BGP Techniques for Internet Service Providers

Philip Smith <pfs@cisco.com>

NANOG 44

12-14 October 2008

Los Angeles

Presentation Slides

- Will be available on

    ftp://ftp-eng.cisco.com

    /pfs/seminars/NANOG44-BGP-Techniques.pdf

    And on the NANOG44 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

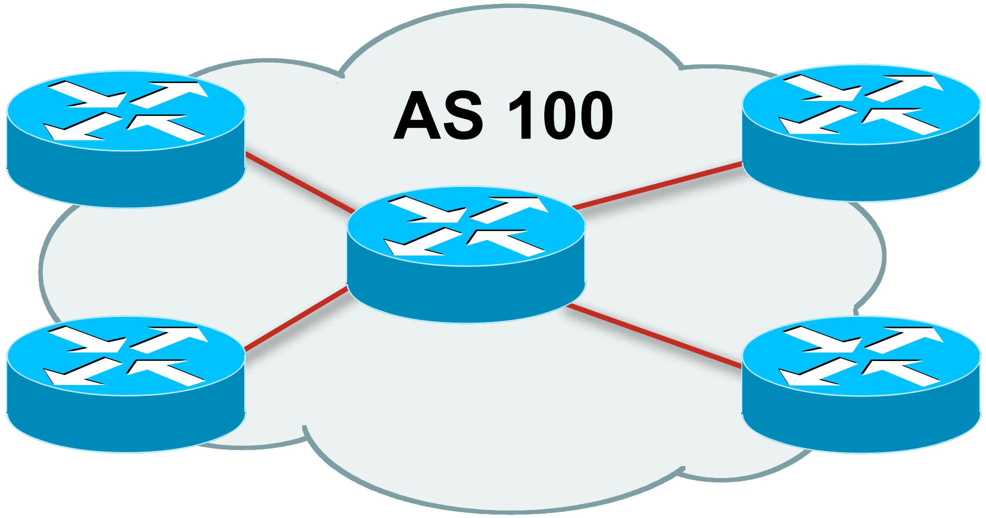
BGP Basics

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 number (ASN)

Autonomous System Number (ASN)

- An ASN is a 32 bit integer
- Two ranges

0-65535	(original 16-bit range)
65536-4294967295	(32-bit range - RFC4893)

- Usage:

1-64511	(public Internet)
64512-65534	(private use only)
23456	(represent 32-bit range in 16-bit world)
0 and 65535	(reserved)
65536-4294967295	(public Internet)

- 32-bit range representation in IETF last call

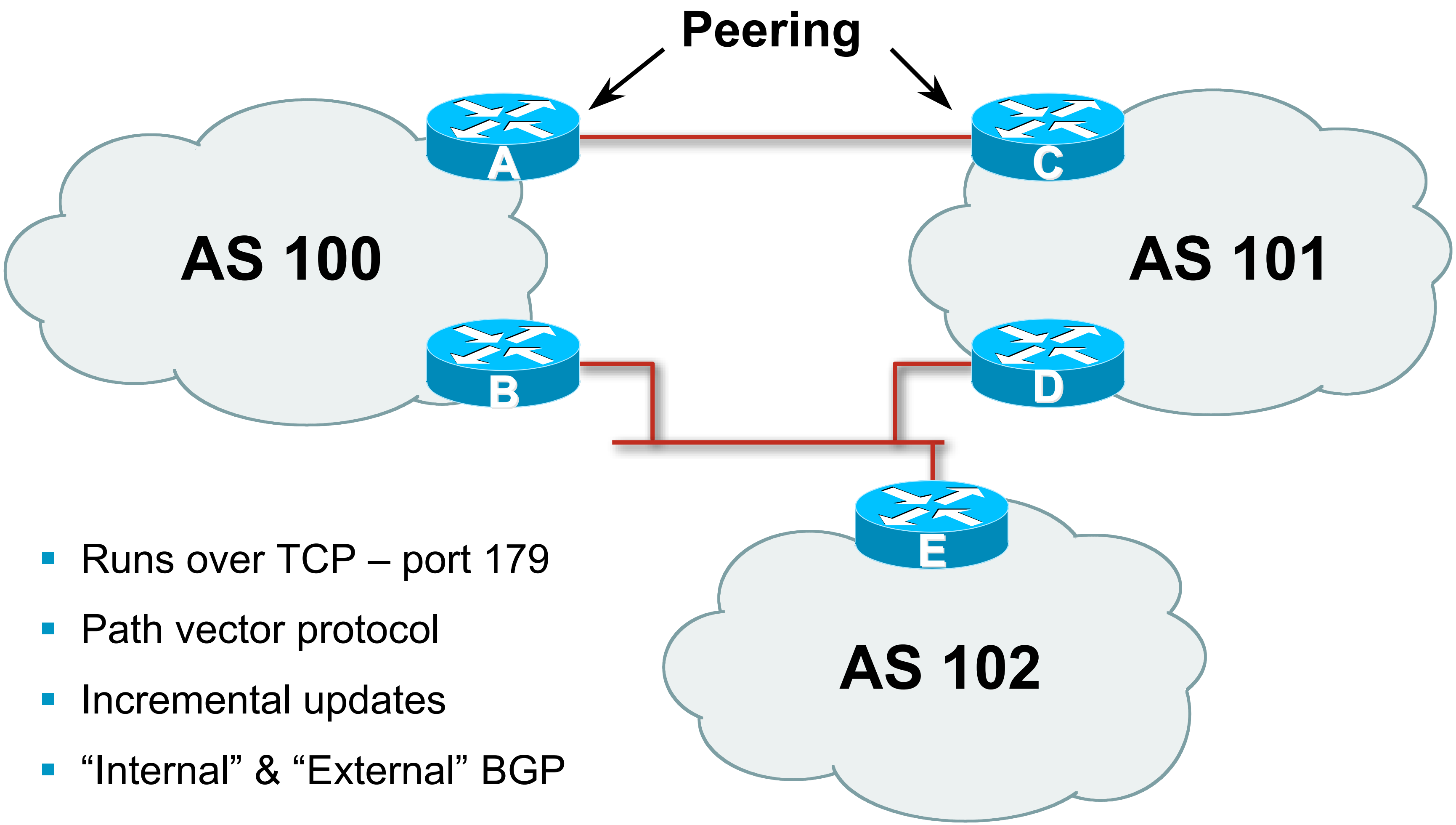
  draft-ietf-idr-as-representation-01.txt

  Defines “asplain” (traditional format) as standard notation

Autonomous System Number (ASN)

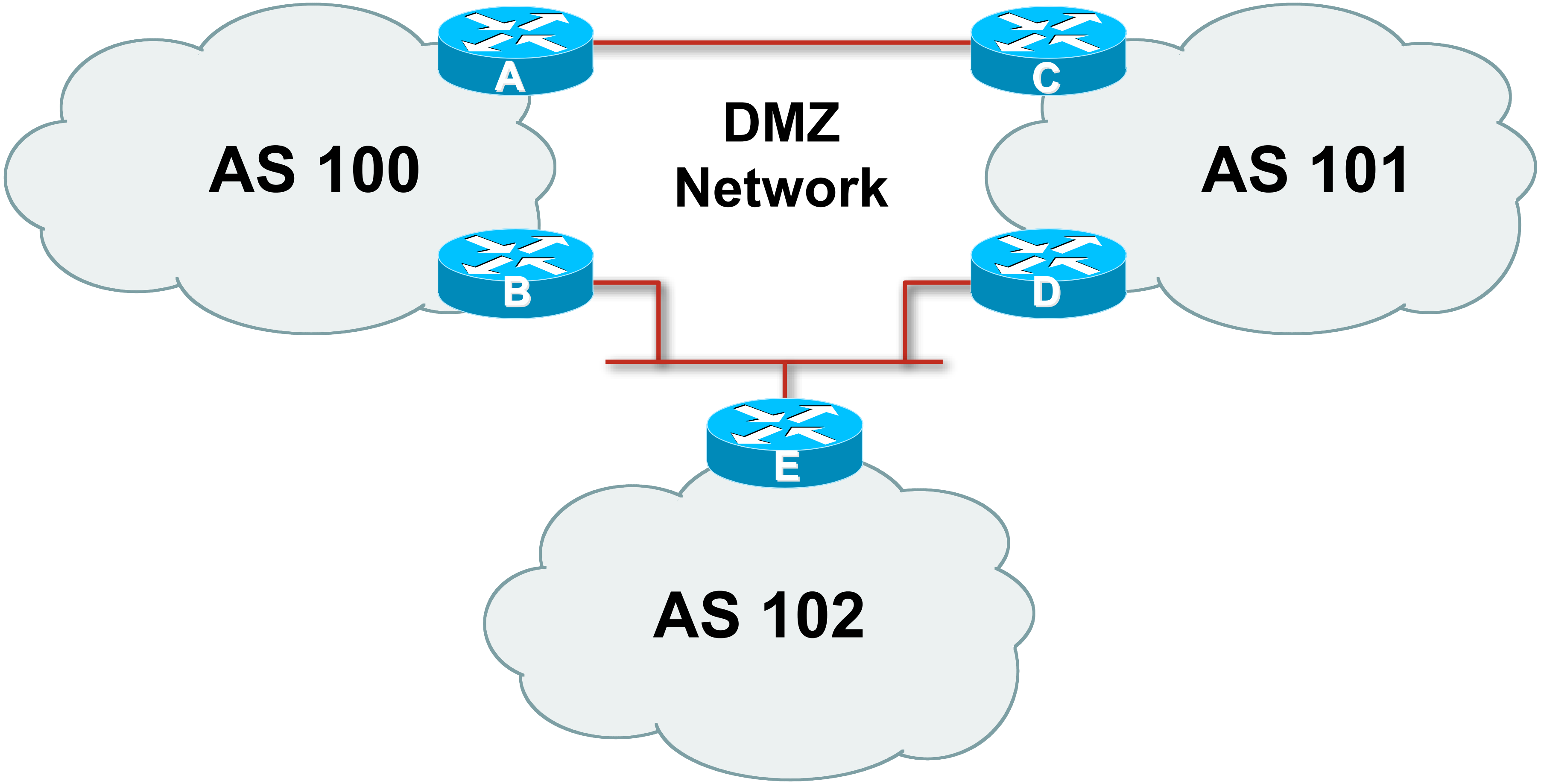
- ASNs are distributed by the Regional Internet Registries
  - They are also available from upstream ISPs who are members of one of the RIRs
- Current 16-bit ASN allocations up to 49151 have been made to the RIRs
  - Around 29400 are visible on the Internet
- The RIRs also have received 1024 32-bit ASNs each
  - 12 are visible on the Internet (early adopters)
- See **www.iana.org/assignments/as-numbers**

