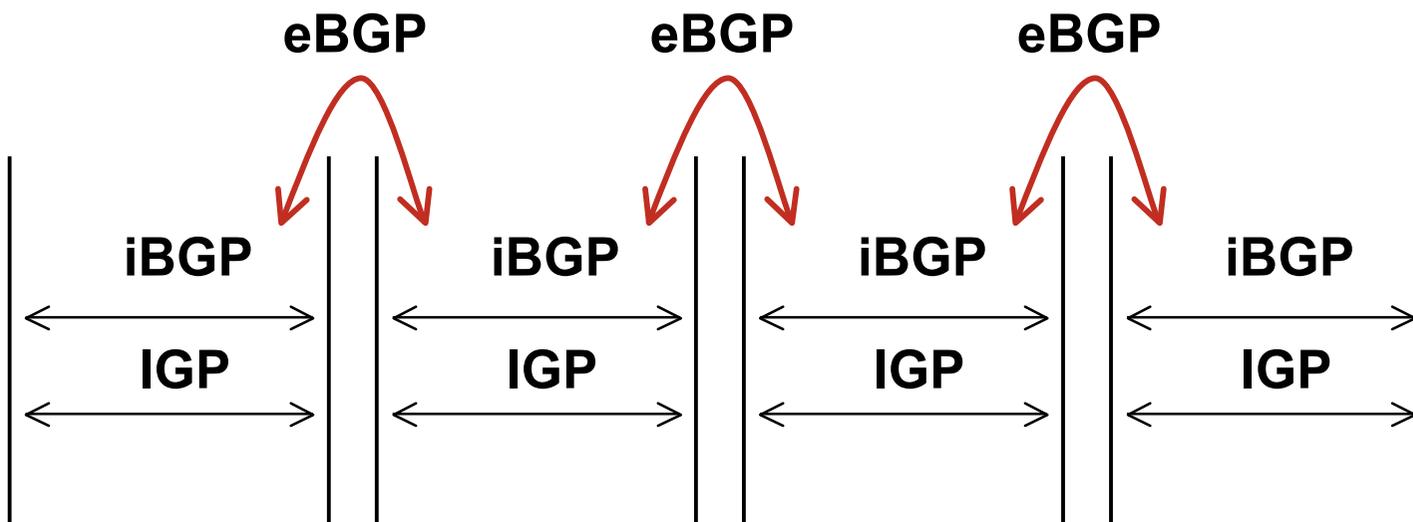
- Learns multiple paths via internal and external BGP speakers
- Picks the best path and installs in the forwarding table
- Best path is sent to external BGP neighbours
- Policies are applied by influencing the best path selection

eBGP & iBGP

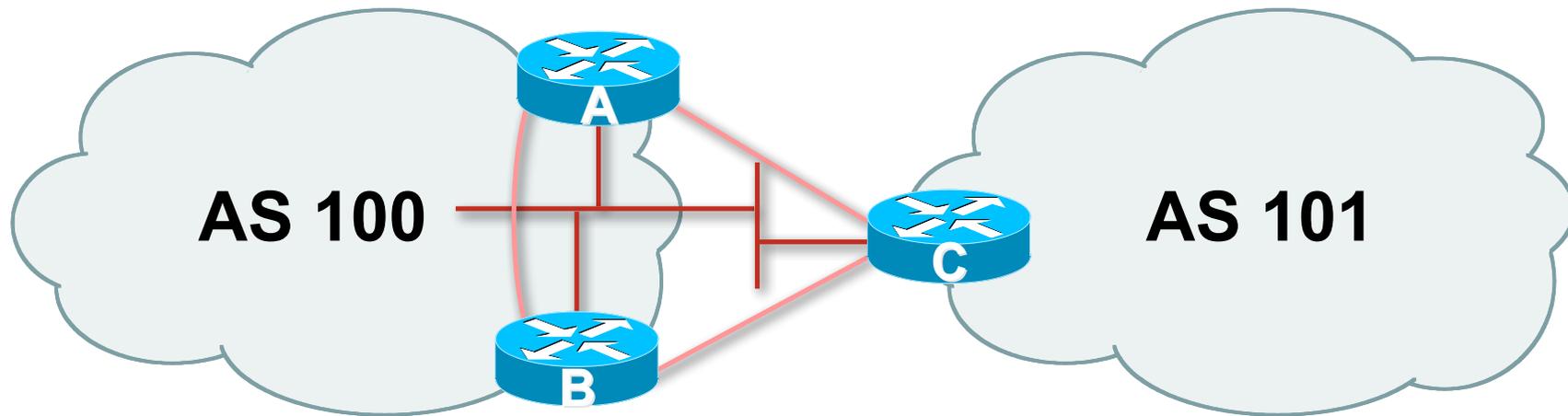
- BGP used internally (iBGP) and externally (eBGP)
- iBGP used to carry
 - some/all Internet prefixes across ISP backbone
 - ISP's customer prefixes
- eBGP used to
 - exchange prefixes with other ASes
 - implement routing policy

BGP/IGP model used in ISP networks

- Model representation



External BGP Peering (eBGP)

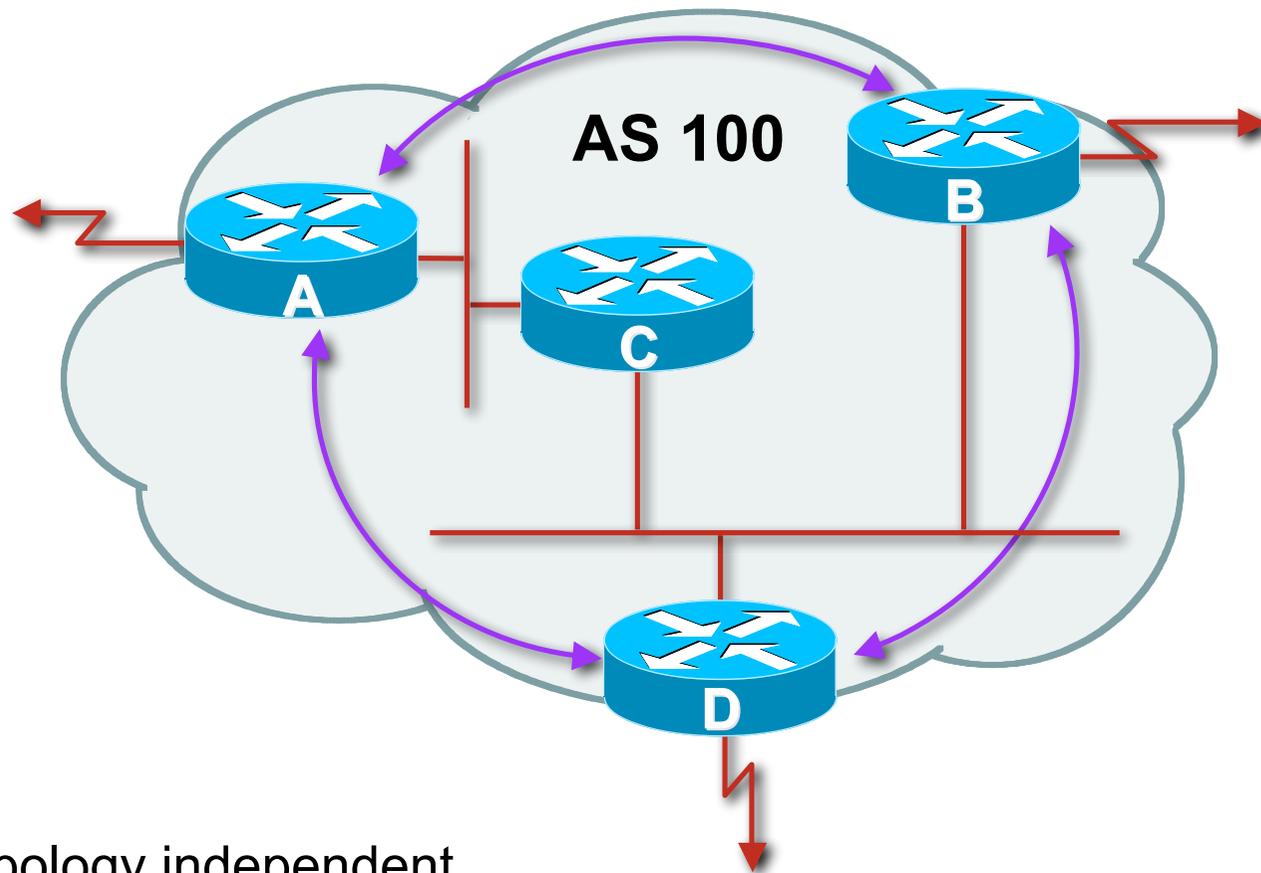


- Between BGP speakers in different AS
- Should be directly connected
- **Never** run an IGP between eBGP peers

Internal BGP (iBGP)

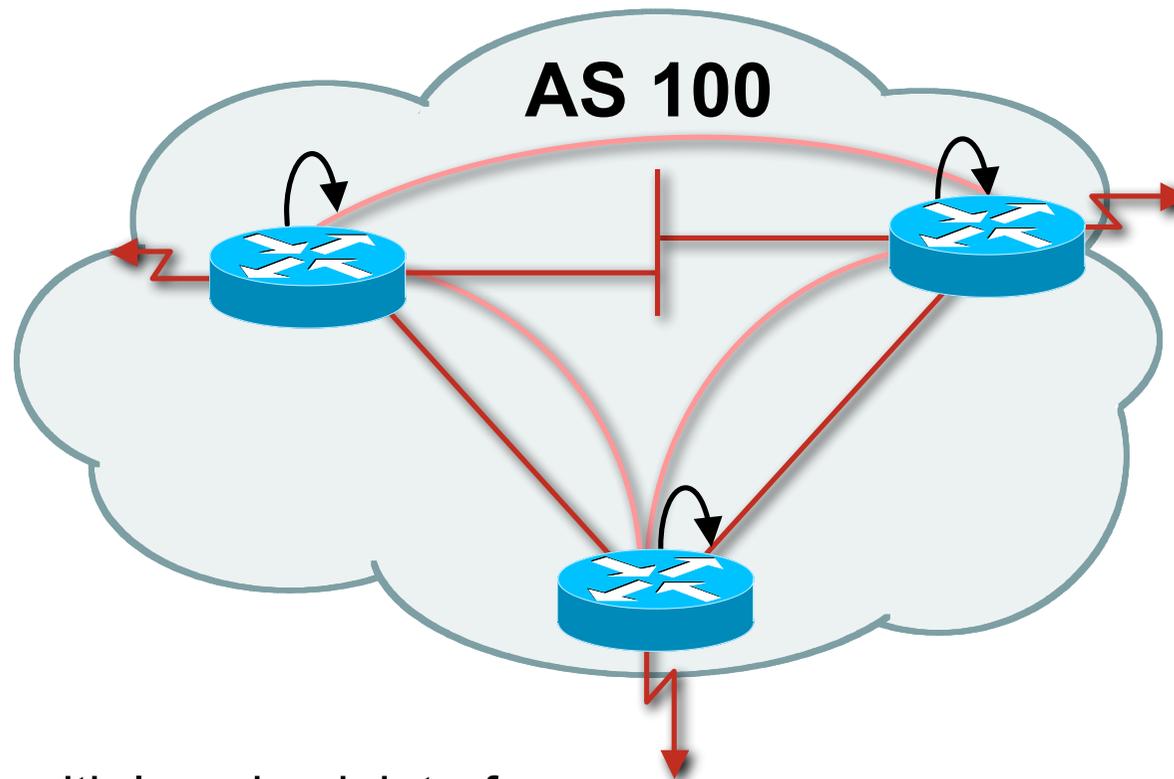
- BGP peer within the same AS
- Not required to be directly connected
 - IGP takes care of inter-BGP speaker connectivity
- iBGP speakers must to be fully meshed:
 - They originate connected networks
 - They pass on prefixes learned from outside the ASN
 - They do **not** pass on prefixes learned from other iBGP speakers

Internal BGP Peering (iBGP)



- Topology independent
- Each iBGP speaker must peer with every other iBGP speaker in the AS

Peering to Loopback Interfaces



- Peer with loop-back interface
 - Loop-back interface does not go down – ever!
- Do not want iBGP session to depend on state of a single interface or the physical topology

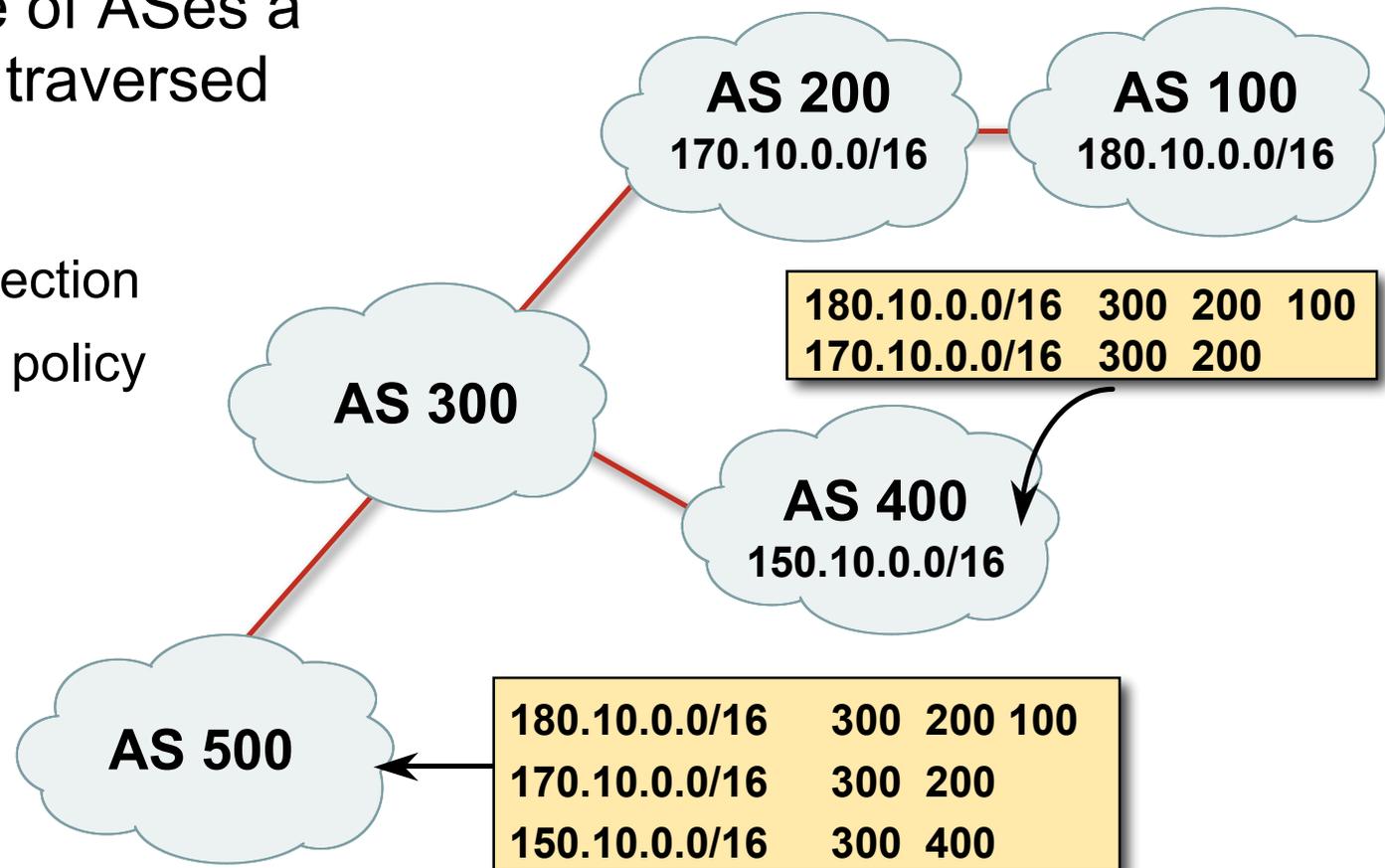


BGP Attributes

Information about BGP

AS-Path

- Sequence of ASes a route has traversed
- Used for:
 - Loop detection
 - Applying policy

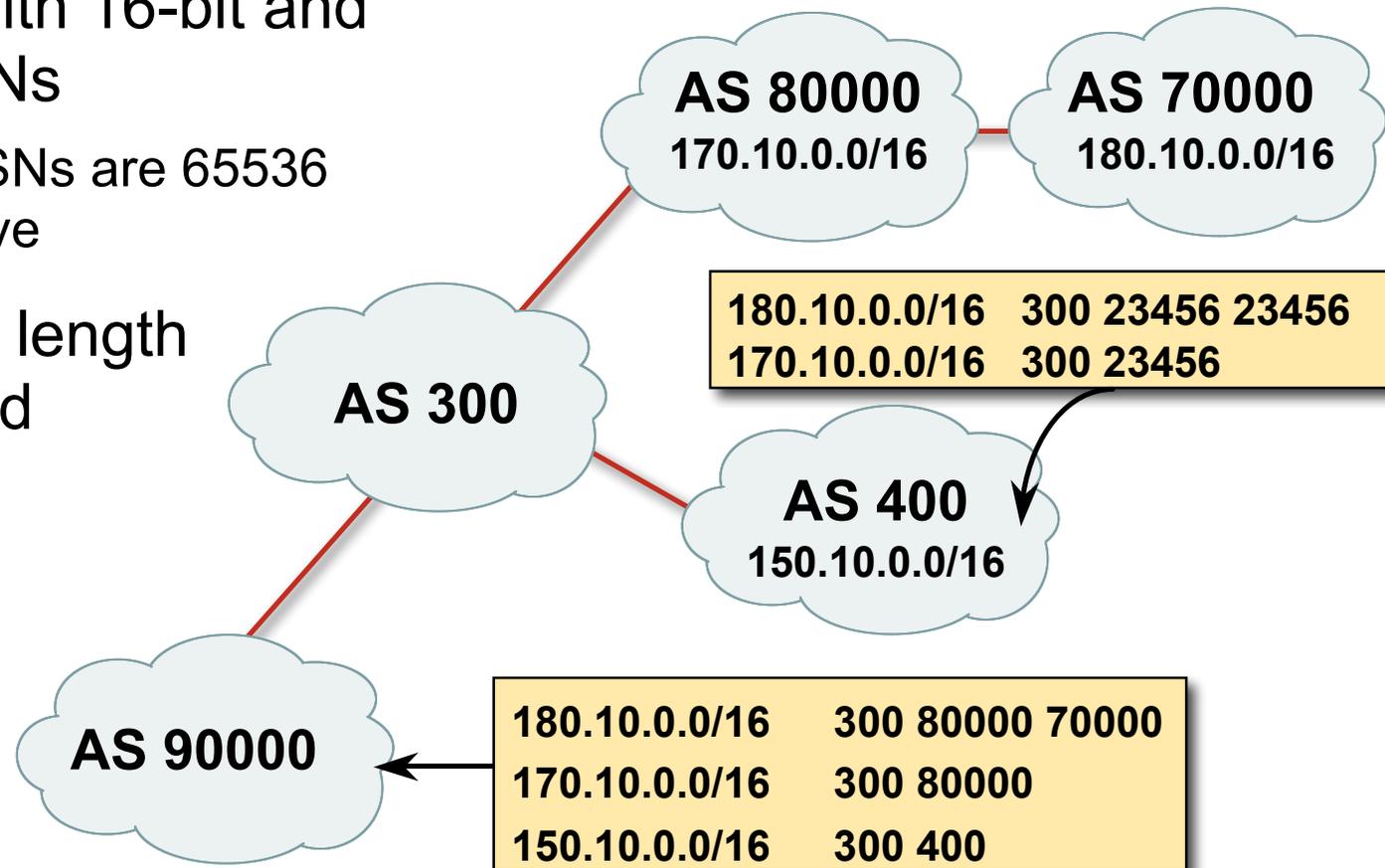


AS-Path (with 16 and 32-bit ASNs)

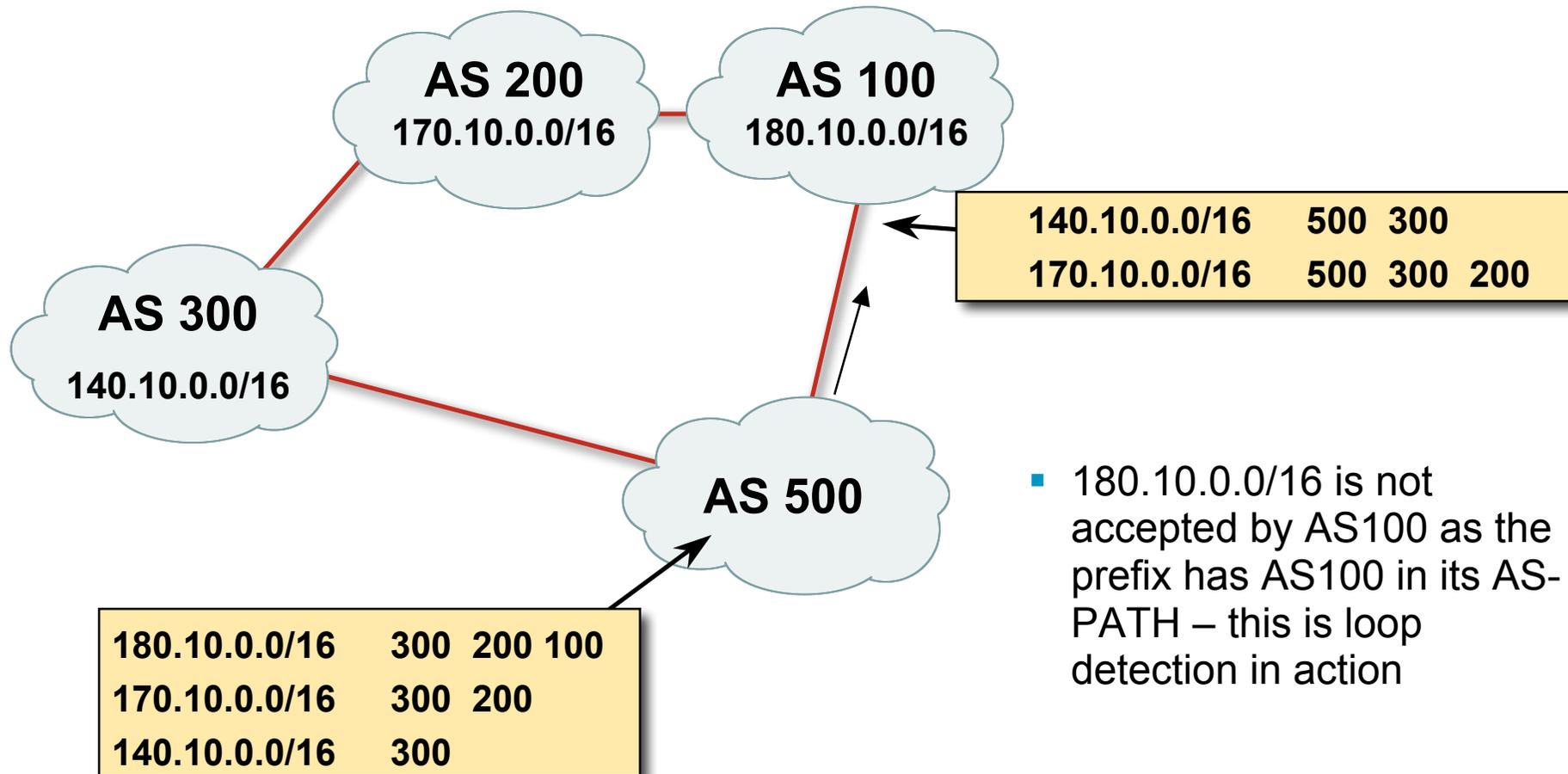
- Internet with 16-bit and 32-bit ASNs

32-bit ASNs are 65536 and above

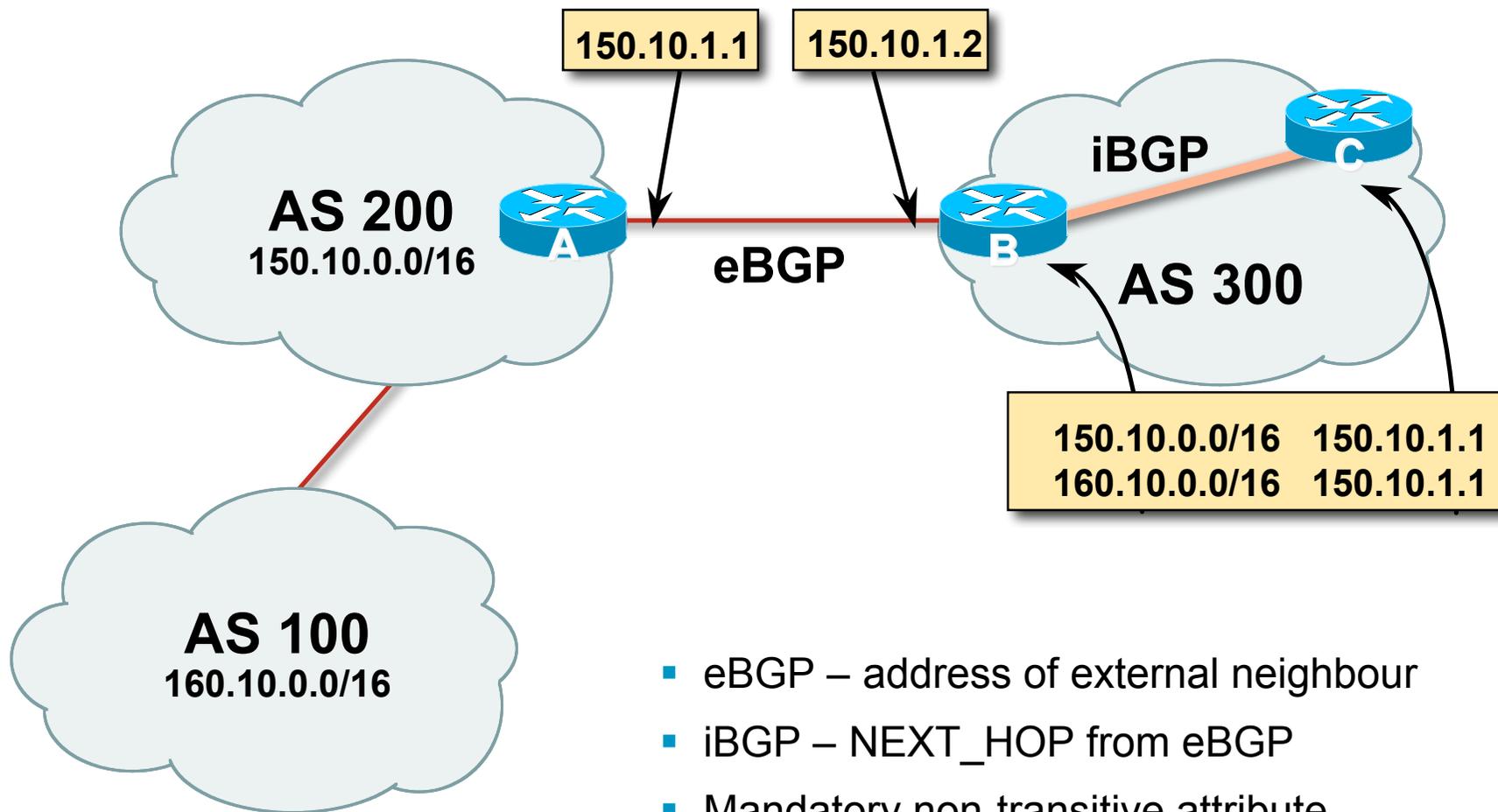
- AS-PATH length maintained



AS-Path loop detection

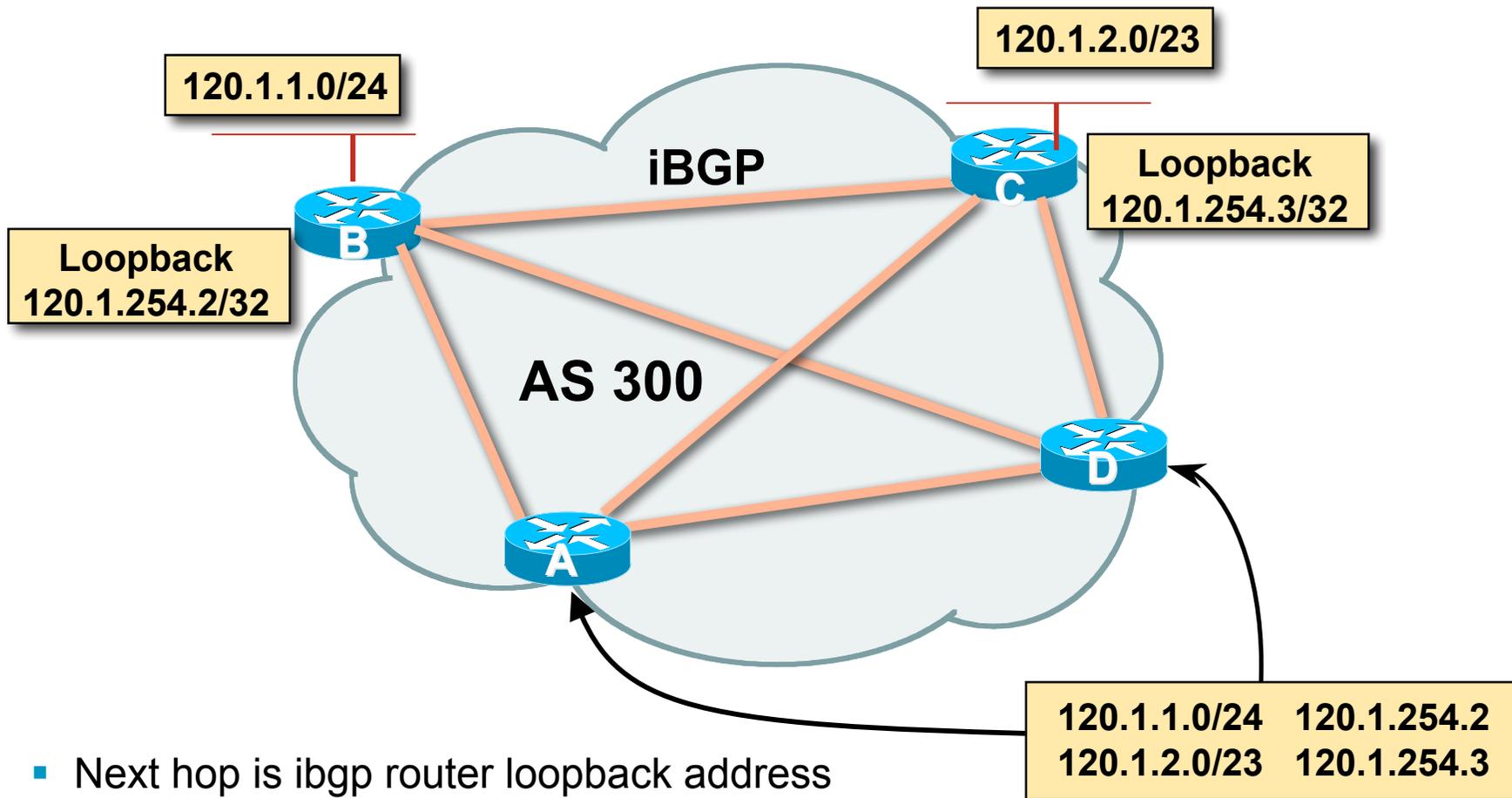


Next Hop



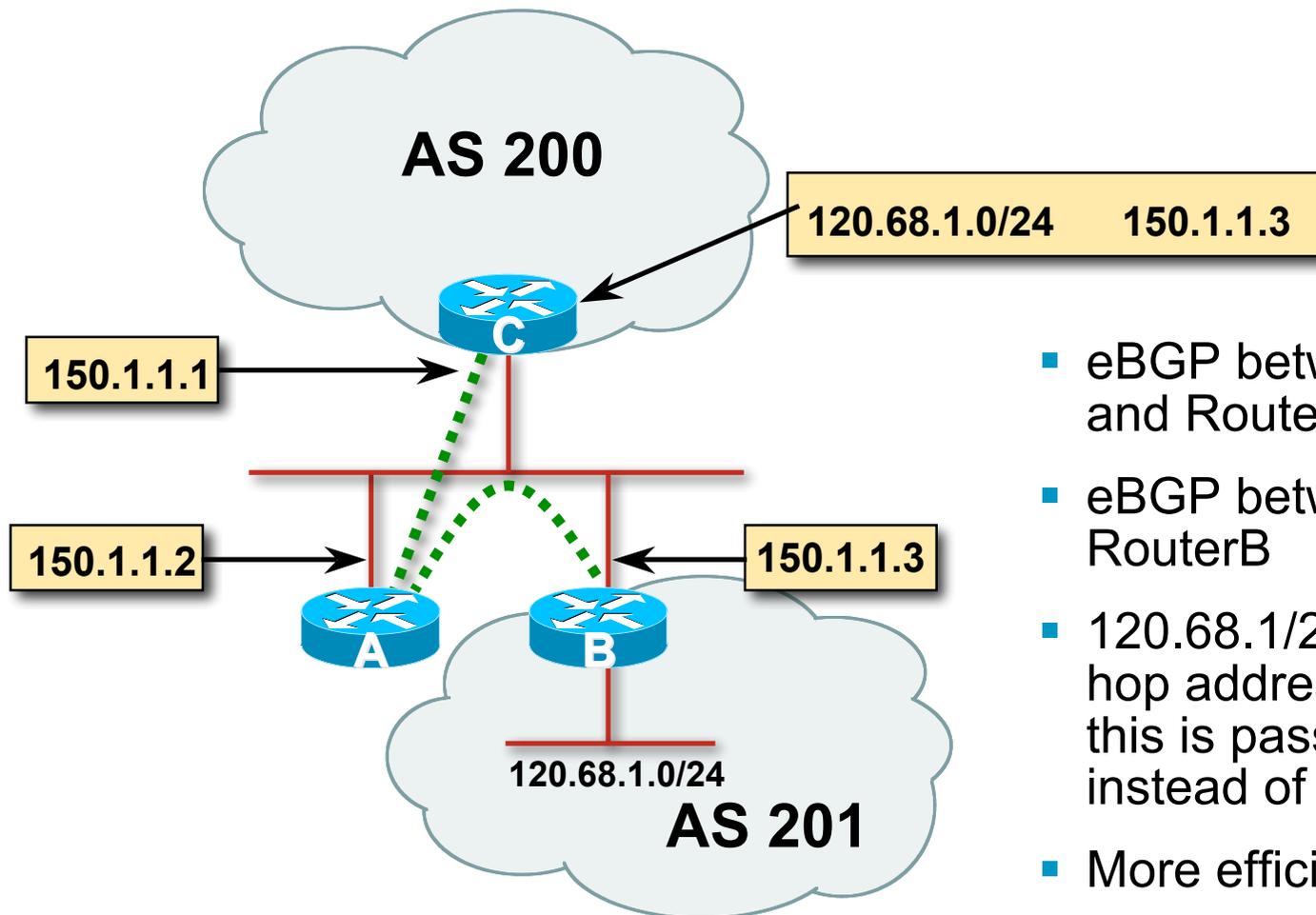
- eBGP – address of external neighbour
- iBGP – NEXT_HOP from eBGP
- Mandatory non-transitive attribute

iBGP Next Hop



- Next hop is ibgp router loopback address
- Recursive route look-up

Third Party Next Hop



- eBGP between Router A and Router C
- eBGP between Router A and Router B
- 120.68.1/24 prefix has next hop address of 150.1.1.3 – this is passed on to Router C instead of 150.1.1.2
- More efficient
- No extra config needed

Next Hop Best Practice

- BGP default is for external next-hop to be propagated unchanged to iBGP peers

This means that IGP has to carry external next-hops

Forgetting means external network is invisible

With many eBGP peers, it is unnecessary extra load on IGP

- ISP Best Practice is to change external next-hop to be that of the local router

Next Hop (Summary)

- IGP should carry route to next hops
- Recursive route look-up
- Unlinks BGP from actual physical topology
- Change external next hops to that of local router
- Allows IGP to make intelligent forwarding decision

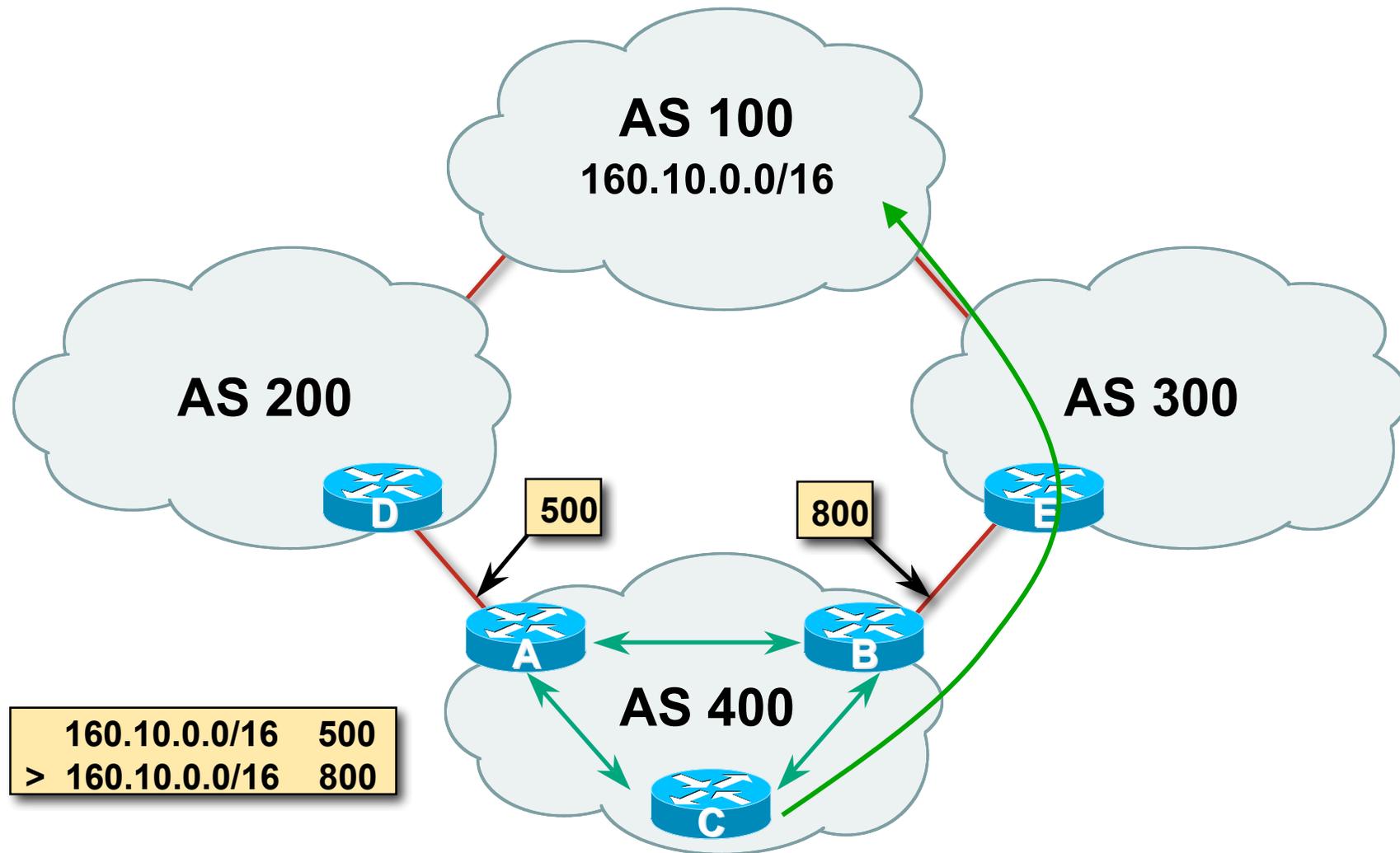
Origin

- Conveys the origin of the prefix
- **Historical** attribute
 - Used in transition from EGP to BGP
- Transitive and Mandatory Attribute
- Influences best path selection
- Three values: IGP, EGP, incomplete
 - IGP – generated by BGP network statement
 - EGP – generated by EGP
 - incomplete – redistributed from another routing protocol