BGP Basics



- Runs over TCP – port 179
- Path vector protocol
- Incremental updates
- “Internal” & “External” BGP

Demarcation Zone (DMZ)



- Shared network between ASes

BGP General Operation

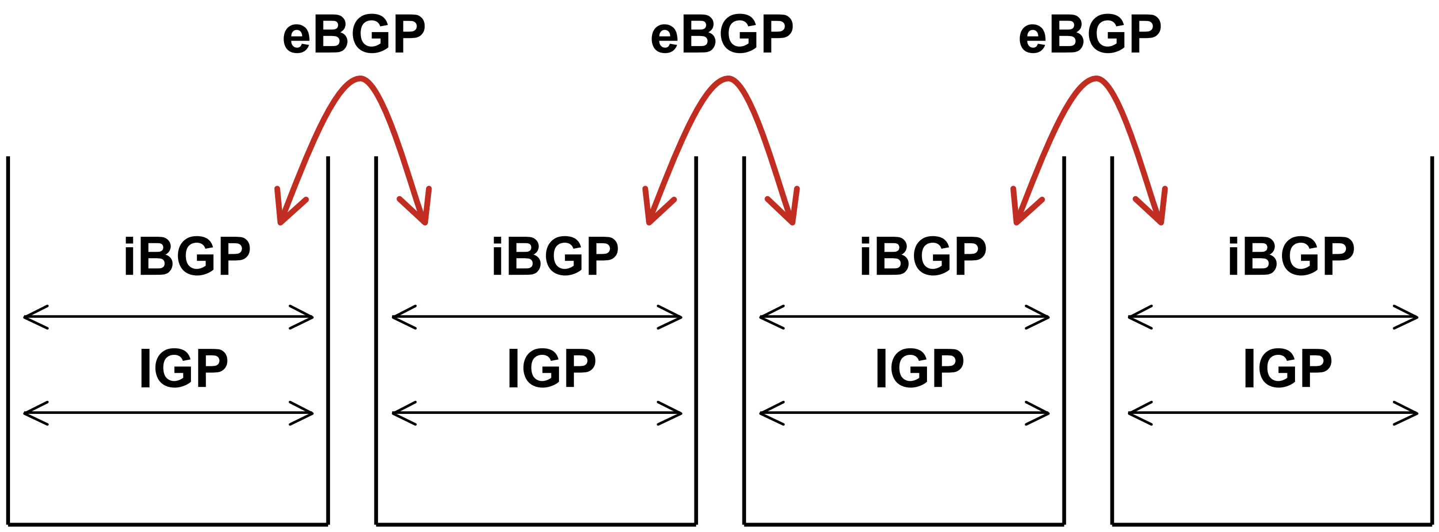
- Learns multiple paths via internal and external BGP speakers
- Picks the best path and installs in the forwarding table
- Best path is sent to external BGP neighbours
- Policies are applied by influencing the best path selection

eBGP & iBGP

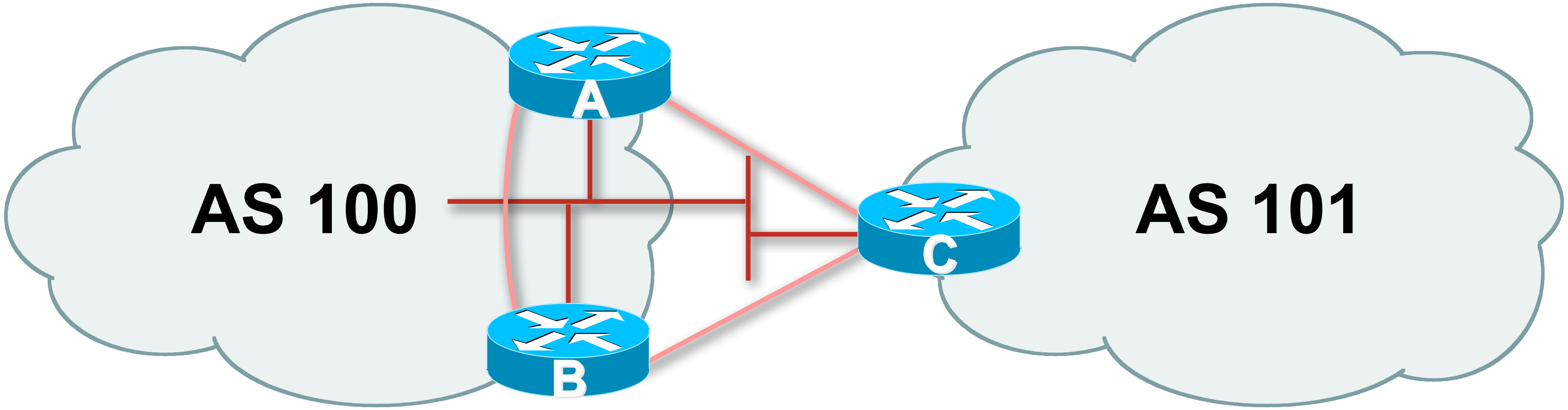
- BGP used internally (iBGP) and externally (eBGP)
- iBGP used to carry
  - some/all Internet prefixes across ISP backbone
  - ISP's customer prefixes
- eBGP used to
  - exchange prefixes with other ASes
  - implement routing policy

BGP/IGP model used in ISP networks

- Model representation



External BGP Peering (eBGP)