Aggregator

- Conveys the IP address of the router or BGP speaker generating the aggregate route
- Optional & transitive attribute
- Useful for debugging purposes
- Does not influence best path selection

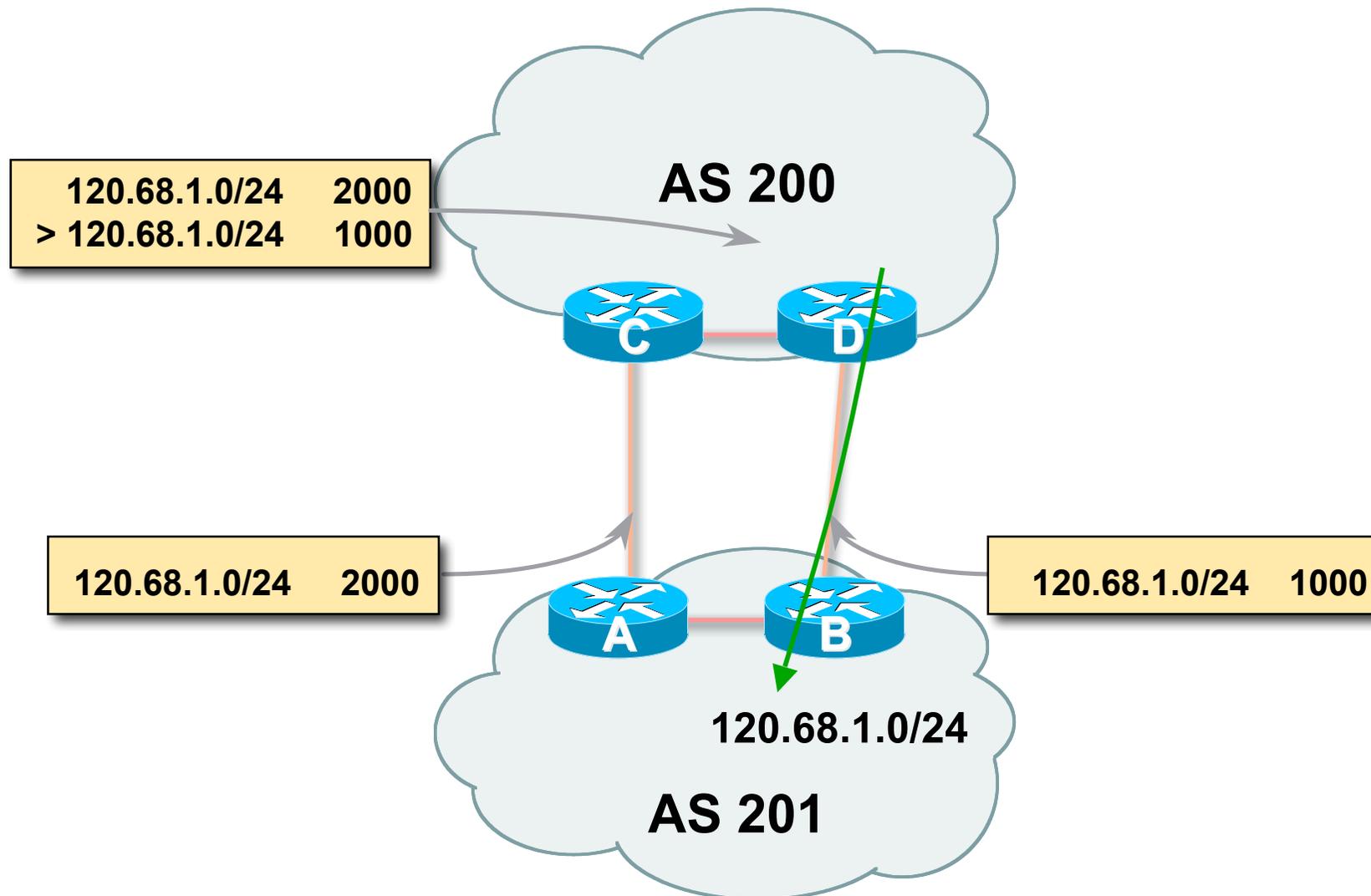
Local Preference



Local Preference

- Non-transitive and optional attribute
- Local to an AS – non-transitive
 - Default local preference is 100 (IOS)
- Used to influence BGP path selection
 - determines best path for *outbound* traffic
- Path with highest local preference wins

Multi-Exit Discriminator (MED)



Multi-Exit Discriminator

- Inter-AS – non-transitive & optional attribute
- Used to convey the relative preference of entry points
determines best path for inbound traffic
- Comparable if paths are from same AS
Implementations have a knob to allow comparisons of MEDs
from different ASes
- Path with lowest MED wins
- Absence of MED attribute implies MED value of **zero**
(RFC4271)

Multi-Exit Discriminator

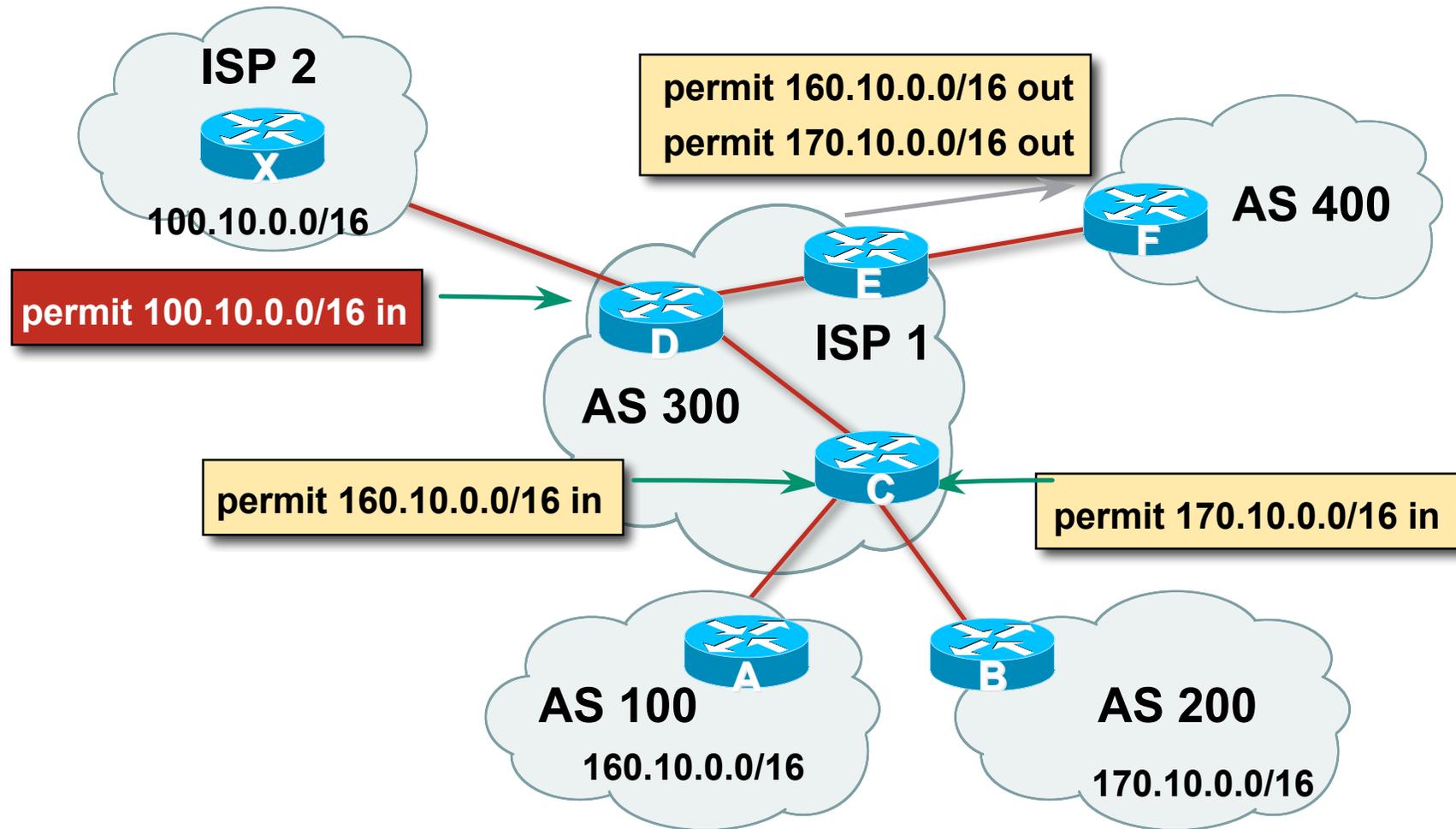
“metric confusion”

- MED is non-transitive and optional attribute
 - Some implementations send learned MEDs to iBGP peers by default, others do not
 - Some implementations send MEDs to eBGP peers by default, others do not
- Default metric varies according to vendor implementation
 - Original BGP spec (RFC1771) made no recommendation
 - Some implementations said that absence of metric was equivalent to 0
 - Other implementations said that absence of metric was equivalent to $2^{32}-1$ (highest possible) or $2^{32}-2$
 - Potential for “metric confusion”

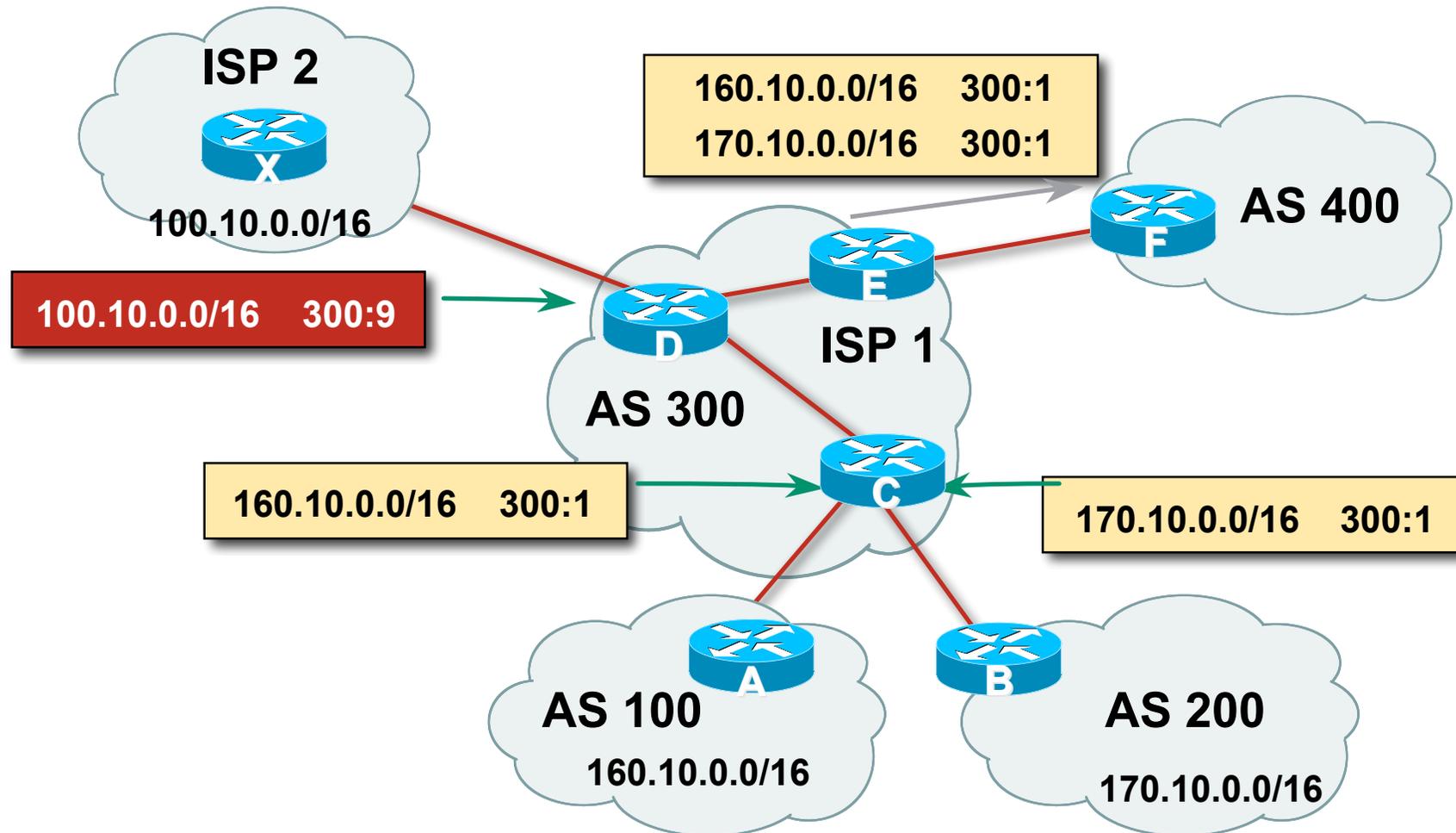
Community

- Communities are described in RFC1997
Transitive and Optional Attribute
- 32 bit integer
Represented as two 16 bit integers (RFC1998)
Common format is <local-ASN>:xx
0:0 to 0:65535 and 65535:0 to 65535:65535 are reserved
- Used to group destinations
Each destination could be member of multiple communities
- Very useful in applying policies within and between ASes

Community Example (before)



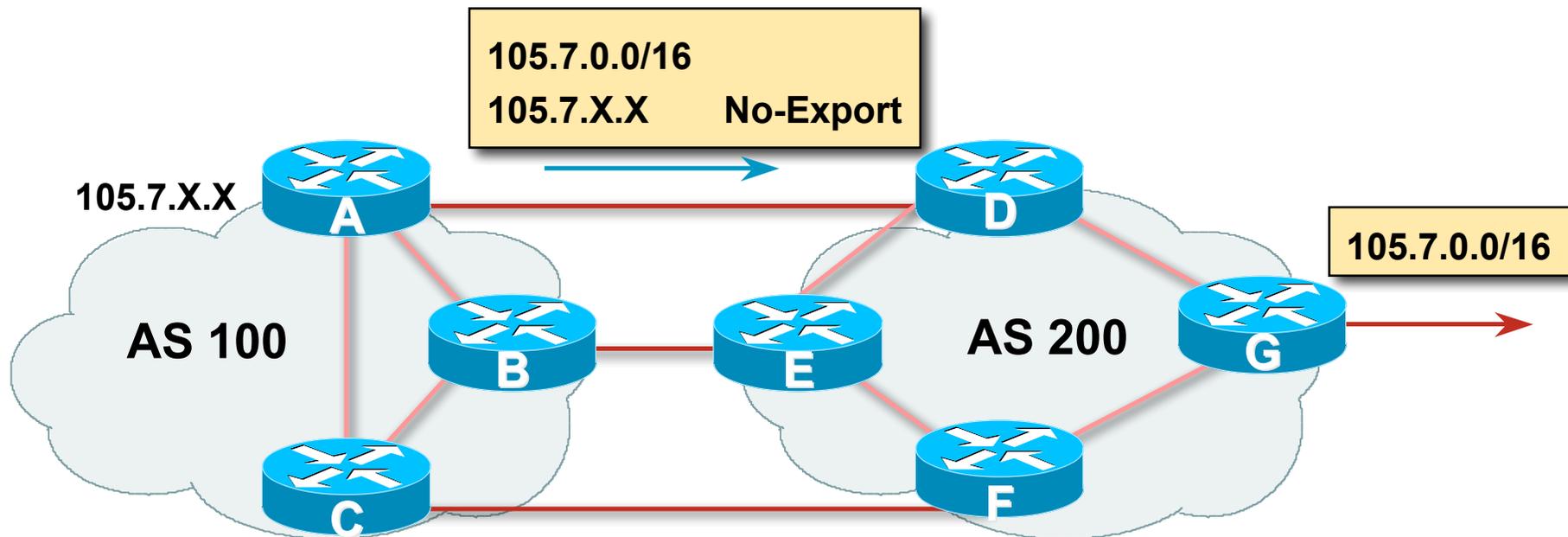
Community Example (after)



Well-Known Communities

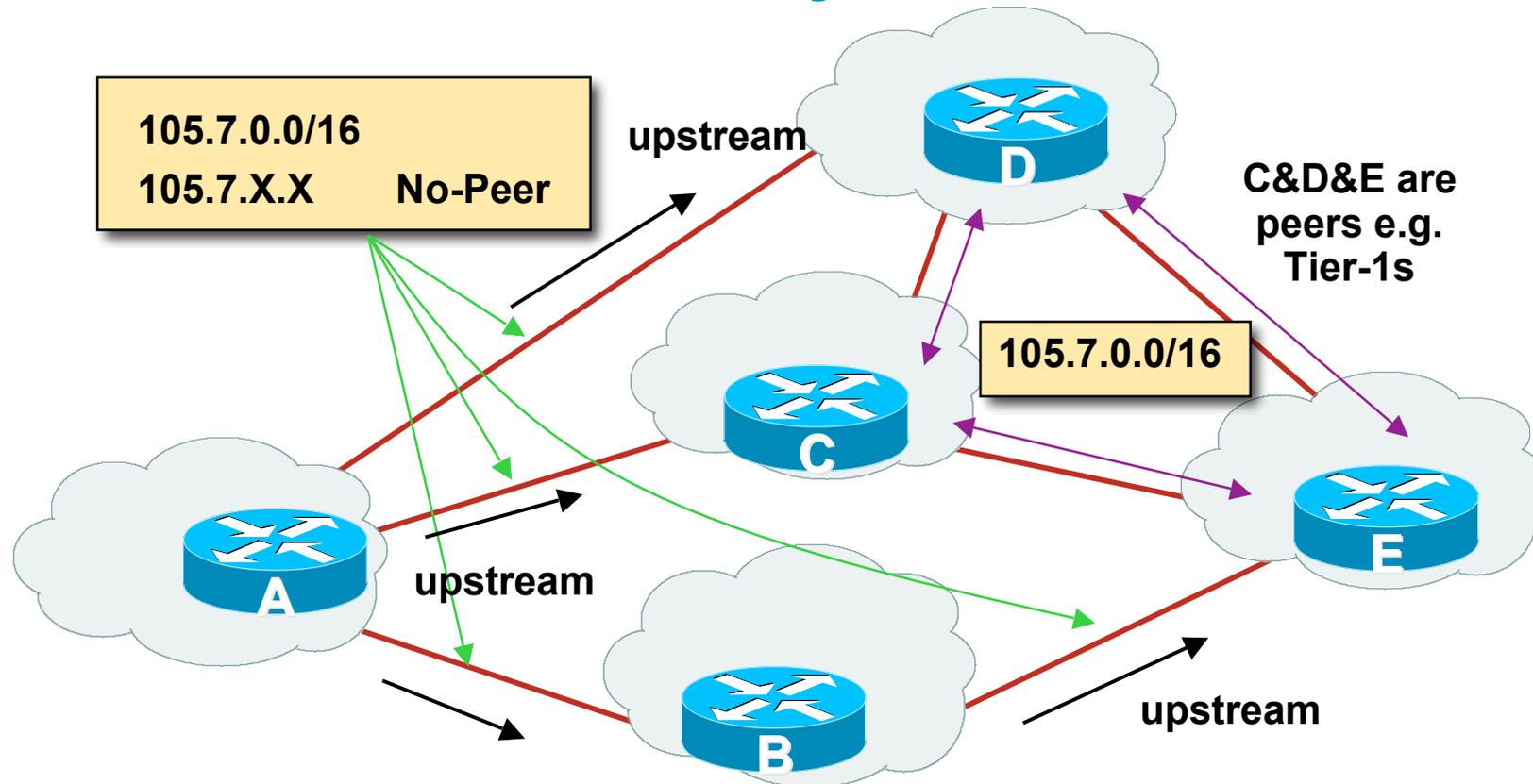
- Several well known communities
www.iana.org/assignments/bgp-well-known-communities
- no-export **65535:65281**
do not advertise to any eBGP peers
- no-advertise **65535:65282**
do not advertise to any BGP peer
- no-export-subconfed **65535:65283**
do not advertise outside local AS (only used with confederations)
- no-peer **65535:65284**
do not advertise to bi-lateral peers (RFC3765)

No-Export Community



- AS100 announces aggregate and subprefixes
Intention is to improve loadsharing by leaking subprefixes
- Subprefixes marked with **no-export** community
- Router G in AS200 does not announce prefixes with **no-export** community set

No-Peer Community



- Sub-prefixes marked with **no-peer** community are not sent to bi-lateral peers

They are only sent to upstream providers

Community

Implementation details

- Community is an optional attribute
 - Some implementations send communities to iBGP peers by default, some do not
 - Some implementations send communities to eBGP peers by default, some do not
- Being careless can lead to community “confusion”
 - ISPs need consistent community policy within their own networks
 - And they need to inform peers, upstreams and customers about their community expectations



BGP Path Selection Algorithm

Why Is This the Best Path?

BGP Path Selection Algorithm for IOS

Part One

- Do not consider path if no route to next hop
- Do not consider iBGP path if not synchronised (Cisco IOS only)
- Highest weight (local to router)
- Highest local preference (global within AS)
- Prefer locally originated route
- Shortest AS path

BGP Path Selection Algorithm for IOS

Part Two

- Lowest origin code
 - IGP < EGP < incomplete
- Lowest Multi-Exit Discriminator (MED)
 - If **bgp deterministic-med**, order the paths before comparing (BGP spec does not specify in which order the paths should be compared. This means best path depends on order in which the paths are compared.)
 - If **bgp always-compare-med**, then compare for all paths otherwise MED only considered if paths are from the same AS (default)

BGP Path Selection Algorithm for IOS

Part Three

- Prefer eBGP path over iBGP path
- Path with lowest IGP metric to next-hop
- Lowest router-id (originator-id for reflected routes)
- Shortest Cluster-List
 - Client **must** be aware of Route Reflector attributes!
- Lowest neighbour IP address

BGP Path Selection Algorithm

- In multi-vendor environments:

- Make sure the path selection processes are understood for each brand of equipment

- Each vendor has slightly different implementations, extra steps, extra features, etc

- Watch out for possible MED confusion



Applying Policy with BGP

Controlling Traffic Flow & Traffic Engineering

Applying Policy in BGP: Why?

- Network operators rarely “plug in routers and go”
- External relationships:
 - Control who they peer with
 - Control who they give transit to
 - Control who they get transit from
- Traffic flow control:
 - Efficiently use the scarce infrastructure resources (external link load balancing)
 - Congestion avoidance
 - Terminology: Traffic Engineering

Applying Policy in BGP: How?

- Policies are applied by:

- Setting BGP attributes (local-pref, MED, AS-PATH, community), thereby influencing the path selection process

- Advertising or Filtering prefixes

- Advertising or Filtering prefixes according to ASN and AS-PATHs

- Advertising or Filtering prefixes according to Community membership

Applying Policy with BGP: Tools

- Most implementations have tools to apply policies to BGP:
 - Prefix manipulation/filtering
 - AS-PATH manipulation/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



BGP Capabilities

Extending BGP

BGP Capabilities

- Documented in RFC2842
- Capabilities parameters passed in BGP open message
- Unknown or unsupported capabilities will result in NOTIFICATION message
- Codes:
 - 0 to 63 are assigned by IANA by IETF consensus
 - 64 to 127 are assigned by IANA “first come first served”
 - 128 to 255 are vendor specific

BGP Capabilities

Current capabilities are:

0	Reserved	[RFC3392]
1	Multiprotocol Extensions for BGP-4	[RFC4760]
2	Route Refresh Capability for BGP-4	[RFC2918]
3	Outbound Route Filtering Capability	[RFC5291]
4	Multiple routes to a destination capability	[RFC3107]
64	Graceful Restart Capability	[RFC4724]
65	Support for 4 octet ASNs	[RFC4893]
66	Deprecated 2003-03-06	
67	Support for Dynamic Capability	[ID]
68	Multisession BGP	[ID]

See www.iana.org/assignments/capability-codes

BGP Capabilities

- Multiprotocol extensions

 - This is a whole different world, allowing BGP to support more than IPv4 unicast routes

 - Examples include: v4 multicast, IPv6, v6 multicast, VPNs

 - Another tutorial (or many!)

- Route refresh is a well known scaling technique – covered shortly

- 32-bit ASNs have recently arrived

- The other capabilities are still in development or not widely implemented or deployed yet

BGP for Internet Service Providers

- BGP Basics
- **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?

BGP Scaling Techniques

- Route Refresh
- Route Reflectors
- Confederations



Dynamic Reconfiguration

Route Refresh

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:
 - Terminates 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 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



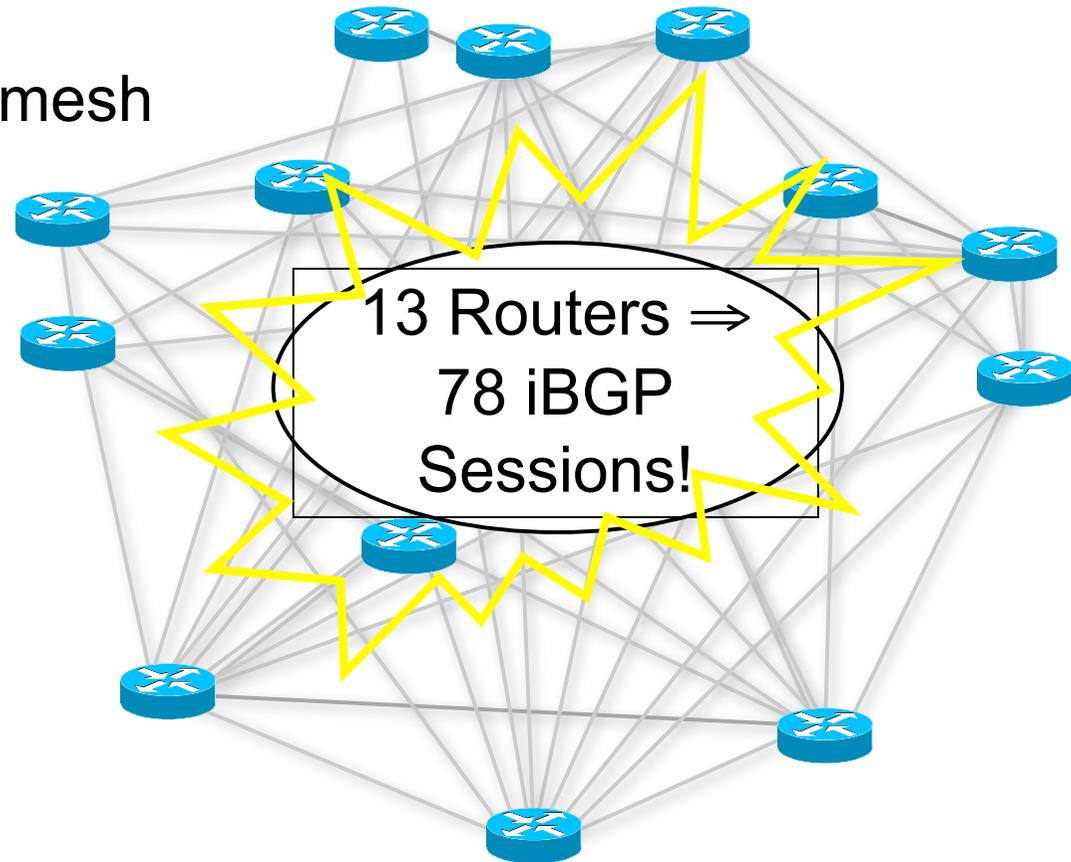
Route Reflectors

Scaling the iBGP mesh

Scaling iBGP mesh

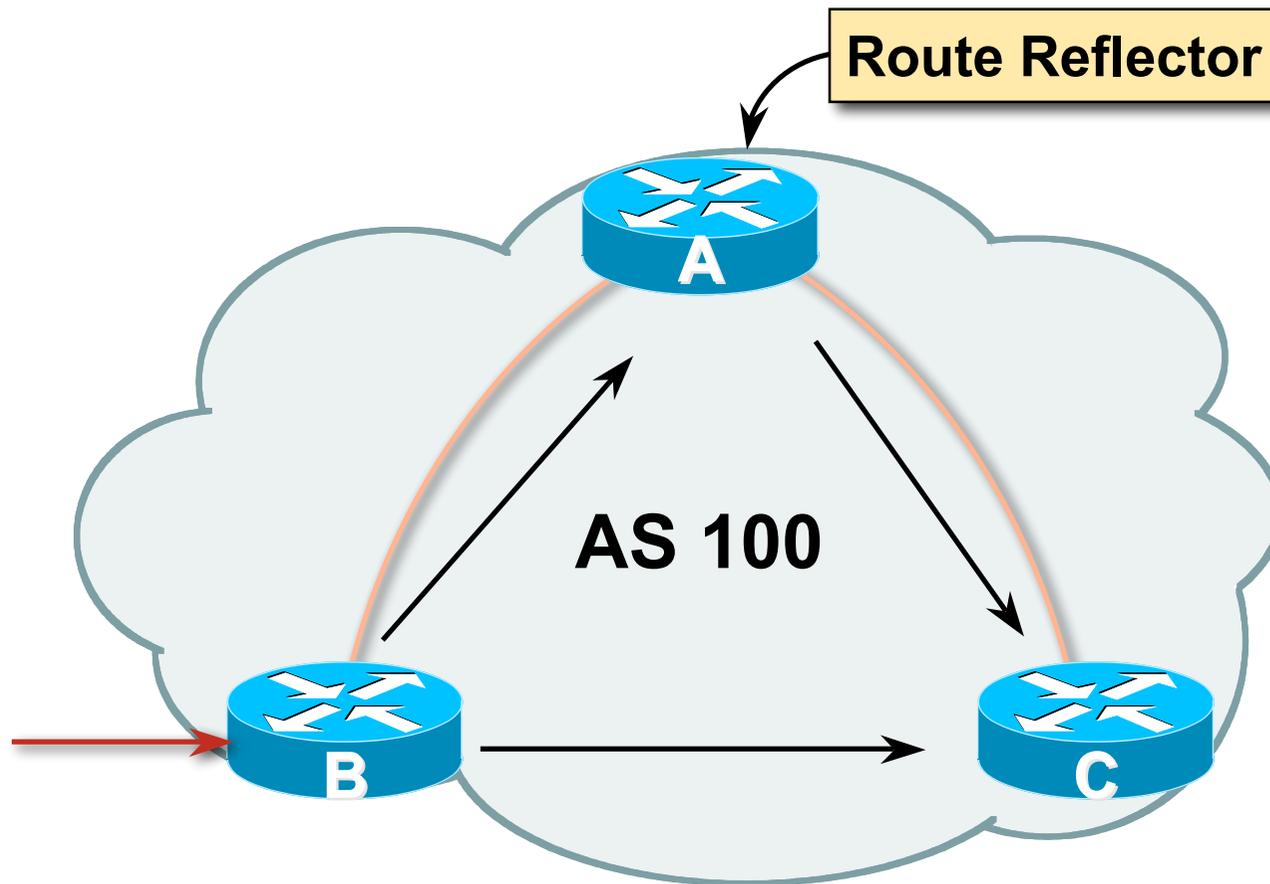
- Avoid $\frac{1}{2}n(n-1)$ iBGP mesh

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



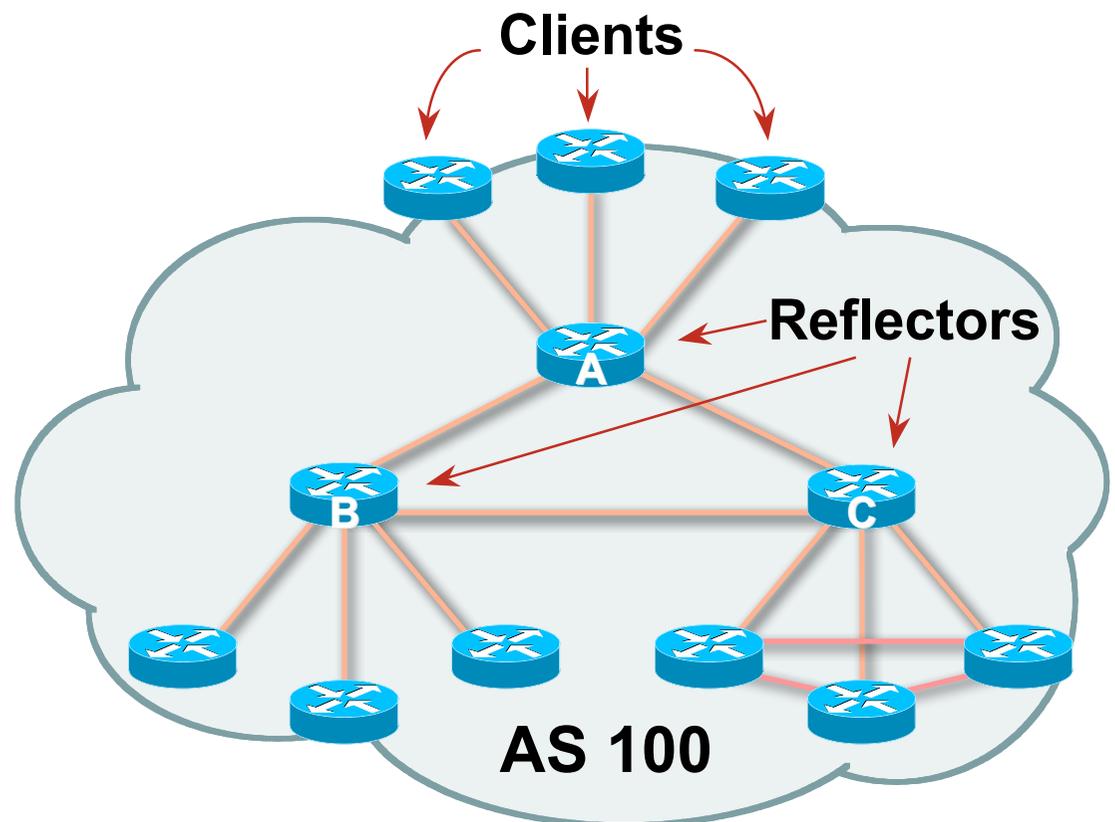
- Two solutions
 - Route reflector – simpler to deploy and run
 - Confederation – more complex, has corner case advantages

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 RFC4456



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 Reflector: 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 Reflector: Redundancy

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

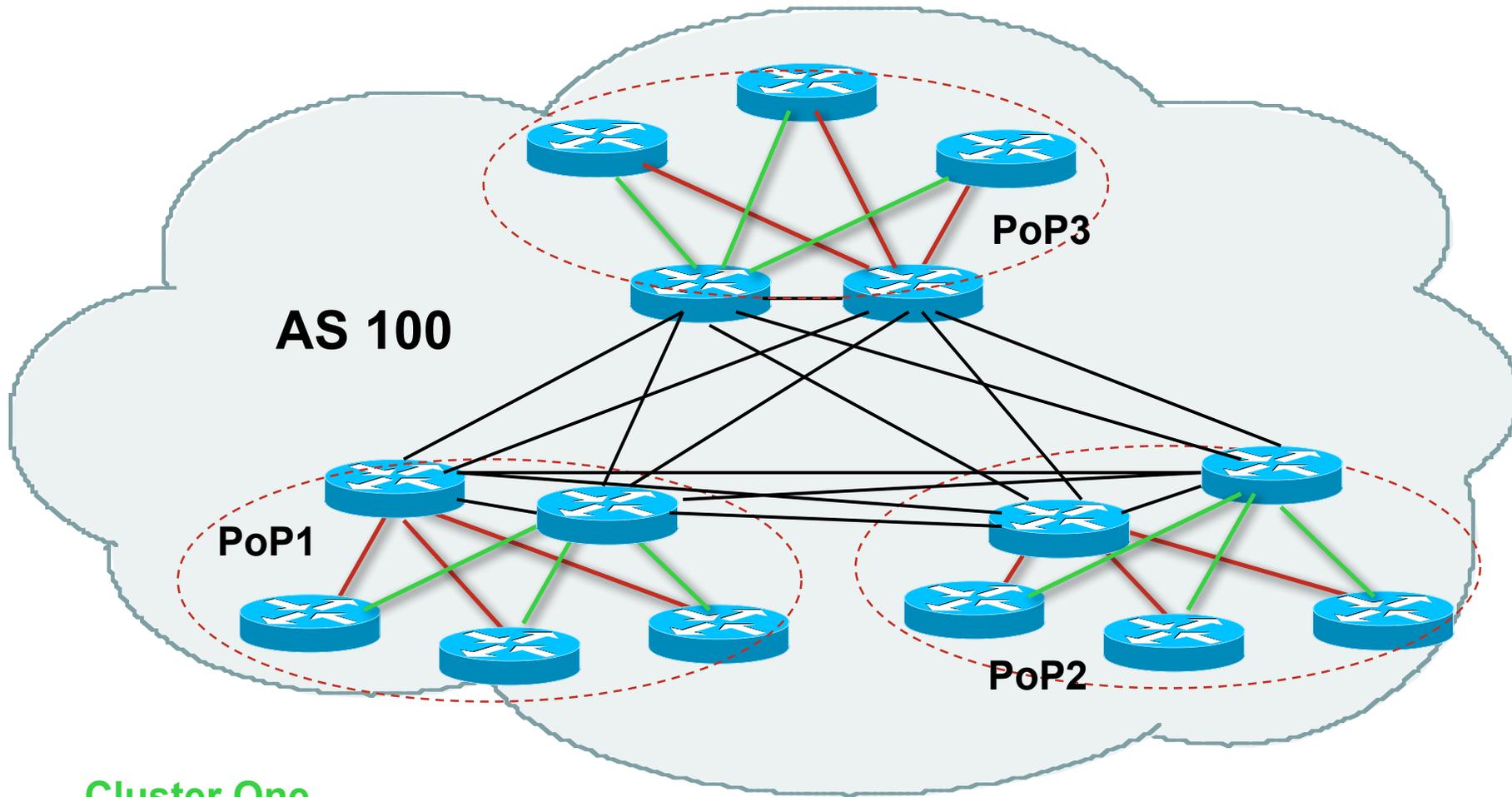
All RRs in the cluster must have the same cluster-id (otherwise it is a different cluster)