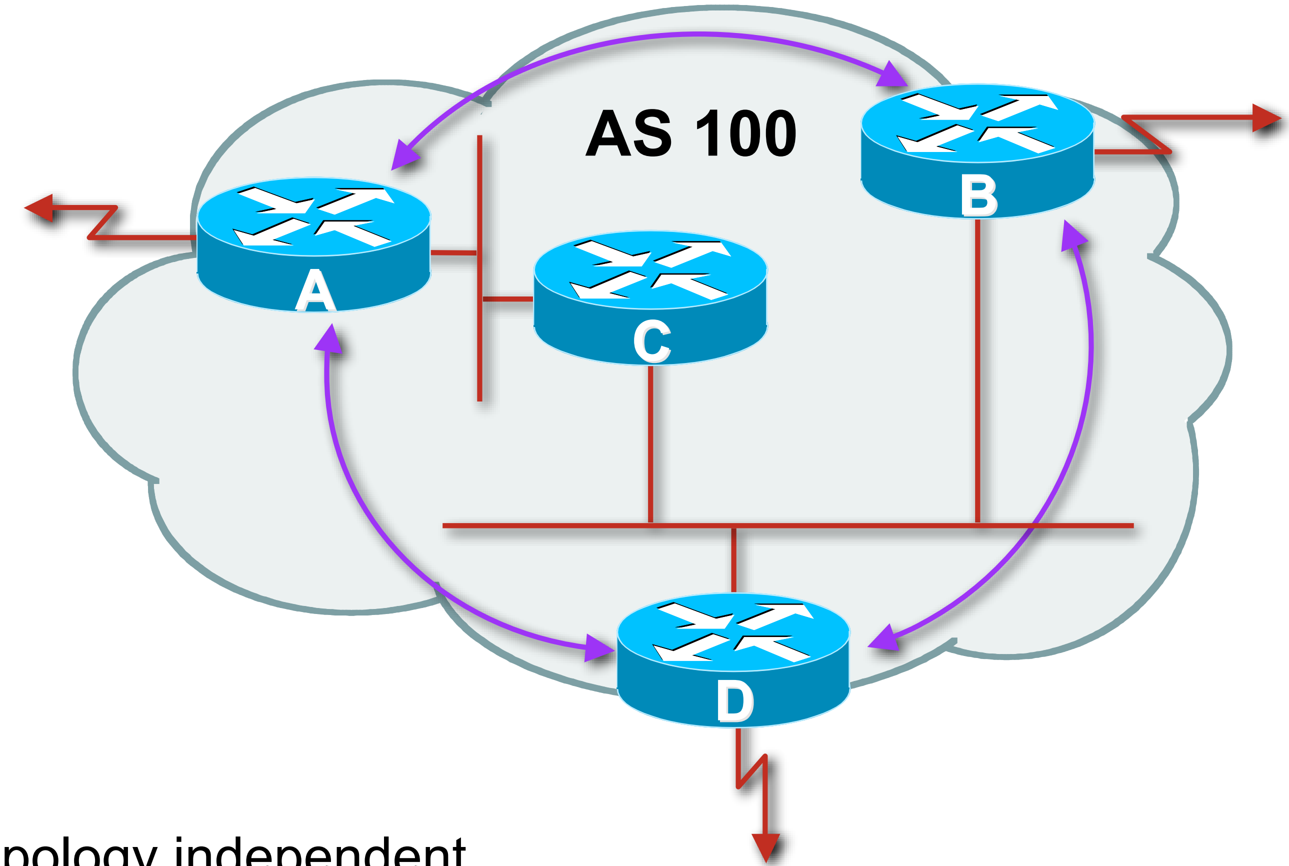


- Between BGP speakers in different AS
- Should be directly connected
- Never run an IGP between eBGP peers

Internal BGP (iBGP)

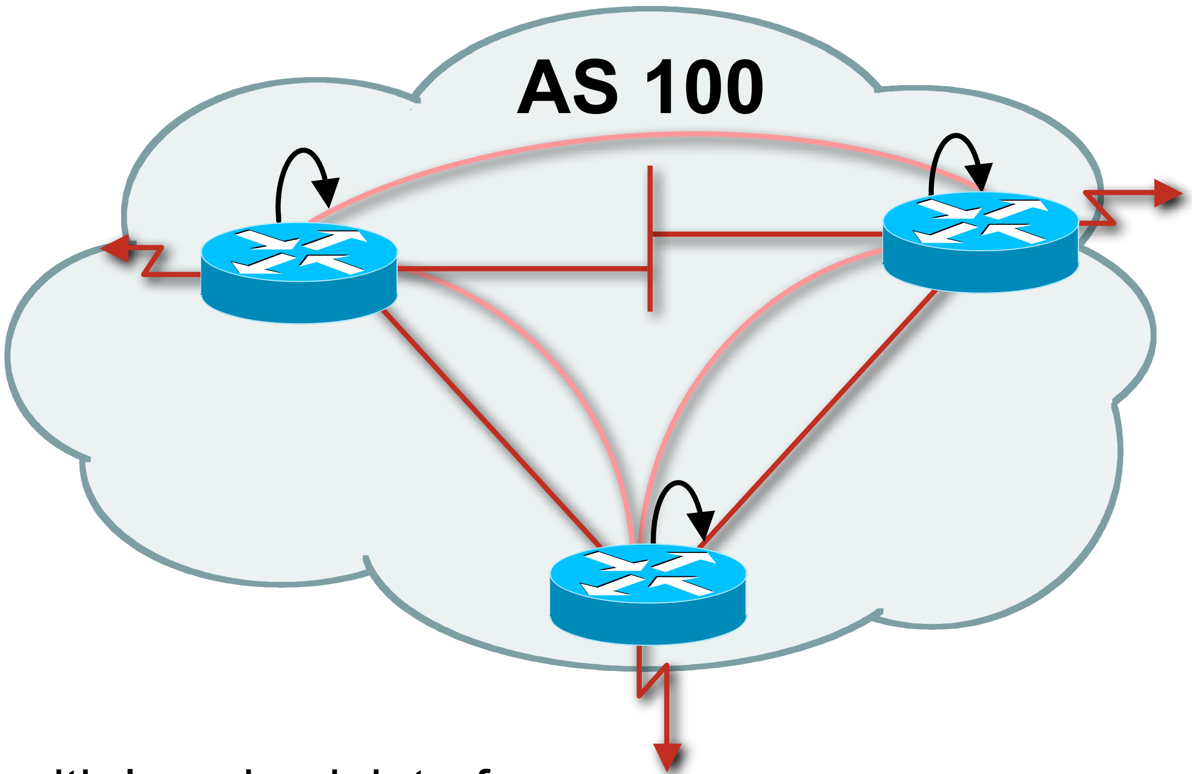
- BGP peer within the same AS
- Not required to be directly connected
  - IGP takes care of inter-BGP speaker connectivity
- iBGP speakers must to be fully meshed:
  - They originate connected networks
  - They pass on prefixes learned from outside the ASN
  - They do not pass on prefixes learned from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

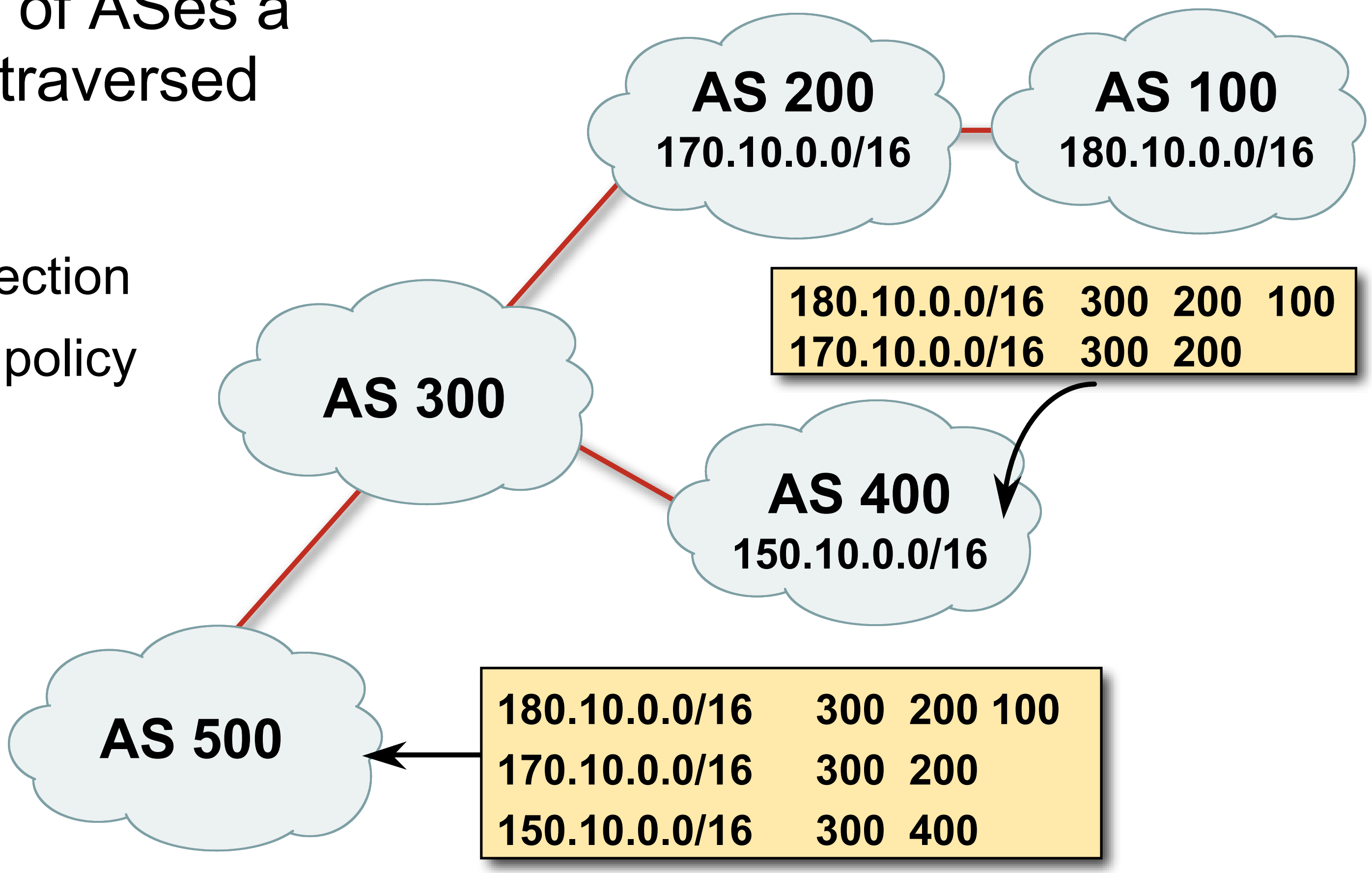
Information about BGP

AS-Path

- Sequence of ASes a route has traversed
- Used for:

  Loop detection

  Applying policy

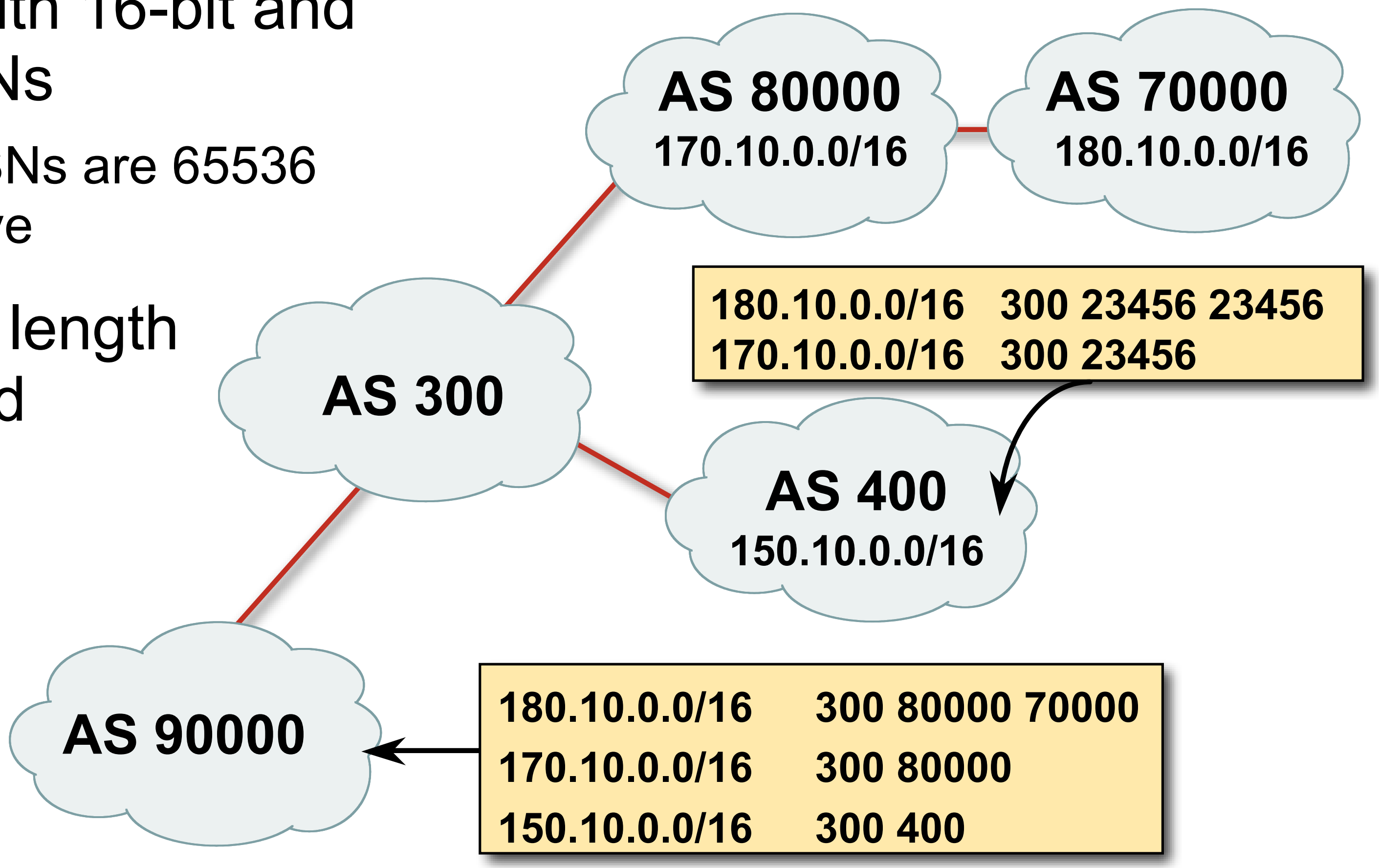


AS-Path (with 16 and 32-bit ASNs)

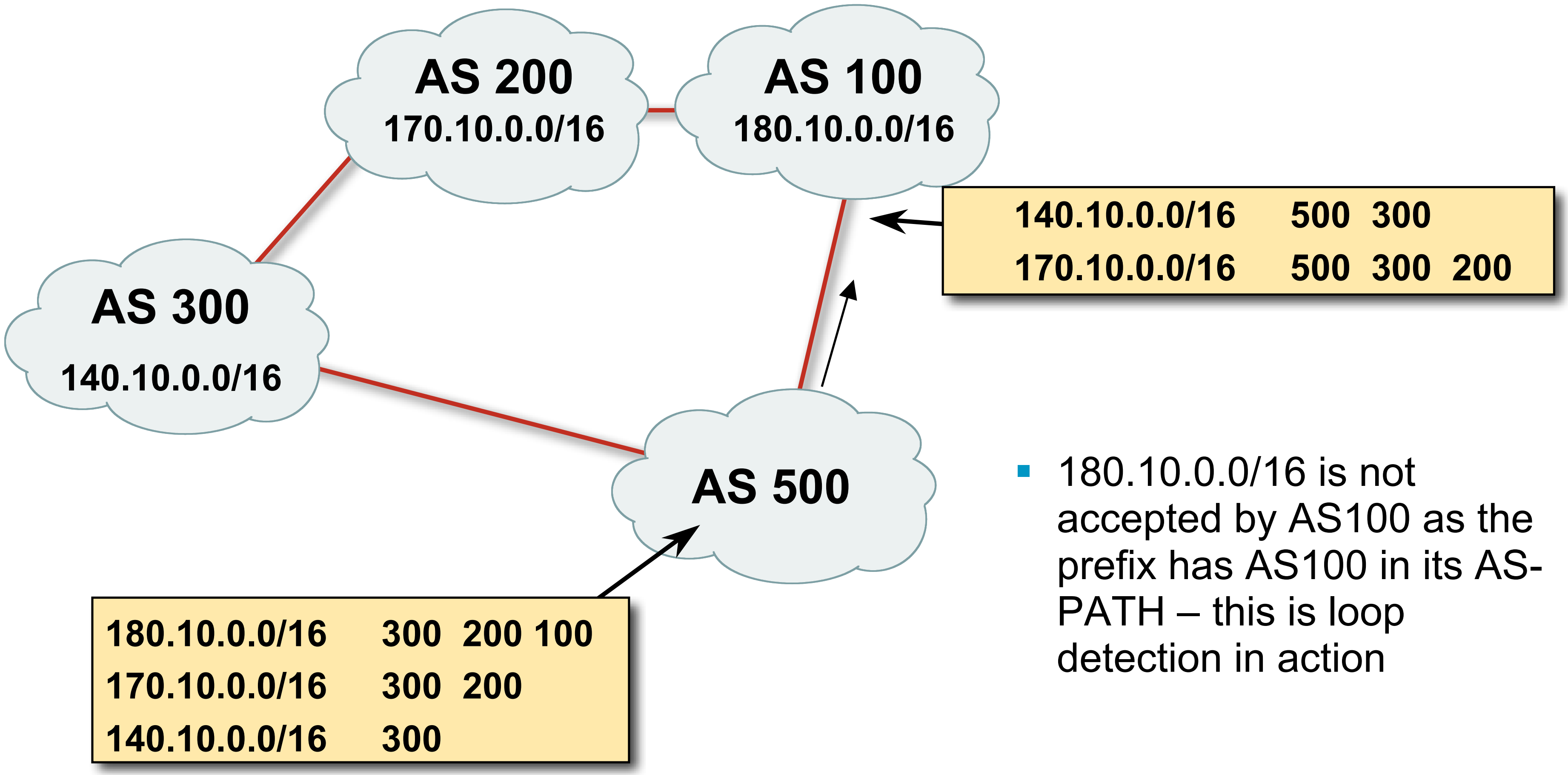
- Internet with 16-bit and 32-bit ASNs

  32-bit ASNs are 65536 and above

- AS-PATH length maintained

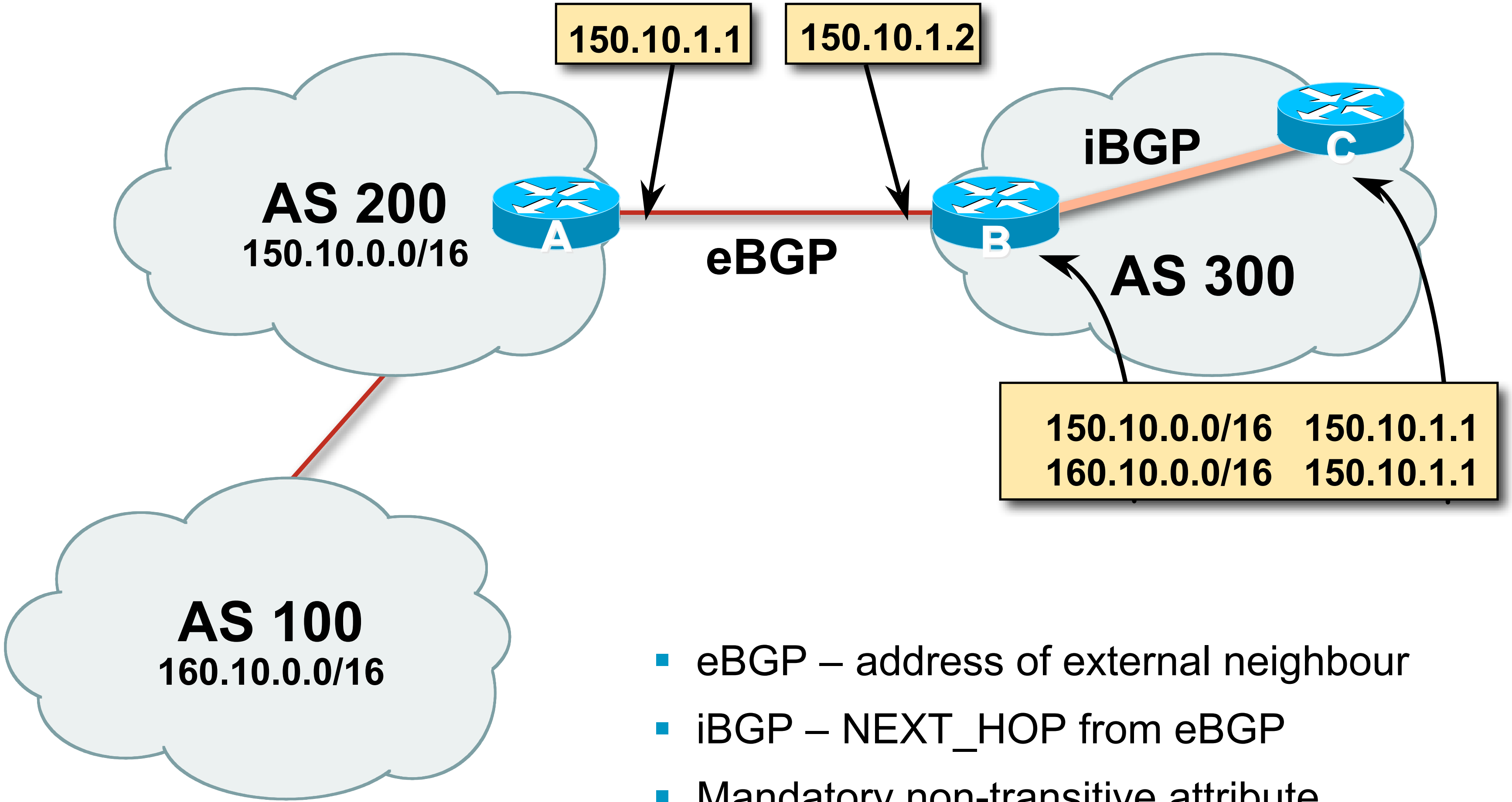


AS-Path loop detection



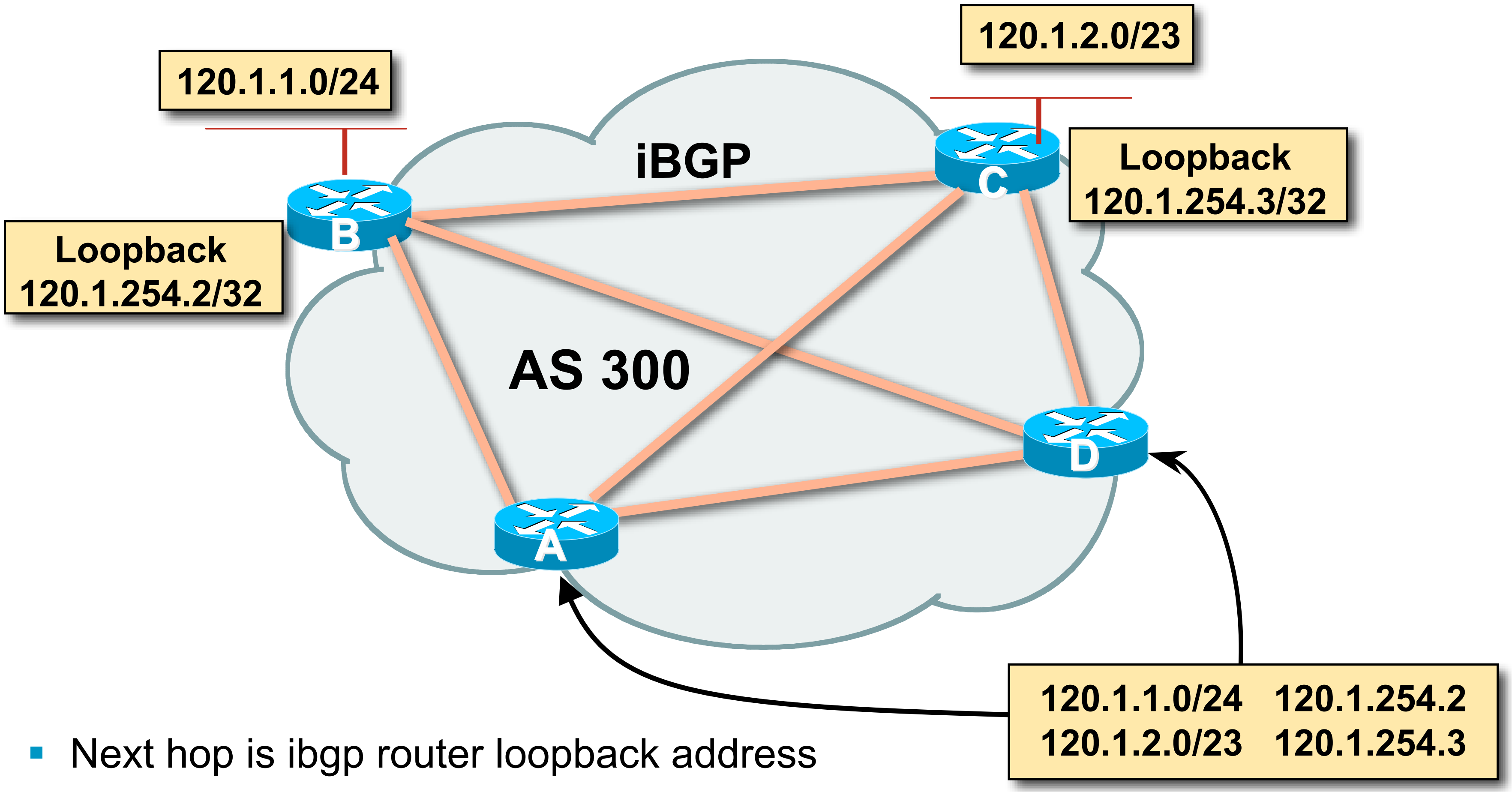
- 180.10.0.0/16 is not accepted by AS100 as the prefix has AS100 in its AS-PATH – this is loop detection in action