- 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 Reflector: Redundancy



Cluster One

Cluster Two

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 Reflector: Deployment

- 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

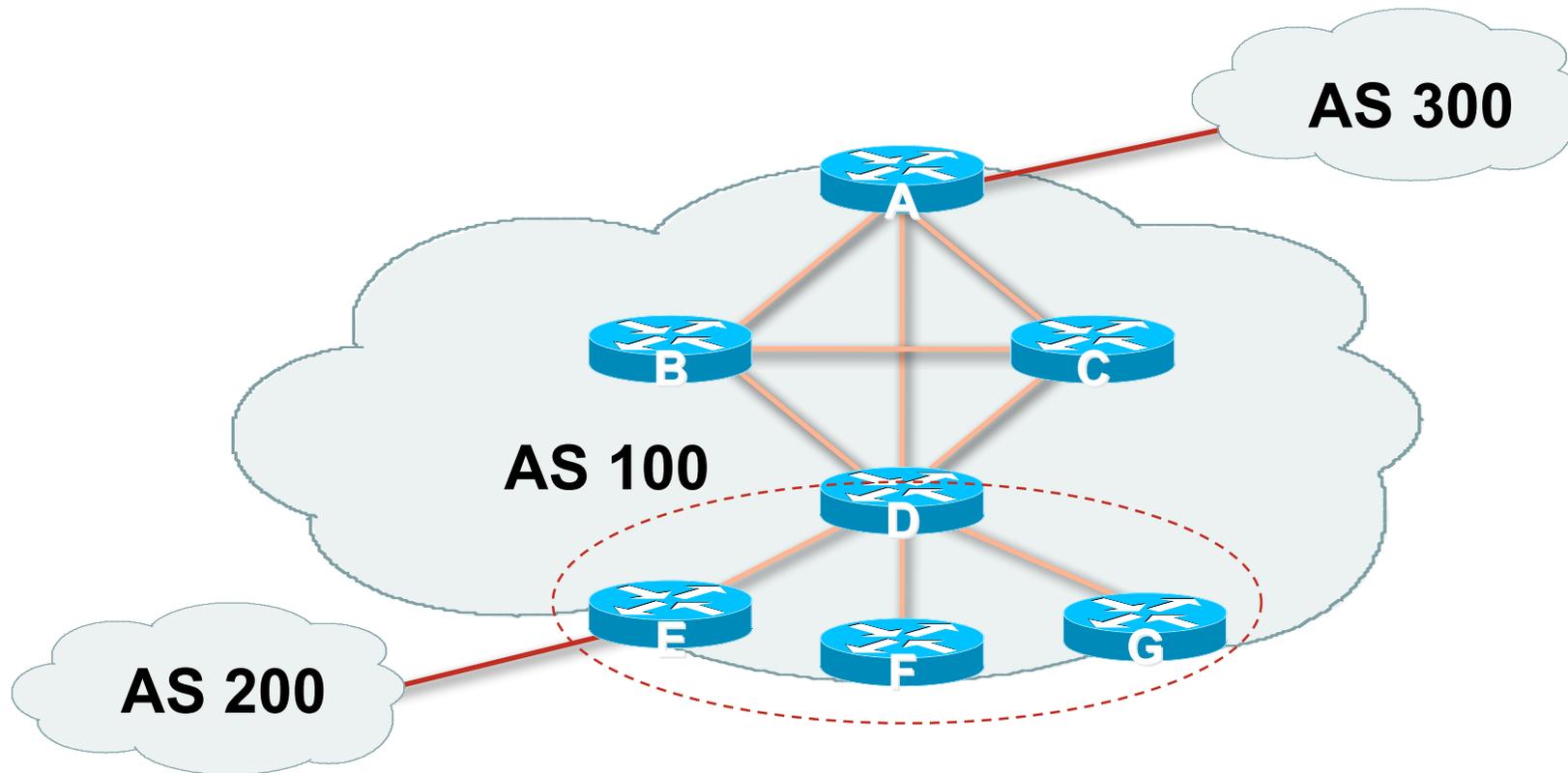
Core routers are RR for the PoP

Two overlaid clusters

Route Reflector: 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 Confederations

Confederations

- Divide the AS into sub-AS
 - eBGP between sub-AS, but some iBGP information is kept
 - Preserve NEXT_HOP across the sub-AS (IGP carries this information)
 - Preserve LOCAL_PREF and MED
- Usually a single IGP
- Described in RFC5065

Confederations (Cont.)

- Visible to outside world as single AS – “Confederation Identifier”

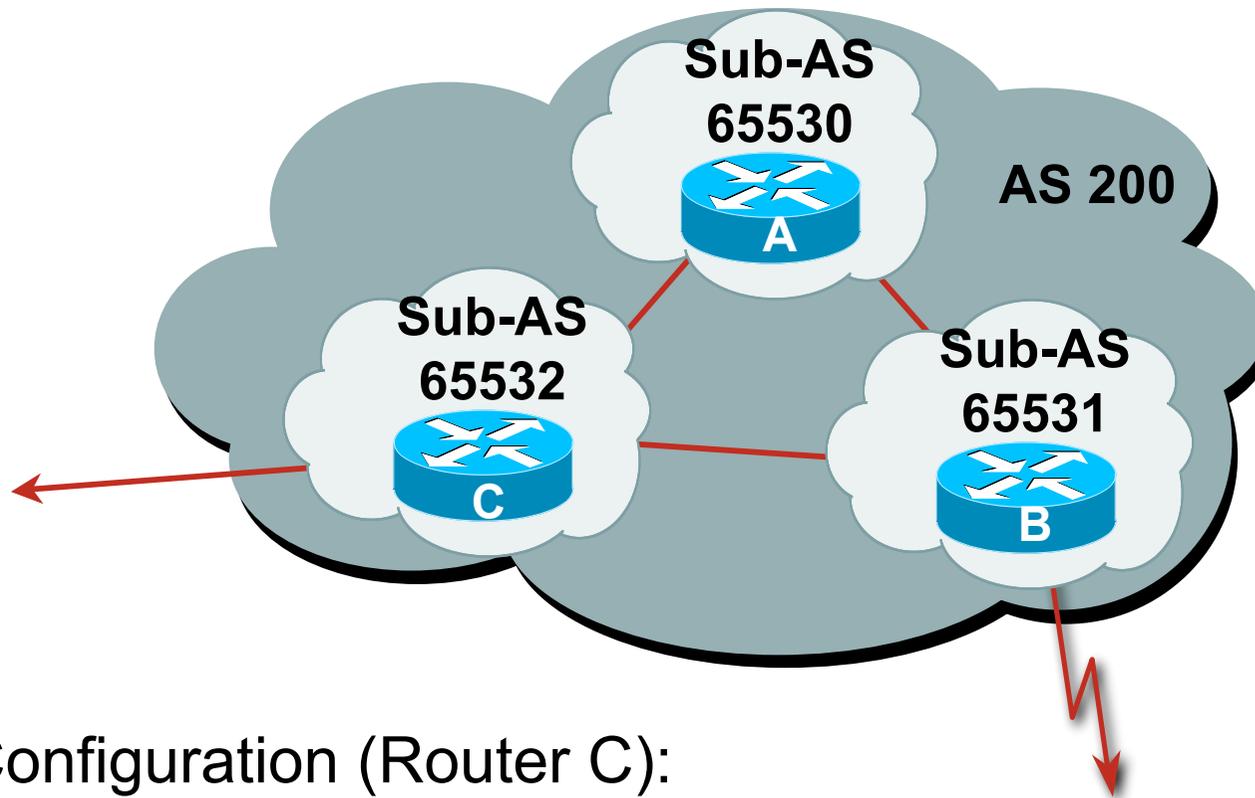
Each sub-AS uses a number from the private AS range (64512-65534)

- iBGP speakers in each sub-AS are fully meshed

The total number of neighbours is reduced by limiting the full mesh requirement to only the peers in the sub-AS

Can also use Route-Reflector within sub-AS

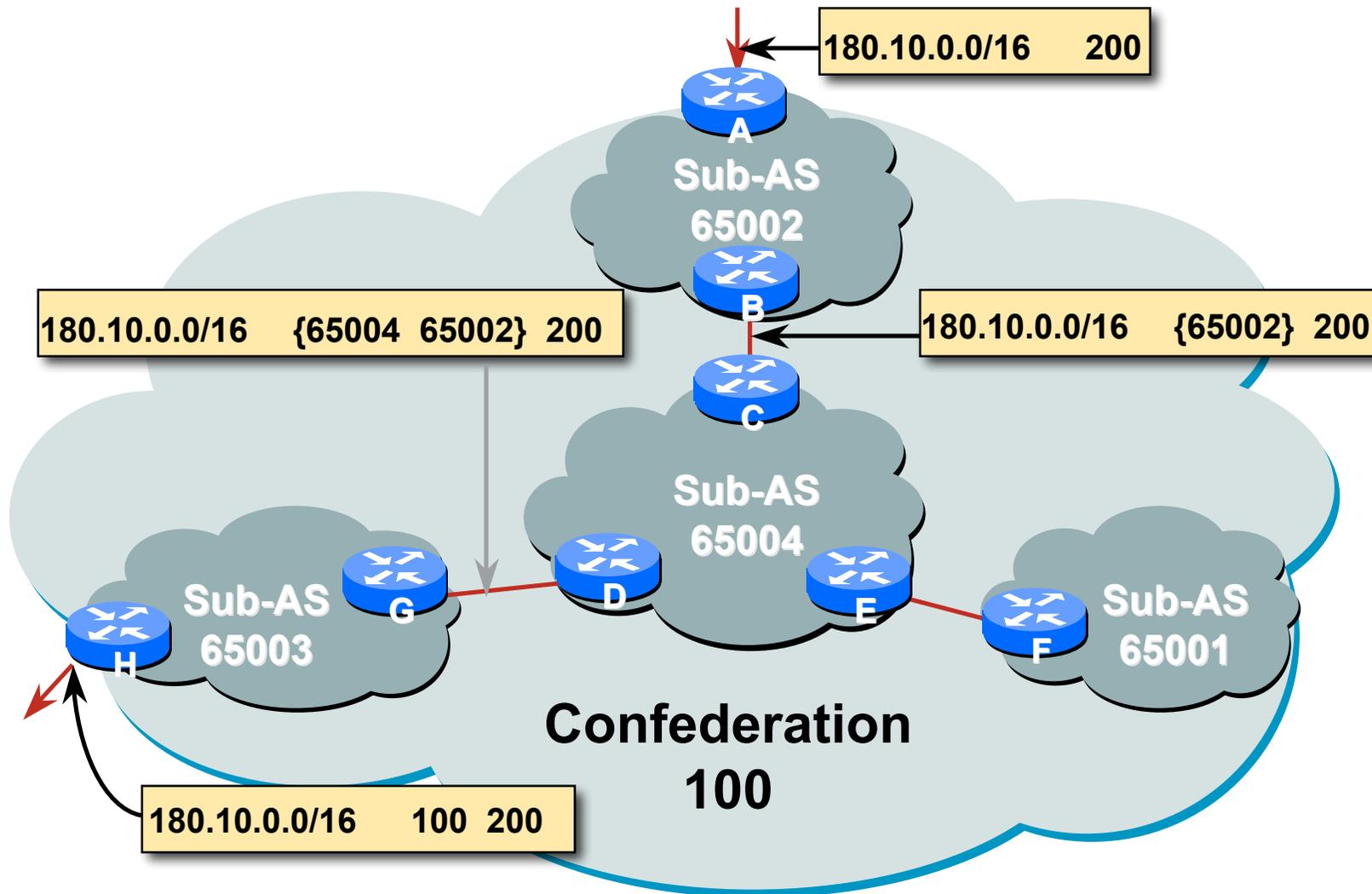
Confederations



- Configuration (Router C):

```
router bgp 65532
  bgp confederation identifier 200
  bgp confederation peers 65530 65531
  neighbor 141.153.12.1 remote-as 65530
  neighbor 141.153.17.2 remote-as 65531
```

Confederations: AS-Sequence



Route Propagation Decisions

- Same as with “normal” BGP:
 - From peer in same sub-AS → only to external peers
 - From external peers → to all neighbors
- “External peers” refers to
 - Peers outside the confederation
 - Peers in a different sub-AS
 - Preserve LOCAL_PREF, MED and NEXT_HOP

RRs or Confederations

	Internet Connectivity	Multi-Level Hierarchy	Policy Control	Scalability	Migration Complexity
Confederations	Anywhere in the Network	Yes	Yes	Medium	Medium to High
Route Reflectors	Anywhere in the Network	Yes	Yes	Very High	Very Low

Most new service provider networks now deploy Route Reflectors from Day One

More points about Confederations

- Can ease “absorbing” other ISPs into you ISP – e.g., if one ISP buys another
 - Or can use AS masquerading feature available in some implementations to do a similar thing
- Can use route-reflectors with confederation sub-AS to reduce the sub-AS iBGP mesh



Route Flap Damping

Network Stability for the 1990s

Network Instability for the 21st Century!

Route Flap Damping

- For many years, Route Flap Damping was a strongly recommended practice
- Now it is strongly discouraged as it appears to cause far greater network instability than it cures
- But first, the theory...

Route Flap Damping

- Route flap

 - Going up and down of path or change in attribute

 - BGP WITHDRAW followed by UPDATE = 1 flap

 - eBGP neighbour going down/up 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

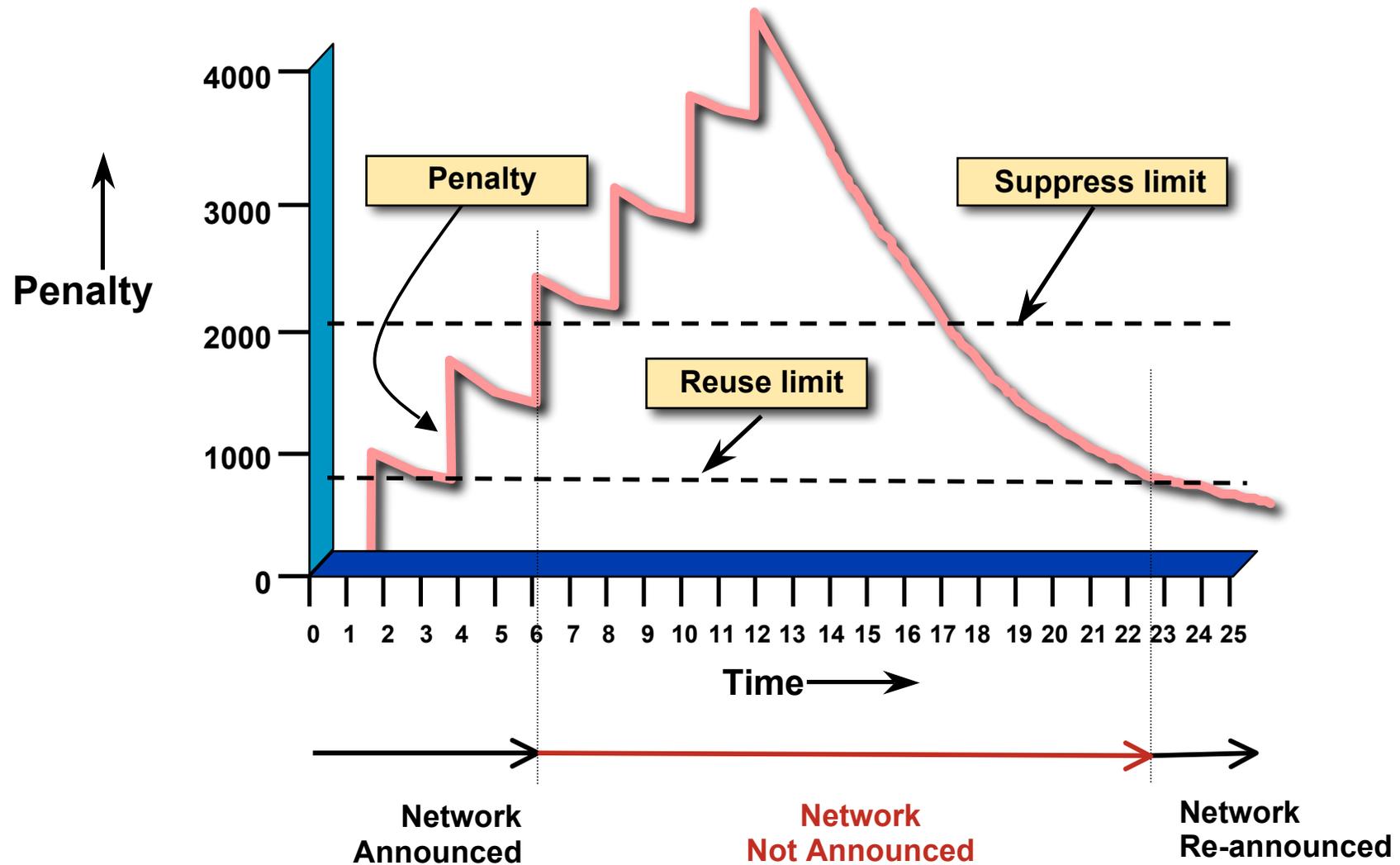
 - Advertise stable routes

- Implementation described in RFC 2439

Operation

- Add penalty (1000) for each flap
 - Change in attribute gets penalty of 500
- 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
 - penalty reset to zero when it is half of reuse-limit

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 and prefix lengths

Route Flap Damping History

- First implementations on the Internet by 1995
- Vendor defaults too severe

RIPE Routing Working Group recommendations in ripe-178, ripe-210, and ripe-229

<http://www.ripe.net/ripe/docs>

But many ISPs simply switched on the vendors' default values without thinking

Serious Problems:

- "Route Flap Damping Exacerbates Internet Routing Convergence"

Zhuoqing Morley Mao, Ramesh Govindan, George Varghese & Randy H. Katz, August 2002

- "What is the sound of one route flapping?"