Next Hop



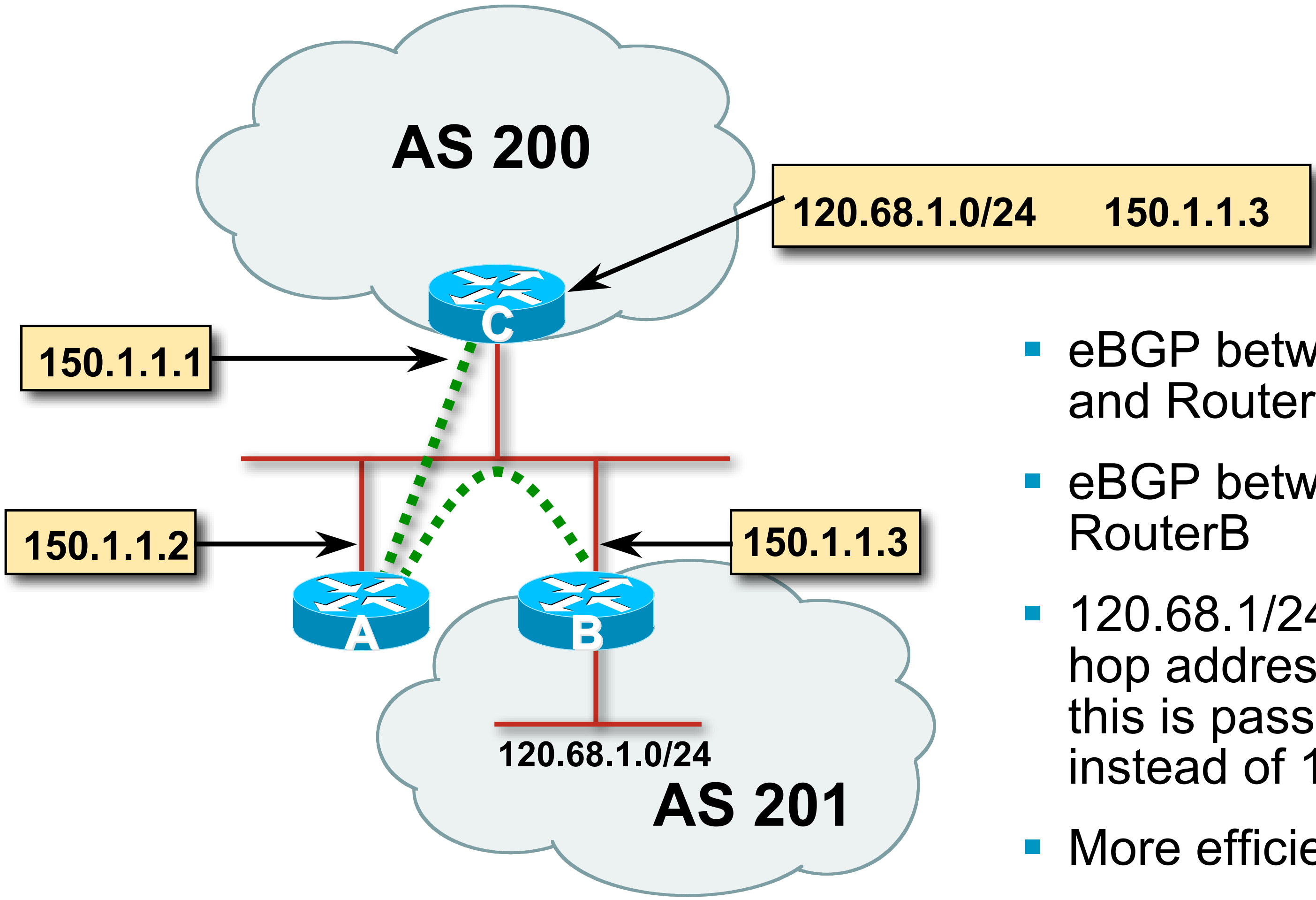
- eBGP – address of external neighbour
- iBGP – NEXT_HOP from eBGP
- Mandatory non-transitive attribute

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 RouterA and RouterB
- 120.68.1/24 prefix has next hop address of 150.1.1.3 – this is passed on to RouterC 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