Tim Griffin, June 2002

- Various work on routing convergence by Craig Labovitz and Abha Ahuja a few years ago

- "Happy Packets"

Closely related work by Randy Bush et al

Problem 1:

- One path flaps:

BGP speakers pick next best path, announce to all peers, flap counter incremented

Those peers see change in best path, flap counter incremented

After a few hops, peers see multiple changes simply caused by a single flap → prefix is suppressed

Problem 2:

- Different BGP implementations have different transit time for prefixes
 - Some hold onto prefix for some time before advertising
 - Others advertise immediately
- Race to the finish line causes appearance of flapping, caused by a simple announcement or path change → prefix is suppressed

Solution:

- Do **NOT** use Route Flap Damping whatever you do!
- RFD will unnecessarily impair access
to your network and
to the Internet
- More information contained in RIPE Routing Working Group recommendations:
[www.ripe.net/ripe/docs/ripe-378.\[pdf,html,txt\]](http://www.ripe.net/ripe/docs/ripe-378.[pdf,html,txt])

BGP for Internet Service Providers

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



Service Provider 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

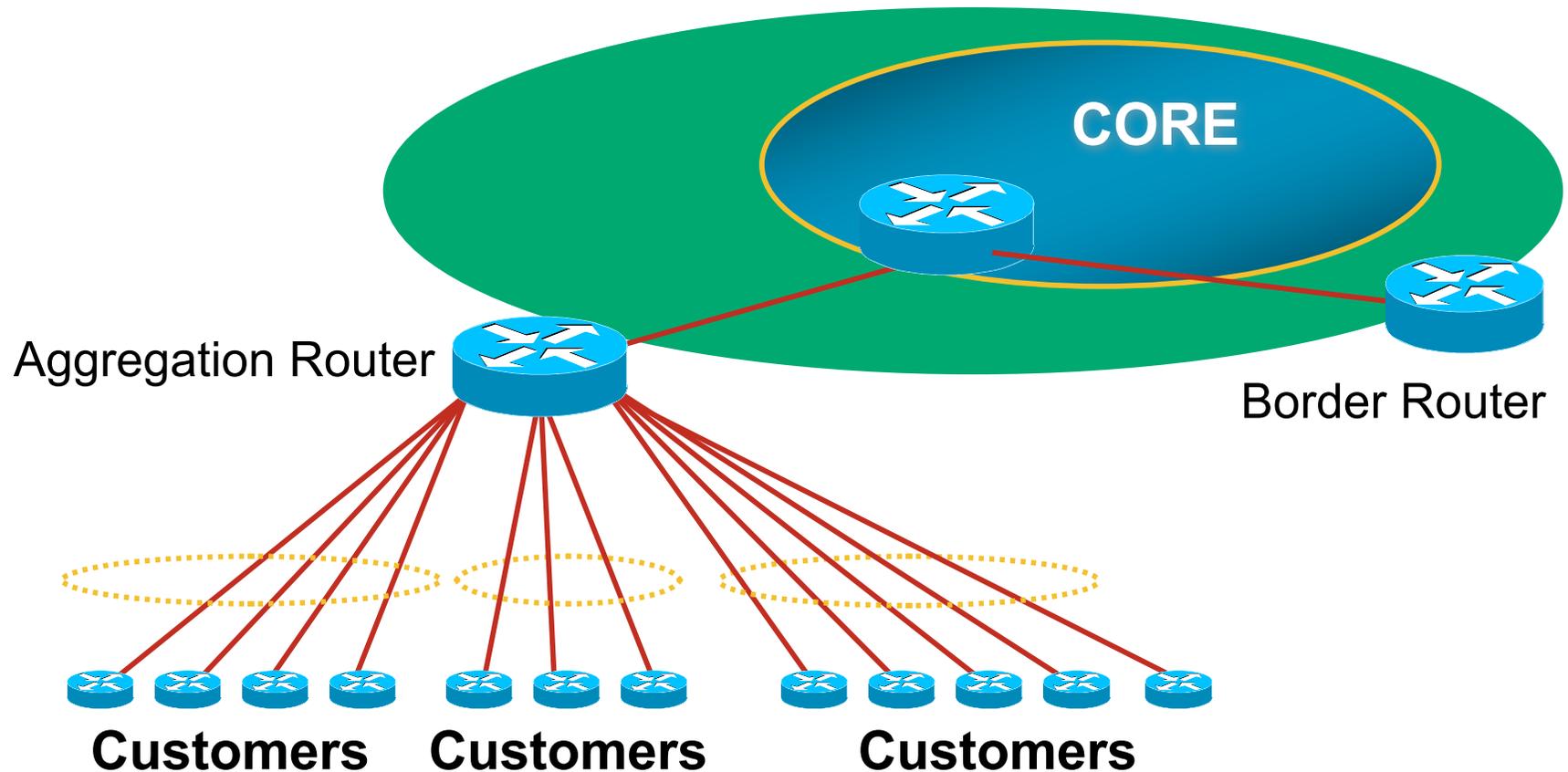
- Community assignments:

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 220k 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

ISP BGP Communities

- There are no recommended ISP BGP communities apart from RFC1998
The five standard communities
www.iana.org/assignments/bgp-well-known-communities
- Efforts have been made to document from time to time
totem.info.ucl.ac.be/publications/papers-elec-versions/draft-quoitin-bgp-comm-survey-00.pdf
But so far... nothing more... ☹️
Collection of ISP communities at www.onesc.net/communities
NANOG Tutorial:
www.nanog.org/meetings/nanog40/presentations/BGPcommunities.pdf
- ISP policy is usually published
On the ISP's website
Referenced in the AS Object in the IRR

Some ISP Examples: Sprintlink

Welcome to Sprint.net

http://www.sprint.net/index.php?module=policies/bgp_policy

Apple .Mac Amazon eBay Yahoo! News (144) Apple (40)

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.

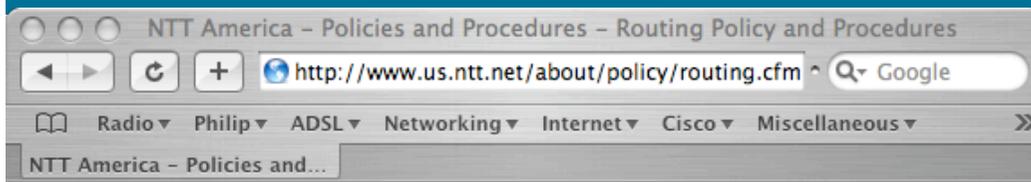
Additionally, Sprint will prepend AS1239 to eBGP sessions with certain autonomous systems depending on a received community. Currently, the following ASes are supported: 1668, 209, 2914, 3300, 3356, 3549, 3561, 4635, 701, 7018, 702 and 8220.

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 ...

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



Some ISP Examples: NTT

BGP customer communities

Customers wanting to alter local preference on their routes.

NTT Communications BGP customers may choose to affect our local preference on their routes by marking their routes with the following communities:

Community	Local-pref	Description
(default)	120	customer
2914:450	96	customer fallback
2914:460	98	peer backup
2914:470	100	peer
2914:480	110	customer backup
2914:490	120	customer default

Customers wanting to alter their route announcements to other customers.

NTT Communications BGP customers may choose to prepend to all other NTT Communications BGP customers with the following communities:

Community	Description
2914:411	prepends o/b to customer 1x
2914:412	prepends o/b to customer 2x
2914:413	prepends o/b to customer 3x

Customers wanting to alter their route announcements to peers.

NTT Communications BGP customers may choose to prepend to all NTT Communications peers with the following communities:

Community	Description
2914:421	prepends o/b to peer 1x
2914:422	prepends o/b to peer 2x

More info at
www.us.ntt.net/about/policy/routing.cfm

Some ISP Examples

AAPT

- Australian ISP
- Run their own Routing Registry
Whois.connect.com.au
- Offer 6 different communities to customers to aid with their traffic engineering

Some ISP Examples

AAPT

```
aut-num:      AS2764
as-name:      ASN-CONNECT-NET
descr:        AAPT Limited
admin-c:      CNO2-AP
tech-c:       CNO2-AP
remarks:      Community support definitions
remarks:
remarks:      Community Definition
remarks:      -----
remarks:      2764:2 Don't announce outside local POP
remarks:      2764:4 Lower local preference by 15
remarks:      2764:5 Lower local preference by 5
remarks:      2764:6 Announce to customers and all peers
                (incl int'l peers), but not transit
remarks:      2764:7 Announce to customers only
remarks:      2764:14 Announce to AANX
notify:       routing@connect.com.au
mnt-by:       CONNECT-AU
changed:      nobody@connect.com.au 20050225
source:       CCAIR
```

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

Some ISP Examples

Verizon Business EMEA

- Verizon Business' European operation
- Permits customers to send communities which determine
 - local preferences within Verizon Business' network
 - Reachability of the prefix
 - How the prefix is announced outside of Verizon Business' network

Some ISP Examples

Verizon Business Europe

```
aut-num: AS702
descr: Verizon Business EMEA - Commercial IP service provider in Eur
remarks: VzBi 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 VzBi AS'es and VzBi customers
       702:30     Keep within Europe, don't announce to other VzBi AS
       702:1      Prepend AS702 once at edges of VzBi to Peers
       702:2      Prepend AS702 twice at edges of VzBi to Peers
       702:3      Prepend AS702 thrice at edges of VzBi 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 VzBi customers.
       702:7001   Prepend AS702 once at edges of VzBi to AS702
                 peers with a scope of National.
       702:7002   Prepend AS702 twice at edges of VzBi to AS702
                 peers with a scope of National.
```

(more)

Some ISP Examples

VzBi Europe

(more)

```
702:7003 Prepend AS702 thrice at edges of VzBi to AS702
        peers with a scope of National.
702:8020 Do not announce to AS702 peers with a scope of
        European but advertise to Global Peers, National
        Peers and VzBi customers.
702:8001 Prepend AS702 once at edges of VzBi to AS702
        peers with a scope of European.
702:8002 Prepend AS702 twice at edges of VzBi to AS702
        peers with a scope of European.
702:8003 Prepend AS702 thrice at edges of VzBi to AS702
        peers with a scope of European.
```

Additional details of the VzBi communities are located at:
<http://www.verizonbusiness.com/uk/customer/bgp/>

```
mnt-by: WCOM-EMEA-RICE-MNT
source: RIPE
```

Some ISP Examples

BT Ignite

- One of the most comprehensive community lists around
 - Seems to be based on definitions originally used in Tiscali's network
 - `whois -h whois.ripe.net AS5400` reveals all
- Extensive community definitions allow sophisticated traffic engineering by customers

Some ISP Examples

BT Ignite

```
aut-num:      AS5400
descr:       BT Ignite European Backbone
remarks:
remarks:     Community to                               Community to
remarks:     Not announce                               To peer:                               AS prepend 5400
remarks:
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:     5400:1506 GlobalCrossing Trans (AS3549)                               5400:2506
remarks:
remarks:     5400:1001 Nexica (AS24592)                               5400:2001
remarks:     5400:1002 Fujitsu (AS3324)                               5400:2002
remarks:     5400:1004 C&W EU (1273)                               5400:2004
<snip>
notify:      notify@eu.bt.net
mnt-by:      CIP-MNT
source:      RIPE
```



Some ISP Examples Level 3

- Highly detailed AS object held on the RIPE Routing Registry
- Also a very comprehensive list of community definitions
`whois -h whois.ripe.net AS3356` reveals all