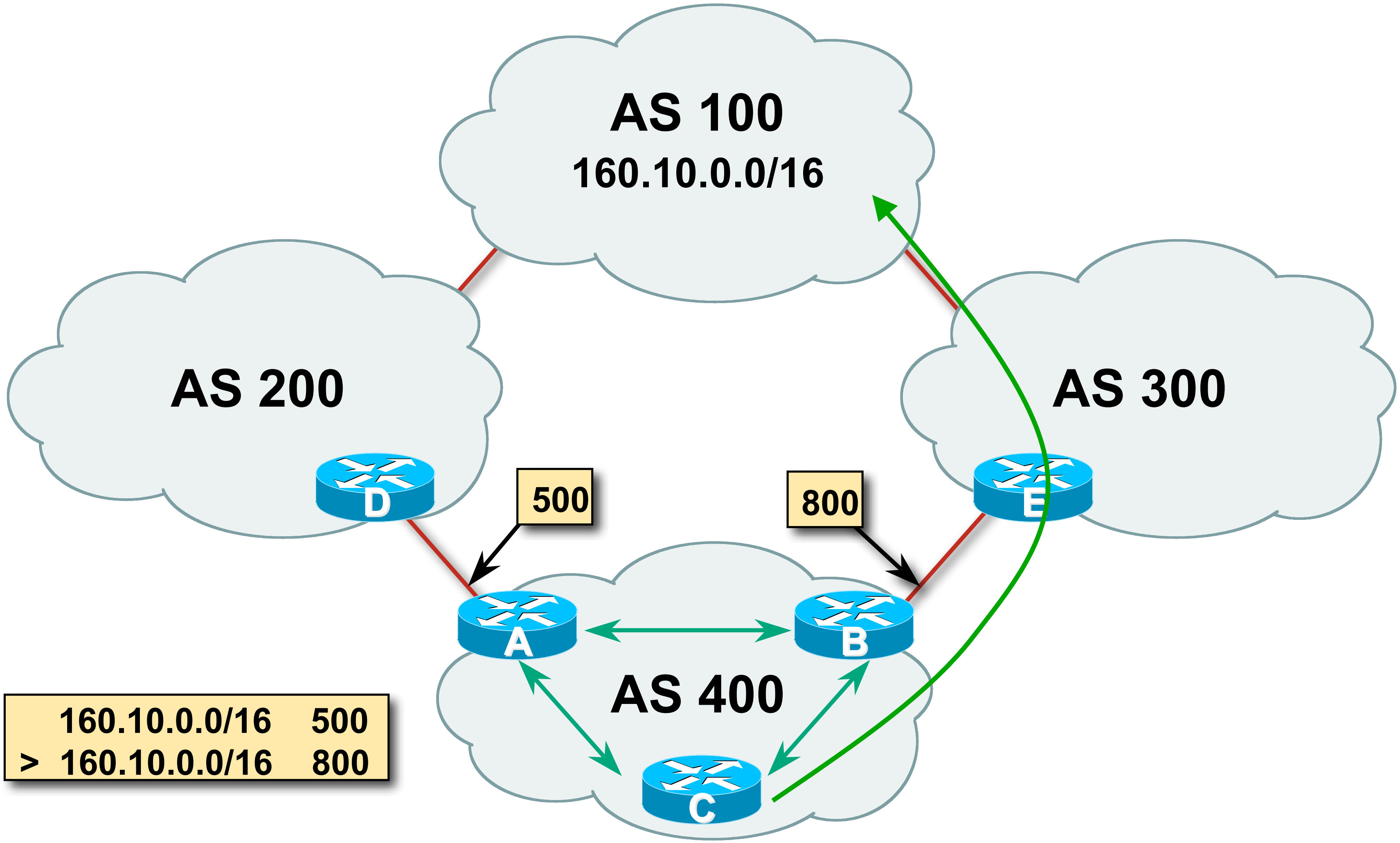
  EGP – generated by EGP

  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



	160.10.0.0/16	500
>	160.10.0.0/16	800

Local Preference

- Non-transitive and optional attribute
- Local to an AS – non-transitive

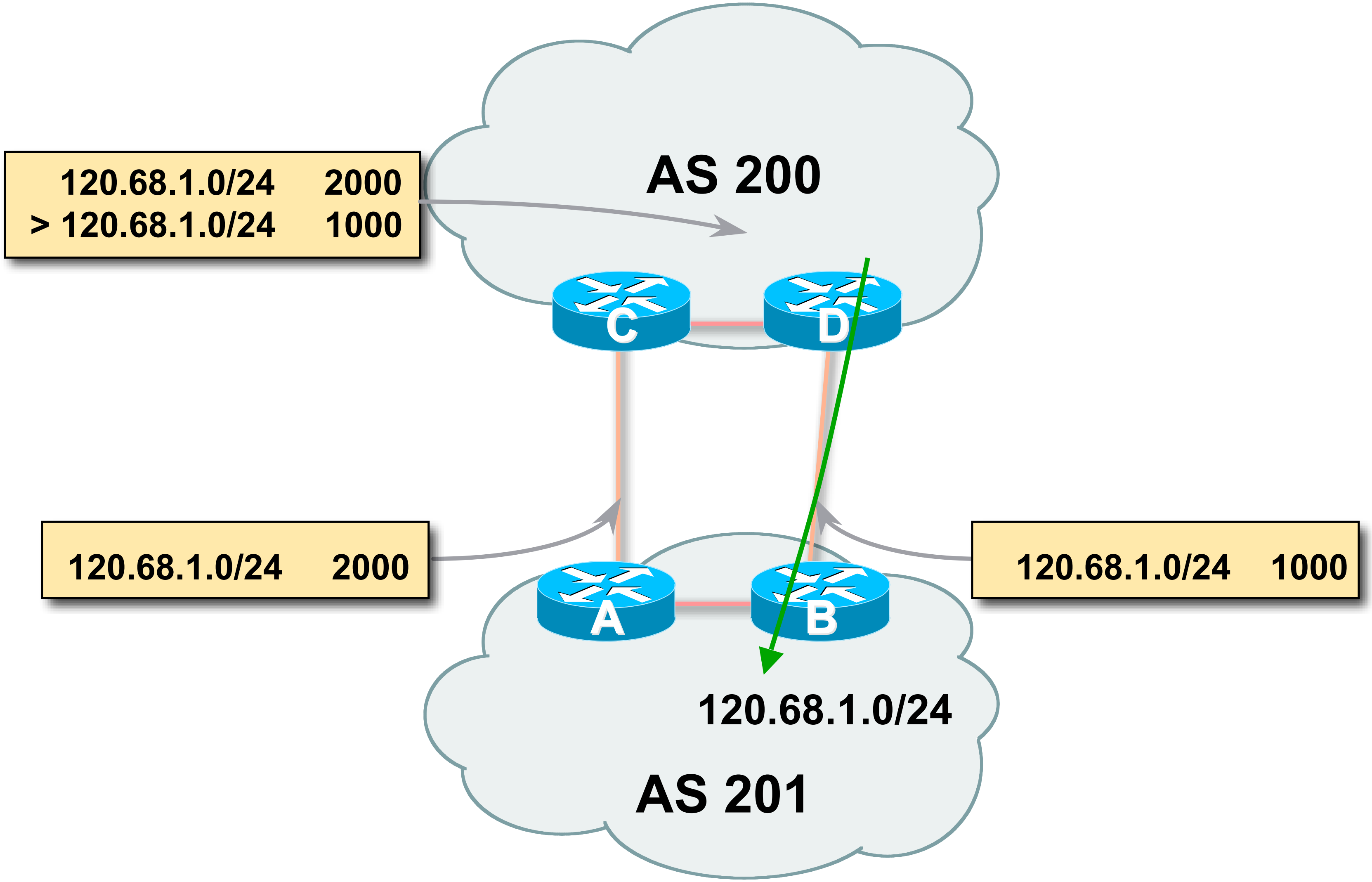
    Default local preference is 100 (IOS)

- 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)

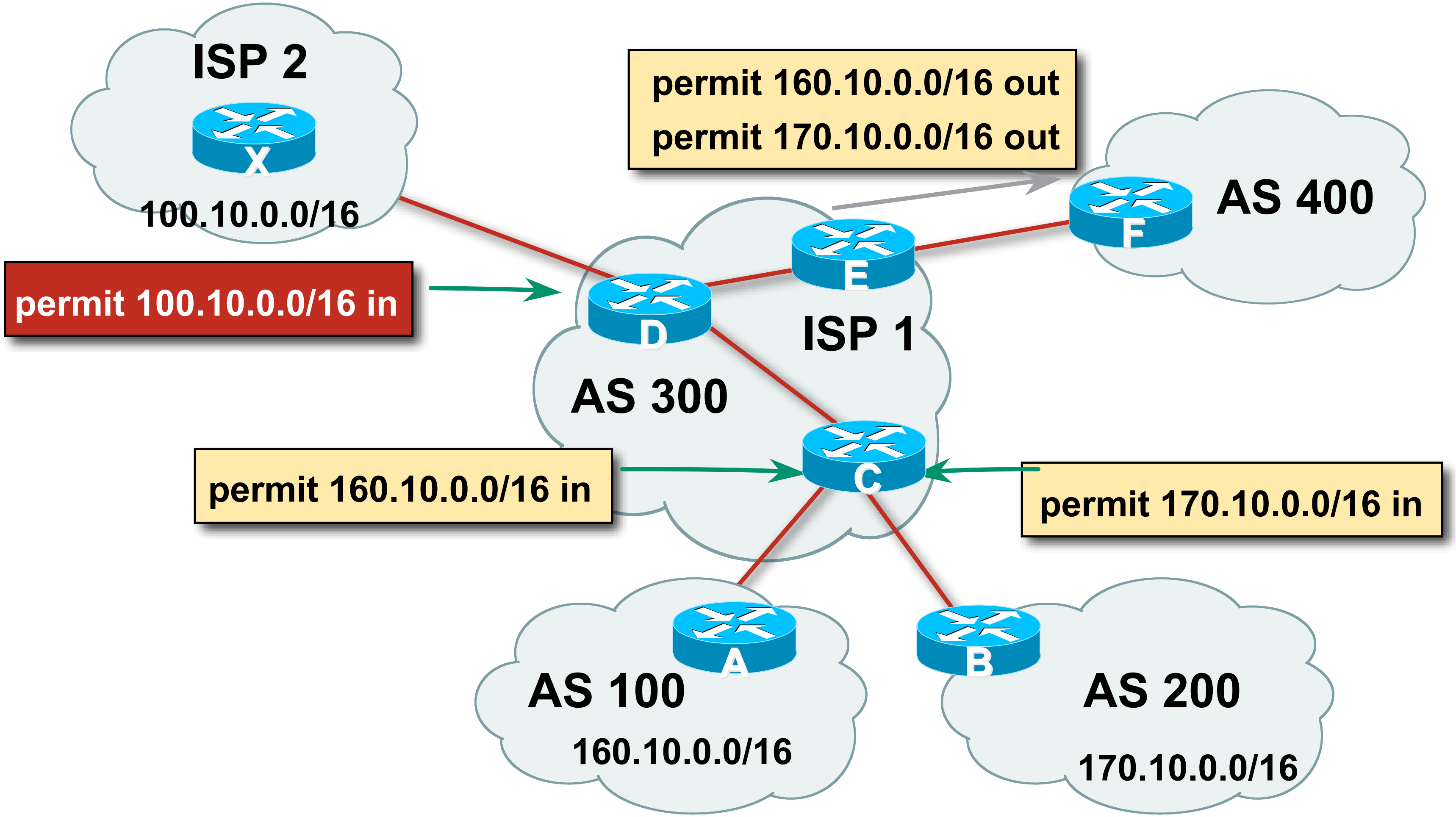
Multi-Exit Discriminator "metric confusion"

- 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"