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
<snip>
remarks:         3356:70   - set local preference to 70
remarks:         3356:80   - set local preference to 80
remarks:         3356:90   - set local preference to 90
remarks:         3356:9999 - blackhole (discard) traffic
<snip>
mnt-by:          LEVEL3-MNT
source:          RIPE
```

And many
many more!

BGP for Internet Service Providers

- BGP Basics
- 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

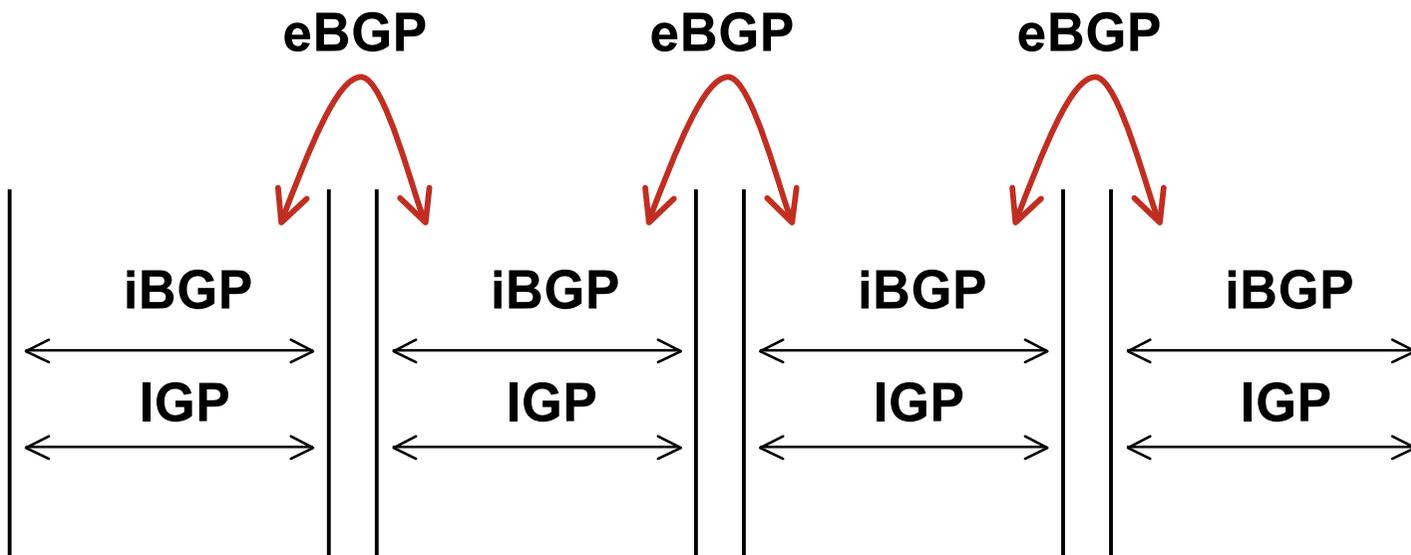
design goal is to **minimise** number of prefixes in IGP to aid scalability and rapid convergence

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

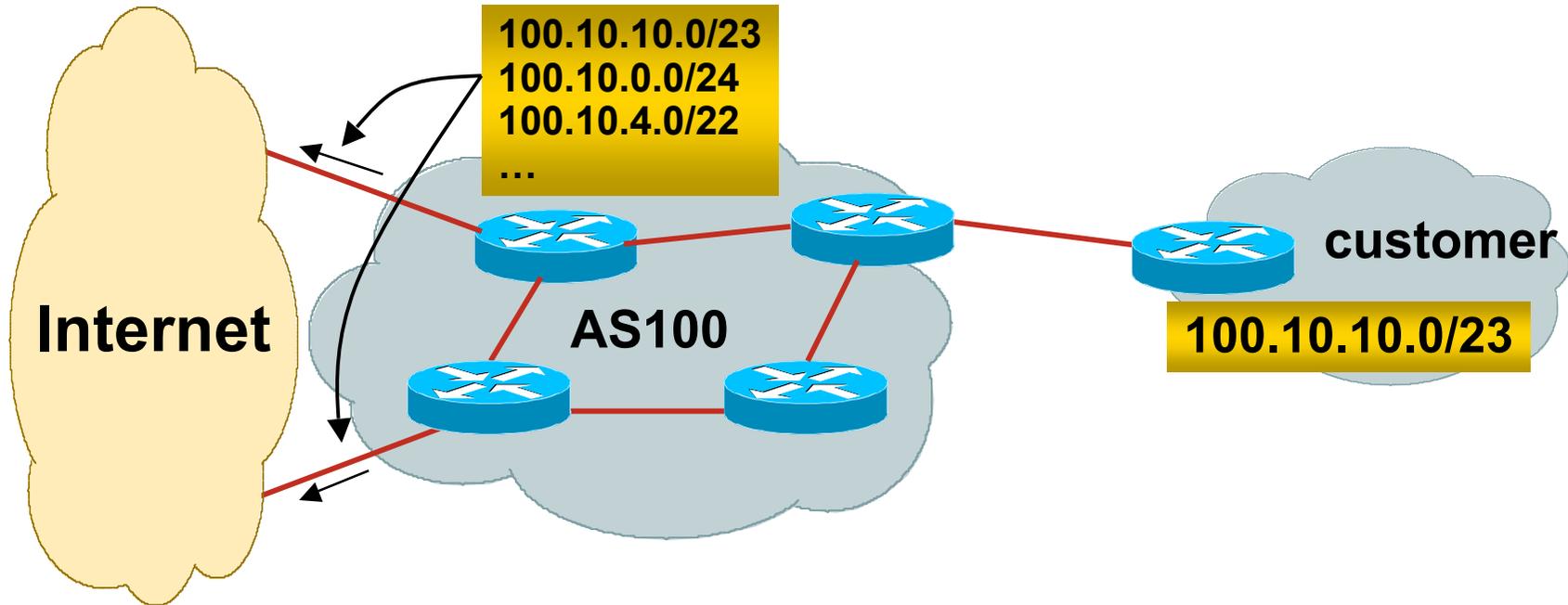
Aggregation

- 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 >146000 /24s!

Aggregation – Example



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

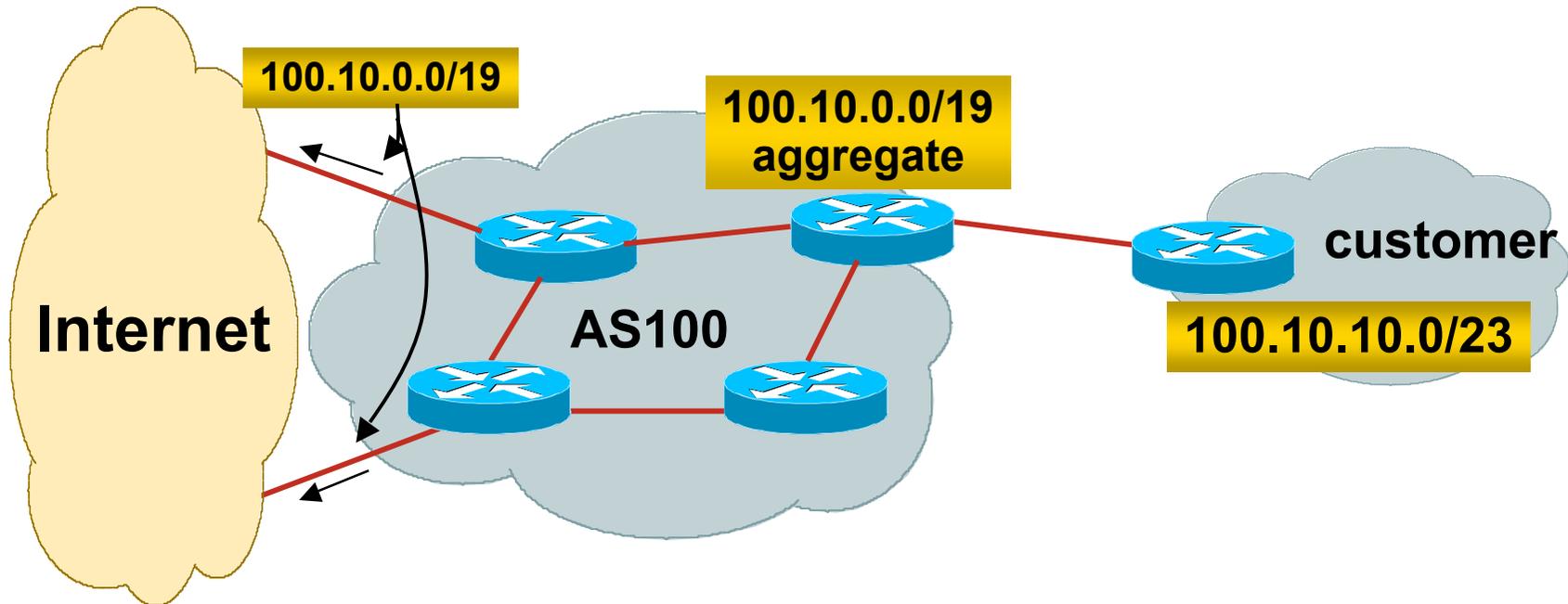
Aggregation – Bad Example

- 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

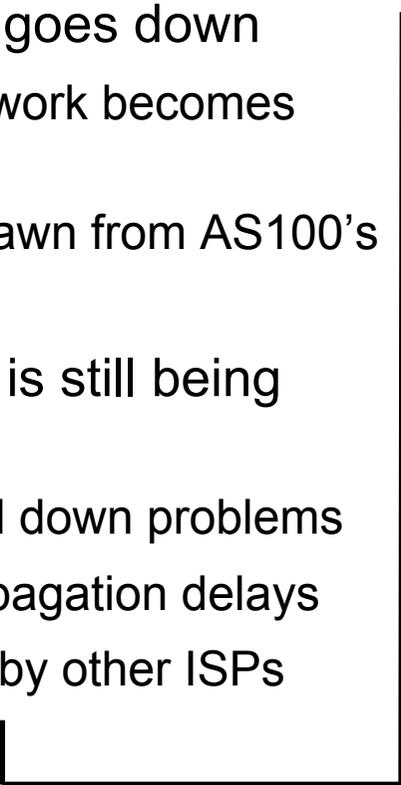
- Their /23 network is now visible to their ISP
- Their /23 network is re-advertised to peers
- Starts rippling through Internet
- Load on Internet backbone routers as network is reinserted into routing table
- Some ISP's suppress the flaps
- Internet may take 10-20 min or longer to be visible
- Where is the Quality of Service???

Aggregation – 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 – 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 2009)

- Current Internet Routing Table Statistics

BGP Routing Table Entries	280535
Prefixes after maximum aggregation	133420
Unique prefixes in Internet	137533
Prefixes smaller than registry alloc	137717
/24s announced	146883
only 5812 /24s are from 192.0.0.0/8	
ASes in use	30610

“The New Swamp”

- Swamp space is name used for areas of poor aggregation

The original swamp was 192.0.0.0/8 from the former class C block

Name given just after the deployment of CIDR

The new swamp is creeping across all parts of the Internet

Not just RIR space, but “legacy” space too

“The New Swamp”

RIR Space – February 1999

RIR blocks contribute 49393 prefixes or 88% of the Internet Routing Table

Block	Networks	Block	Networks	Block	Networks	Block	Networks
24/8	165	79/8	0	118/8	0	201/8	0
41/8	0	80/8	0	119/8	0	202/8	2276
58/8	0	81/8	0	120/8	0	203/8	3622
59/8	0	82/8	0	121/8	0	204/8	3792
60/8	0	83/8	0	122/8	0	205/8	2584
61/8	3	84/8	0	123/8	0	206/8	3127
62/8	87	85/8	0	124/8	0	207/8	2723
63/8	20	86/8	0	125/8	0	208/8	2817
64/8	0	87/8	0	126/8	0	209/8	2574
65/8	0	88/8	0	173/8	0	210/8	617
66/8	0	89/8	0	174/8	0	211/8	0
67/8	0	90/8	0	186/8	0	212/8	717
68/8	0	91/8	0	187/8	0	213/8	1
69/8	0	96/8	0	189/8	0	216/8	943
70/8	0	97/8	0	190/8	0	217/8	0
71/8	0	98/8	0	192/8	6275	218/8	0
72/8	0	99/8	0	193/8	2390	219/8	0
73/8	0	112/8	0	194/8	2932	220/8	0
74/8	0	113/8	0	195/8	1338	221/8	0
75/8	0	114/8	0	196/8	513	222/8	0
76/8	0	115/8	0	198/8	4034		
77/8	0	116/8	0	199/8	3495		
78/8	0	117/8	0	200/8	1348		

“The New Swamp”

RIR Space – February 2009

RIR blocks contribute 245261 prefixes or 87% of the Internet Routing Table

Block	Networks	Block	Networks	Block	Networks	Block	Networks
24/8	3132	79/8	1018	118/8	1084	201/8	3847
41/8	2642	80/8	2271	119/8	1282	202/8	11142
58/8	1531	81/8	1740	120/8	418	203/8	11261
59/8	1582	82/8	1473	121/8	1480	204/8	5527
60/8	932	83/8	1297	122/8	2008	205/8	3129
61/8	2814	84/8	1327	123/8	1753	206/8	3810
62/8	2443	85/8	2522	124/8	1942	207/8	4484
63/8	3447	86/8	790	125/8	2302	208/8	6536
64/8	6249	87/8	1430	126/8	71	209/8	5600
65/8	4413	88/8	962	173/8	991	210/8	4700
66/8	7112	89/8	3168	174/8	199	211/8	2810
67/8	3514	90/8	311	186/8	305	212/8	3353
68/8	2681	91/8	3698	187/8	516	213/8	3449
69/8	4698	96/8	518	189/8	2259	216/8	7728
70/8	1870	97/8	567	190/8	5319	217/8	2710
71/8	1490	98/8	875	192/8	6990	218/8	1360
72/8	3882	99/8	249	193/8	6382	219/8	1332
73/8	8	112/8	229	194/8	5056	220/8	2308
74/8	3926	113/8	409	195/8	5003	221/8	1028
75/8	1181	114/8	615	196/8	1753	222/8	1223
76/8	1030	115/8	856	198/8	4673		
77/8	1822	116/8	1808	199/8	4177		
78/8	1370	117/8	1237	200/8	8822		

“The New Swamp” Summary

- RIR space shows creeping deaggregation

It seems that an RIR /8 block averages around 5000 prefixes once fully allocated

So their existing 95 /8s will eventually cause 440000 prefix announcements

- Food for thought:

Remaining 32 unallocated /8s and the 88 RIR /8s combined will cause:

635000 prefixes with 5000 prefixes per /8 density

762000 prefixes with 6000 prefixes per /8 density

Plus 12% due to “non RIR space deaggregation”

→ Routing Table size of 853440 prefixes

“The New Swamp” Summary

- Rest of address space is showing similar deaggregation too 😞
- What are the reasons?
 - Main justification is traffic engineering
- Real reasons are:
 - Lack of knowledge
 - Laziness
 - Deliberate & knowing actions

BGP Report (bgp.potaroo.net)

- 199336 total announcements in October 2006
- 129795 prefixes
 - After aggregating including full AS PATH info
i.e. including each ASN's traffic engineering
 - 35% saving possible
- 109034 prefixes
 - After aggregating by Origin AS
i.e. ignoring each ASN's traffic engineering
 - 10% saving possible

Deaggregation: The Excuses

- Traffic engineering causes 10% of the Internet Routing table
- Deliberate deaggregation causes 35% of the Internet Routing table

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

- RIPE Routing WG aggregation recommendation

RIPE-399 — <http://www.ripe.net/ripe/docs/ripe-399.html>

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

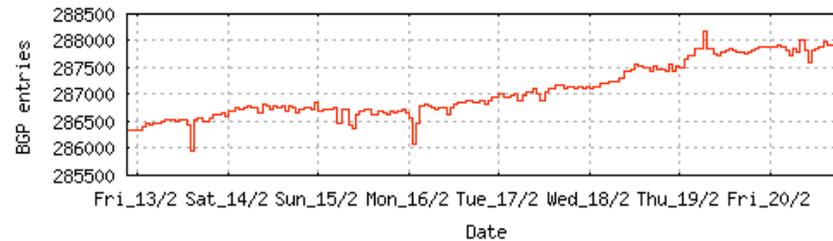
Very effectively challenges the traffic engineering excuse

Status Summary

Table History

Date	Prefixes	CIDR Aggregated
13-02-09	286337	178222
14-02-09	286606	178525
15-02-09	286864	178547
16-02-09	286675	178825
17-02-09	286961	178737
18-02-09	287148	178948
19-02-09	287534	179208
20-02-09	287886	179260

Plot: [BGP Table Size](#)



AS Summary

30724	Number of ASes in routing system
13062	Number of ASes announcing only one prefix
4368	Largest number of prefixes announced by an AS
	AS6389 : BELLSOUTH-NET-BLK - BellSouth.net Inc.
89816320	Largest address span announced by an AS (/32s)
	AS27064 : DDN-ASNBLK1 - DoD Network Information Center

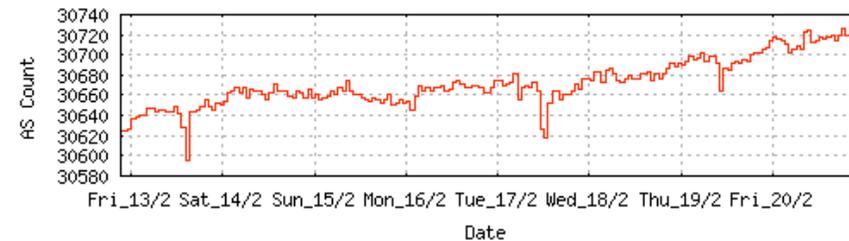
Plot: [AS count](#)

Plot: [Average announcements per origin AS](#)

Report: [ASes ordered by originating address span](#)

Report: [ASes ordered by transit address span](#)

Report: [Autonomous System number-to-name mapping \(from Registry WHOIS data\)](#)



Aggregation Summary

The algorithm used in this report proposes aggregation only when there is a precise match using AS path so as to preserve traffic transit policies. Aggregation is also proposed across non-advertised address space ('holes').

--- 20Feb09 ---

ASnum NetsNow NetsAggr NetGain % Gain Description

ASnum	NetsNow	NetsAggr	NetGain	% Gain	Description
Table	287785	179251	108534	37.7%	All ASes
AS6389	4368	353	4015	91.9%	BELLSOUTH-NET-BLK - BellSouth.net Inc.
AS4323	4224	1816	2408	57.0%	TWTC - tw telecom holdings, inc.
AS209	2833	1258	1575	55.6%	ASN-QWEST - Qwest Communications Corporation
AS4766	1815	524	1291	71.1%	KIXS-AS-KR Korea Telecom
AS17488	1521	368	1153	75.8%	HATHWAY-NET-AP Hathway IP Over Cable Internet
AS4755	1212	233	979	80.8%	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
AS22773	1019	60	959	94.1%	ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
AS8452	1225	307	918	74.9%	TEDATA TEDATA
AS8151	1477	609	868	58.8%	Uninet S.A. de C.V.
AS1785	1748	928	820	46.9%	AS-PAETEC-NET - PaeTec Communications, Inc.
AS19262	953	244	709	74.4%	VZGNI-TRANSIT - Verizon Internet Services Inc.
AS11492	1144	476	668	58.4%	CABLEONE - CABLE ONE, INC.
AS18566	1061	411	650	61.3%	COVAD - Covad Communications Co.
AS18101	752	165	587	78.1%	RIL-IDC Reliance Infocom Ltd Internet Data Centre,
AS3356	1141	558	583	51.1%	LEVEL3 Level 3 Communications
AS6478	1250	681	569	45.5%	ATT-INTERNET3 - AT&T WorldNet Services
AS7545	746	191	555	74.4%	TPG-INTERNET-AP TPG Internet Pty Ltd
AS17908	604	112	492	81.5%	TCISL Tata Communications
AS22047	605	118	487	80.5%	VTR BANDA ANCHA S.A.
AS2706	550	80	470	85.5%	HKSUPER-HK-AP Pacific Internet (Hong Kong) Limited
AS855	619	161	458	74.0%	CANET-ASN-4 - Bell Aliant
AS4808	615	157	458	74.5%	CHINA169-BJ CNCGROUP IP network China169 Beijing Province Network
AS24560	673	238	435	64.6%	AIRTELBROADBAND-AS-AP Bharti Airtel Ltd., Telemedia Services
AS7018	1444	1011	433	30.0%	ATT-INTERNET4 - AT&T WorldNet Services
AS4134	927	499	428	46.2%	CHINANET-BACKBONE No.31,Jin-rong Street
AS4668	703	285	418	59.5%	LGNET-AS-KR LG CNS
AS9443	507	91	416	82.1%	INTERNETPRIMUS-AS-AP Primus Telecommunications



Top 20 Added Routes this week per Originating AS

Prefixes	ASnum	AS Description
159	AS8452	TEDATA TEDATA
86	AS3644	SPR-VPN - Sprint
64	AS9658	ETPI-IDS-AS-AP Eastern Telecoms Phils., Inc.
56	AS237	MERIT-AS-14 - Merit Network Inc.
50	AS4766	KIXS-AS-KR Korea Telecom
49	AS5056	INS-NET-2 - Iowa Network Services
35	AS9583	SIFY-AS-IN Sify Limited
32	AS20299	Newcom Limited
30	AS13789	INTERNAP-BLK3 - Internap Network Services Corporation
29	AS1785	AS-PAETEC-NET - PaeTec Communications, Inc.
25	AS18207	IQARA-INDIA-AP Iqara Telecom India Pvt Ltd.
25	AS12182	INTERNAP-2BLK - Internap Network Services Corporation
25	AS32035	CCDT-AS - Telekenex
25	AS47931	ALENETWORK A.L.E. COM NETWORK S.R.L
24	AS4134	CHINANET-BACKBONE No.31,Jin-rong Street
22	AS3216	SOVAM-AS Golden Telecom, Moscow, Russia
20	AS20858	EGYNET-AS
19	AS45379	UNIST-AS-KR Ulsan National Institute of Science and Technology
18	AS23693	TELKOMSEL-ASN-ID PT. Telekomunikasi Selular
17	AS6458	Telgua

Top 20 Withdrawn Routes this week per Originating AS

Prefixes	ASnum	AS Description
-87	AS11492	CABLEONE - CABLE ONE, INC.
-86	AS15471	SNR-RO SNR - Societatea Nationala de Radiocomunicatii
-66	AS1785	AS-PAETEC-NET - PaeTec Communications, Inc.
-46	AS18809	Cable Onda
-36	AS14363	OUTFITTERS - Infobahn Outfitters, Inc.
-21	AS17816	CHINA169-GZ CNCGROUP IP network China169 Guangzhou MAN
-20	AS29545	IPLACE iPlace Internet & Network Services GmbH
-19	AS4787	ASN-CBN Internet Service Provider
-18	AS36351	SOFTLAYER - SoftLayer Technologies Inc.
-18	AS8452	TEDATA TEDATA
-18	AS15691	Leonet Srl Autonomous System
-18	AS12654	BDFE-NCC-BIS-AS BDFE NCC BIS project

More Specifics

A list of route advertisements that appear to be more specific than the original Class-based prefix mask, or more specific than the registry allocation size.

Top 20 ASes advertising more specific prefixes

More Specifics	Total Prefixes	ASnum	AS Description
4220	4368	AS6389	BELLSOUTH-NET-BLK - BellSouth.net Inc.
4018	4224	AS4323	TWTC - tw telecom holdings, inc.
2629	2833	AS209	ASN-QWEST - Qwest Communications Corporation
1766	1815	AS4766	KIXS-AS-KR Korea Telecom
1660	1748	AS1785	AS-PAETEC-NET - PaeTec Communications, Inc.
1536	1583	AS20115	CHARTER-NET-HKY-NC - Charter Communications
1520	1521	AS17488	HATHWAY-NET-AP Hathway IP Over Cable Internet
1470	1477	AS8151	Uninet S.A. de C.V.
1250	1250	AS6478	ATT-INTERNET3 - AT&T WorldNet Services
1198	1212	AS4755	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
1178	1444	AS7018	ATT-INTERNET4 - AT&T WorldNet Services
1156	1265	AS2386	INS-AS - AT&T Data Communications Services
1137	1144	AS11492	CABLEONE - CABLE ONE, INC.
1090	1091	AS9583	SIFY-AS-IN Sify Limited
1051	1061	AS18566	COVAD - Covad Communications Co.
981	1019	AS22773	ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
960	966	AS7011	FRONTIER-AND-CITIZENS - Frontier Communications of America, Inc.
947	947	AS23577	ATM-MPLS-AS-KR Korea Telecom
884	1141	AS3356	LEVEL3 Level 3 Communications
882	953	AS19262	VZGNI-TRANSIT - Verizon Internet Services Inc.

Report: [ASes ordered by number of more specific prefixes](#)

Report: [More Specific prefix list \(by AS\)](#)

Report: [More Specific prefix list \(ordered by prefix\)](#)

Announced Prefixes

Rank	AS	Type	Originate	Addr Space (pfx)	Transit	Addr space (pfx)	Description
141	AS4755		ORG+TRN Originate:	2434560 /10.78	Transit:	5148928 /9.70	TATACOMM-AS TATA Communications formerly VSNL

Aggregation Suggestions

This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

Rank	AS	AS Name	Current	Wthdw	Aggte	Annce	Redctn	%
7	AS4755	TATACOMM-AS TATA Communications formerly VSNL	1212	1029	50	233	979	80.78%

Prefix	AS Path	Aggregation Suggestion
59.151.144.0/22	4777 2516 4755	
59.160.0.0/16	4777 2516 4755	
59.160.0.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.4.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.5.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.8.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.12.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.15.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.16.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.24.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.24.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.28.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.32.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.38.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.40.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.44.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.46.0/23	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.48.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.48.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.56.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.64.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.71.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.72.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.73.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.81.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.82.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.83.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.88.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.88.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.89.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755
59.160.96.0/20	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 4755

Announced Prefixes

Rank	AS	Type	Originate	Addr Space (pfx)	Transit	Addr space (pfx)	Description
149	AS18566	ORIGIN	Originate:	2352896 /10.83	Transit:	0 /0.00	COVAD - Covad Communications Co.

Aggregation Suggestions

This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

Rank	AS	AS Name	Current	Withdw	Aggte	Annce	Redctn	%
14	AS18566	COVAD - Covad Communications Co.	1061	735	85	411	650	61.26%

Prefix	AS Path	Aggregation Suggestion
64.105.0.0/16	4777 2497 2828 18566	
64.105.0.0/23	4777 2516 3356 18566	
64.105.4.0/23	4777 2516 3356 18566	
64.105.6.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.8.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.10.0/23	4777 2516 3356 18566	
64.105.14.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.16.0/24	4777 2516 3356 18566	
64.105.17.0/24	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.18.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.20.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.22.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.24.0/21	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.32.0/21	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.40.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.42.0/23	4777 2516 3356 18566	
64.105.44.0/23	4777 2516 3356 18566	
64.105.46.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.48.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.50.0/23	4777 2516 3356 18566	
64.105.52.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.54.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.56.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.58.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.60.0/22	4777 2516 3356 18566	+ Announce - aggregate of 64.105.60.0/23 (4777 2516 3356 18566) and 64.105.62.0/23 (4777 2516 3356 18566)
64.105.60.0/23	4777 2516 3356 18566	- Withdrawn - aggregated with 64.105.62.0/23 (4777 2516 3356 18566)
64.105.62.0/23	4777 2516 3356 18566	- Withdrawn - aggregated with 64.105.60.0/23 (4777 2516 3356 18566)
64.105.64.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566
64.105.66.0/23	4777 2516 3356 18566	
64.105.68.0/23	4777 2516 3356 18566	
64.105.70.0/23	4777 2497 2828 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2497 2828 18566

Importance of Aggregation

- Size of routing table

 - Memory is no longer a problem

 - Routers can be specified to carry 1 million prefixes

- Convergence of the Routing System

 - This is a problem

 - Bigger table takes longer for CPU to process

 - BGP updates take longer to deal with

 - BGP Instability Report tracks routing system update activity

 - <http://bgpupdates.potaroo.net/instability/bgpupd.html>

The BGP Instability Report

The BGP Instability Report is updated daily. This report was generated on 20 February 2009 06:54 (UTC+1000)

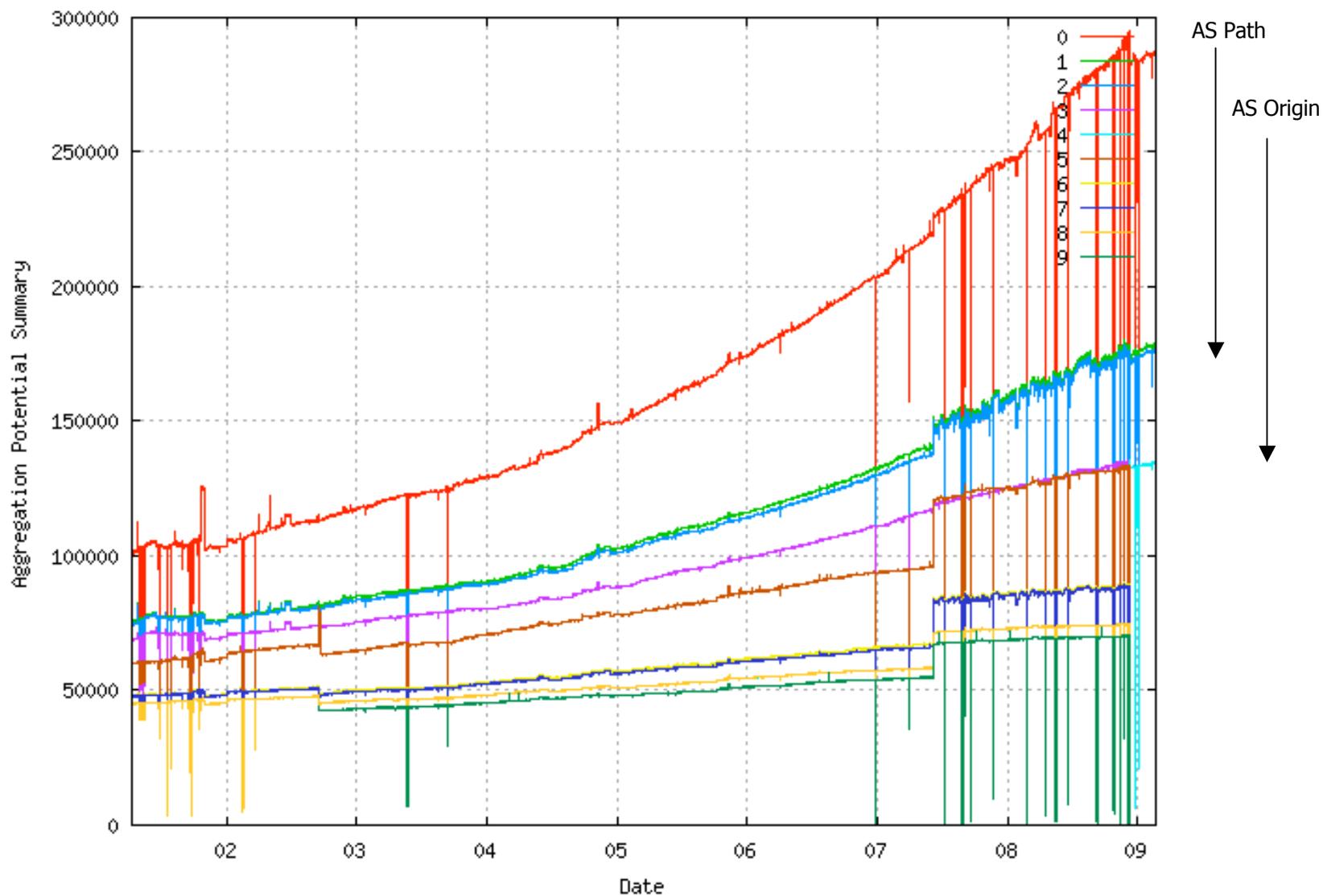
50 Most active ASes for the past 31 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	9583	252388	5.41%	1494	168.93	SIFY-AS-IN Sify Limited
2	7643	167194	3.59%	601	278.19	VNN-AS-AP Vietnam Posts and Telecommunications (VNPT)
3	30890	52239	1.12%	445	117.39	EVOLVA Evolva Telecom
4	6629	48437	1.04%	65	745.18	NOAA-AS - NOAA
5	35805	35061	0.75%	377	93.00	UTG-AS United Telecom AS
6	17974	33362	0.72%	504	66.19	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
7	6458	30697	0.66%	465	66.02	Telgua
8	27757	30429	0.65%	124	245.40	ANDINATEL S.A.
9	30306	28205	0.61%	4	7051.25	AfOL-Sz-AS
10	12500	26532	0.57%	4	6633.00	RCS-AS RCS Autonomus System
11	5050	24213	0.52%	59	410.39	PSC-EXT - Pittsburgh Supercomputing Center
12	30969	23596	0.51%	8	2949.50	TAN-NET TransAfrica Networks
13	16559	22991	0.49%	9	2554.56	REALCONNECT-01 - RealConnect, Inc
14	5056	21623	0.46%	116	186.41	INS-NET-2 - Iowa Network Services
15	8452	21514	0.46%	1469	14.65	TEDATA TEDATA
16	14420	21438	0.46%	244	87.86	ANDINATEL S.A.
17	20115	21021	0.45%	2069	10.16	CHARTER-NET-HKY-NC - Charter Communications
18	17488	20962	0.45%	1542	13.59	HATHWAY-NET-AP Hathway IP Over Cable Internet
19	8151	20618	0.44%	1497	13.77	Uninet S.A. de C.V.
20	6389	20014	0.43%	4400	4.55	BELLSOUTH-NET-BLK - BellSouth.net Inc.
21	9829	19606	0.42%	639	30.68	BSNL-NIB National Internet Backbone
22	1785	19399	0.42%	1891	10.26	AS-PAETEC-NET - PaeTec Communications, Inc.
23	23966	19069	0.41%	368	51.82	LDN-AS-AP LINKdotNET Telecom Limited

50 Most active Prefixes for the past 31 days

RANK	PREFIX	UPDs	%	Origin AS -- AS NAME
1	72.23.246.0/24	23982	0.47%	5050 -- PSC-EXT - Pittsburgh Supercomputing Center
2	221.134.32.0/24	23761	0.47%	9583 -- SIFY-AS-IN Sify Limited
3	190.152.103.0/24	17617	0.35%	27757 -- ANDINATEL S.A.
4	192.35.129.0/24	16220	0.32%	6629 -- NOAA-AS - NOAA
5	192.102.88.0/24	16080	0.32%	6629 -- NOAA-AS - NOAA
6	198.77.177.0/24	15939	0.31%	6629 -- NOAA-AS - NOAA
7	221.135.107.0/24	15250	0.30%	9583 -- SIFY-AS-IN Sify Limited
8	210.214.177.0/24	15104	0.30%	9583 -- SIFY-AS-IN Sify Limited
9	210.214.146.0/24	14974	0.29%	9583 -- SIFY-AS-IN Sify Limited
10	210.214.222.0/24	14862	0.29%	9583 -- SIFY-AS-IN Sify Limited
11	210.214.232.0/24	14844	0.29%	9583 -- SIFY-AS-IN Sify Limited
12	210.214.132.0/24	14821	0.29%	9583 -- SIFY-AS-IN Sify Limited
13	210.214.117.0/24	14773	0.29%	9583 -- SIFY-AS-IN Sify Limited
14	64.162.116.0/24	14428	0.28%	5033 -- ISW - Internet Specialties West Inc.
15	212.85.223.0/24	14010	0.28%	30306 -- AfOL-Sz-AS
16	212.85.220.0/24	13975	0.27%	30306 -- AfOL-Sz-AS
17	210.214.184.0/24	13450	0.26%	9583 -- SIFY-AS-IN Sify Limited
18	210.18.10.0/24	13306	0.26%	9583 -- SIFY-AS-IN Sify Limited
19	210.210.127.0/24	13025	0.26%	9583 -- SIFY-AS-IN Sify Limited
20	66.63.32.0/19	12603	0.25%	16559 -- REALCONNECT-01 - RealConnect, Inc
21	221.135.105.0/24	12200	0.24%	9583 -- SIFY-AS-IN Sify Limited
22	210.214.156.0/24	11926	0.23%	9583 -- SIFY-AS-IN Sify Limited
23	196.27.104.0/21	11521	0.23%	30969 -- TAN-NET TransAfrica Networks
24	196.27.108.0/22	11468	0.23%	30969 -- TAN-NET TransAfrica Networks
25	41.204.2.0/24	11066	0.22%	32398 -- REALNET-ASN-1
27	221.135.80.0/24	10192	0.20%	9583 -- SIFY-AS-IN Sify Limited
28	198.92.192.0/21	10076	0.20%	16559 -- REALCONNECT-01 - RealConnect, Inc
29	192.12.120.0/24	9938	0.20%	5691 -- MITRE-AS-5 - The MITRE Corporation

Aggregation Potential (source: bgp.potaroo.net/as2.0/)



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 routers 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

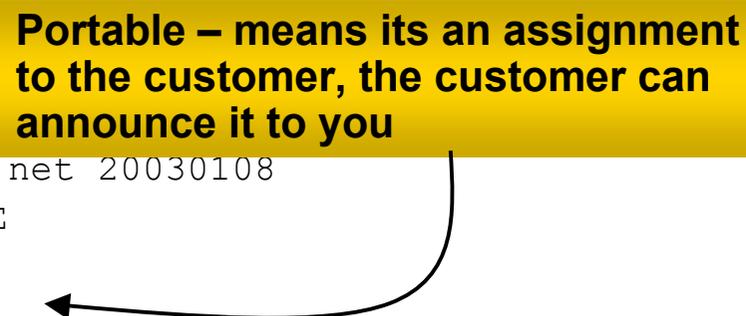
- ISPs should only accept prefixes which have been assigned or allocated to their downstream customer
- If ISP has assigned address space to its customer, then the customer IS entitled to announce it back to his ISP
- If the ISP has NOT assigned address space to its customer, then:
 - Check the five RIR databases to see if this address space really has been assigned to the customer
 - The tool: **whois**

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

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 later
- 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
 - <ftp://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 Team Cymru's bogon pages
 - <http://www.team-cymru.org/Services/Bogons/>
 - <http://www.team-cymru.org/Services/Bogons/routeserver.html> – bogon route server

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

Before we begin...

Preparing the Network

- We will deploy BGP across the network before we try and multihome
- BGP will be used therefore an ASN is required
- If multihoming to different ISPs, public ASN needed:
 - 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

Initial Assumptions

- 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

First Step: IGP

- Decide on an IGP: OSPF or ISIS 😊
- Assign loopback interfaces and /32 address 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 110m 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

WAN point to point links go in IGP

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

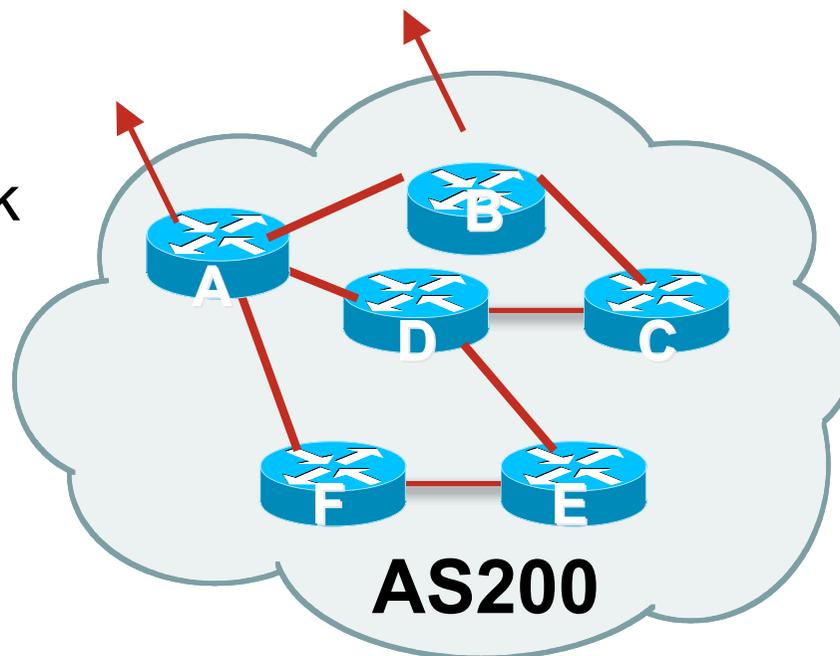
Summarise on area/level boundaries (if possible) – i.e. think about your IGP address plan

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 Second Step: iBGP

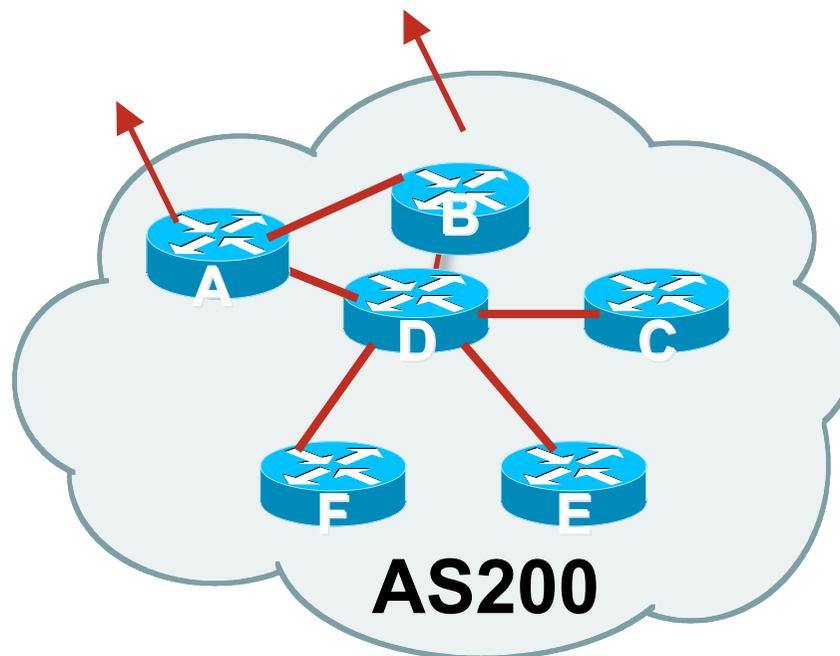
- Second step is to configure the local network to use iBGP
- iBGP can run on
 - all routers, or
 - a subset of routers, or
 - just on the upstream edge
- *iBGP must run on all routers which are in the transit path between external connections*



Preparing the Network

Second Step: 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 and peer-groups 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 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 groups and 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 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 by the 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

Example Two

- The network is not running any BGP at the moment
single statically routed connection to upstream ISP
- The network is running an IGP though
All internal routing information is in the IGP
By IGP, OSPF or ISIS is assumed

Preparing the Network IGP

- If not already done, assign loopback interfaces and /32 addresses to each router which is running the IGP
 - Loopback is used for OSPF and BGP router id anchor
 - Used for iBGP and route origination
- Ensure that the loopback /32s are appearing in the IGP

Preparing the Network iBGP

- Go through the iBGP decision process as in Example One
- Decide full or partial, and the extent of the iBGP reach in the network

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: Reduce BGP distance to be less than the IGP
(so that iBGP routes take priority)

- Step 4: Carefully remove the “customer” prefixes from the IGP

- Check!** Does the network still work?

- Step 5: Restore BGP distance to less than IGP

- Step 6: 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 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 “customer” routes from IGP?

- Work on one router at a time:
 - Check that IGP route for a particular destination is also learned by iBGP
 - If so, remove it from the IGP
 - 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 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 passwords, tricks and 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!

iBGP: Next-hop-self

- BGP speaker announces external network to iBGP peers using router's local address (loopback) as next-hop
- 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
 - Many ISPs consider this “best practice”

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          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

```
*> 194.146.180.0/22      2497 3257 29686 16327 16327
16327 16327 16327 16327 16327 16327 16327 16327
16327 16327 16327 16327 16327 16327 16327 16327
16327 16327 16327 i
```

This example shows 20 prepends (for no obvious reason)

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

BGP TTL “hack”

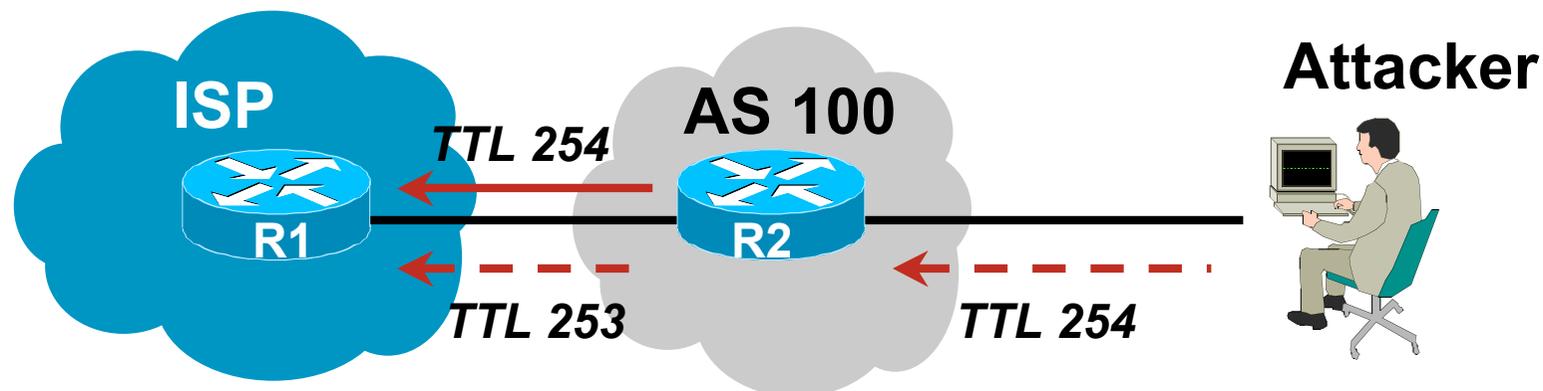
- Implement RFC5082 on BGP peerings

(Generalised TTL Security Mechanism)

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

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 Team Cymru's BGP templates
 - <http://www.team-cymru.org/ReadingRoom/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!

iBGP Template

Example continued

- Use passwords on iBGP session
 - Not being paranoid, **VERY** necessary
 - 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"

eBGP Template Example

- BGP damping
 - Do **NOT** use it unless you understand the impact
 - Do **NOT** 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
- Limit maximum as-path length inbound
- 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!!

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 Techniques for Internet Service Providers

Philip Smith <pfs@cisco.com>

APRICOT 2009

18th-27th February 2009

Manila, Philippines