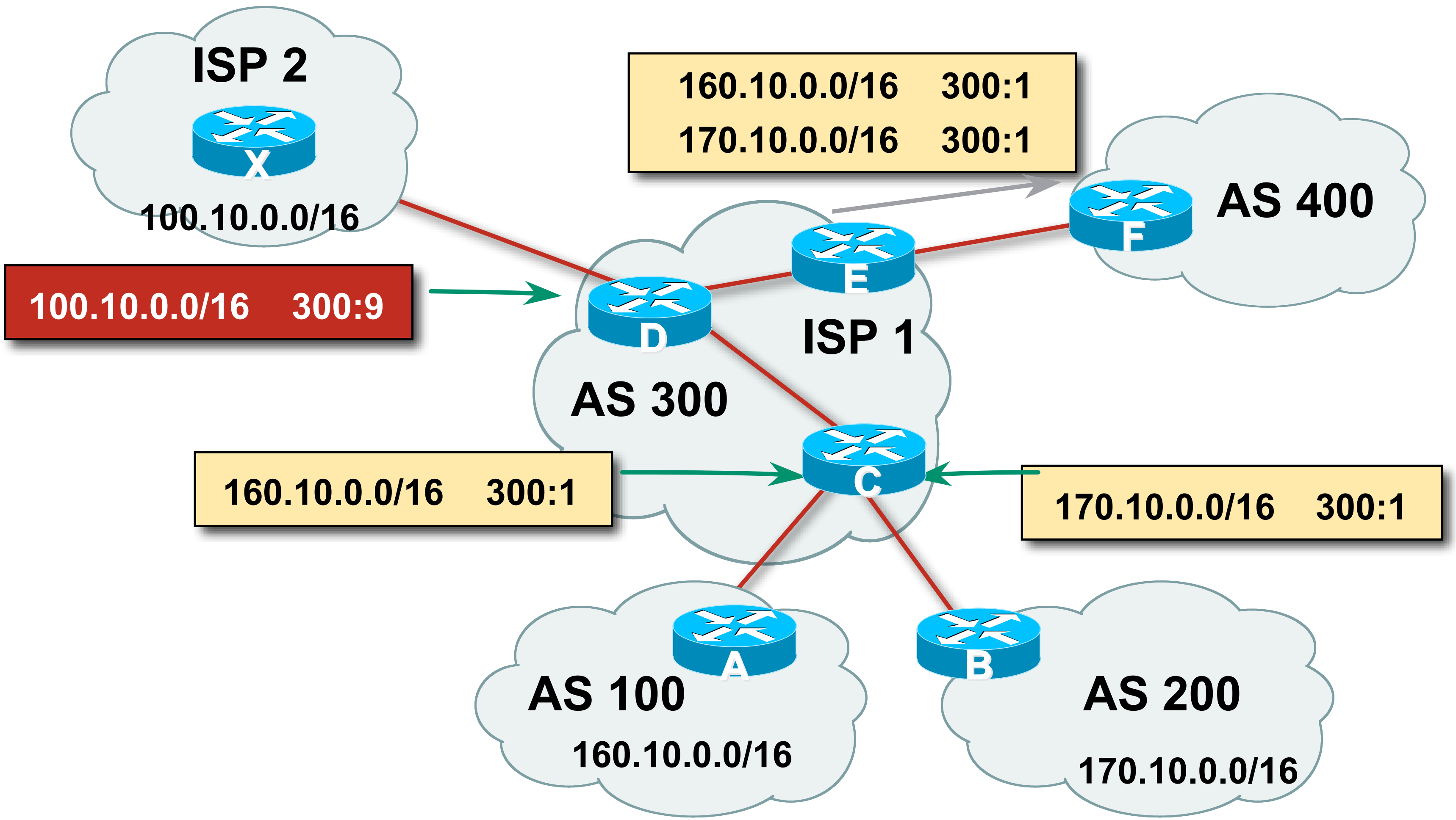
Community

- 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

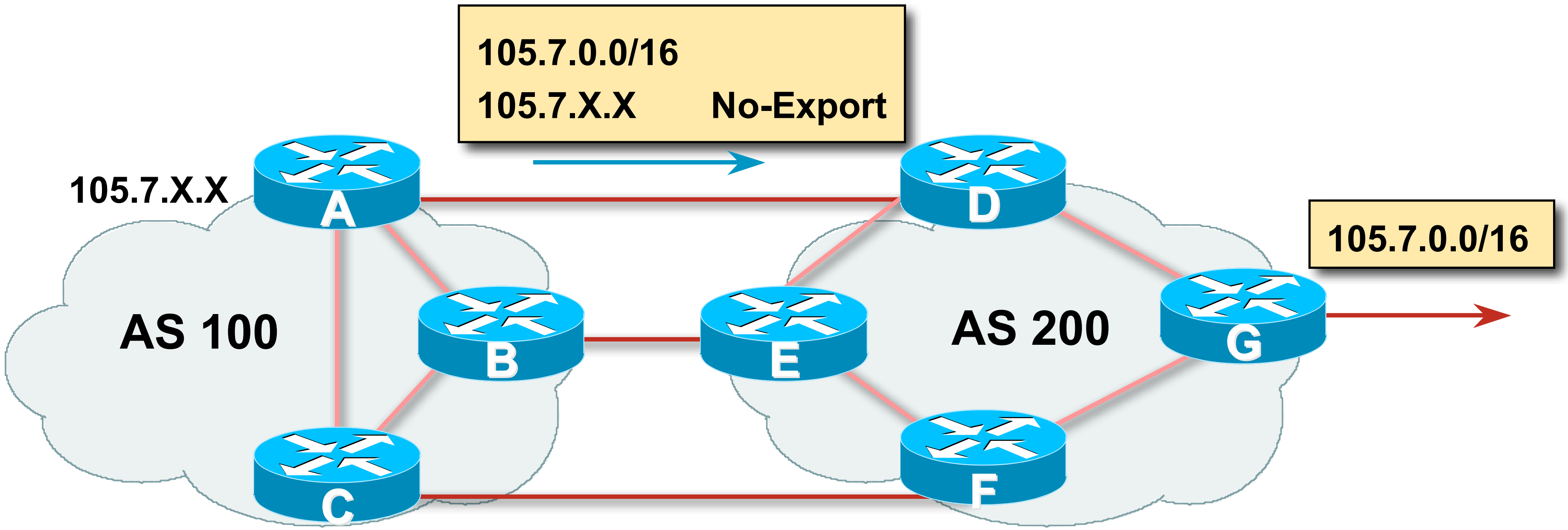
- no-export-subconfed 65535:65283

  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

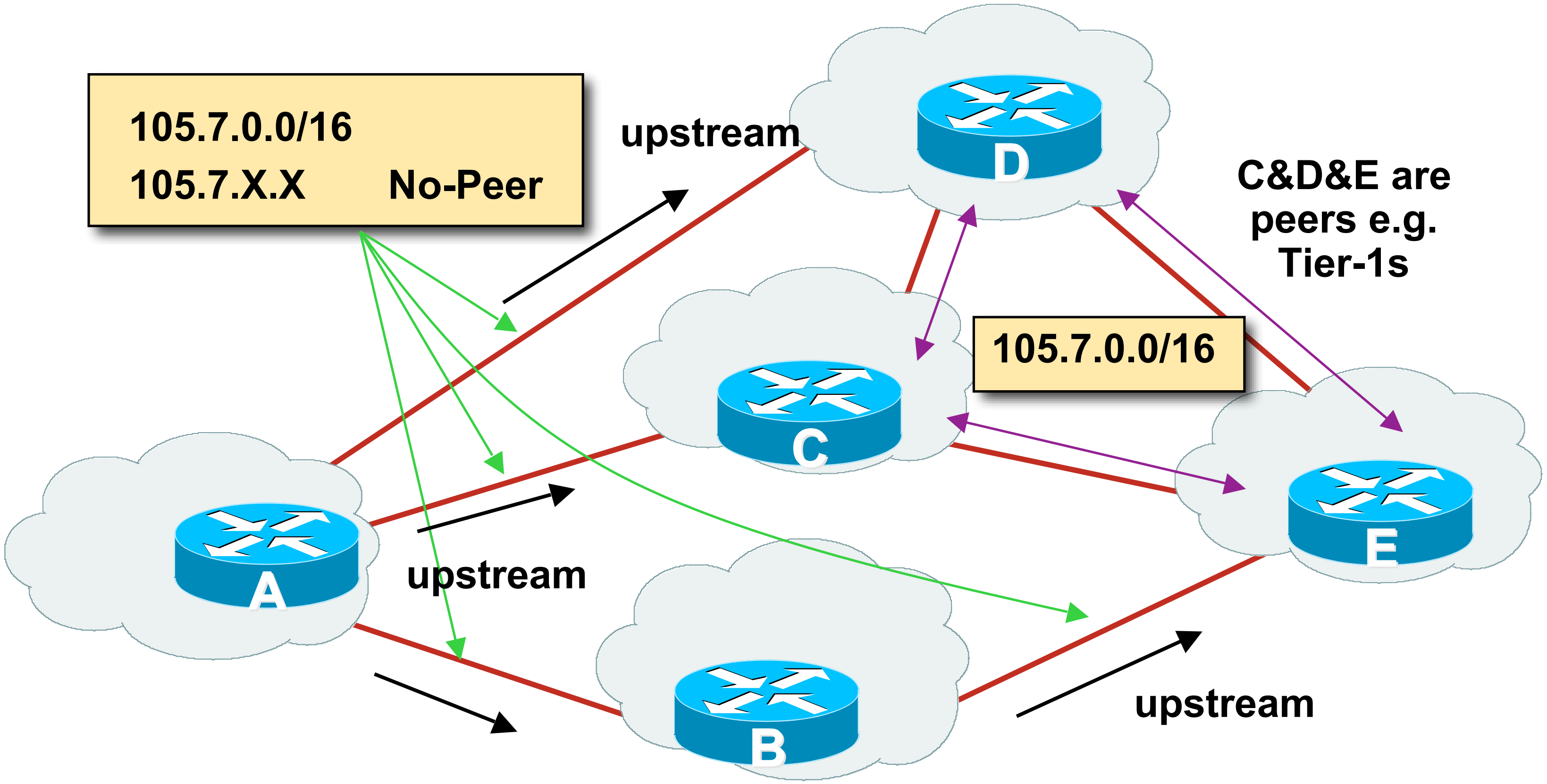


- AS100 announces aggregate and subprefixes

  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

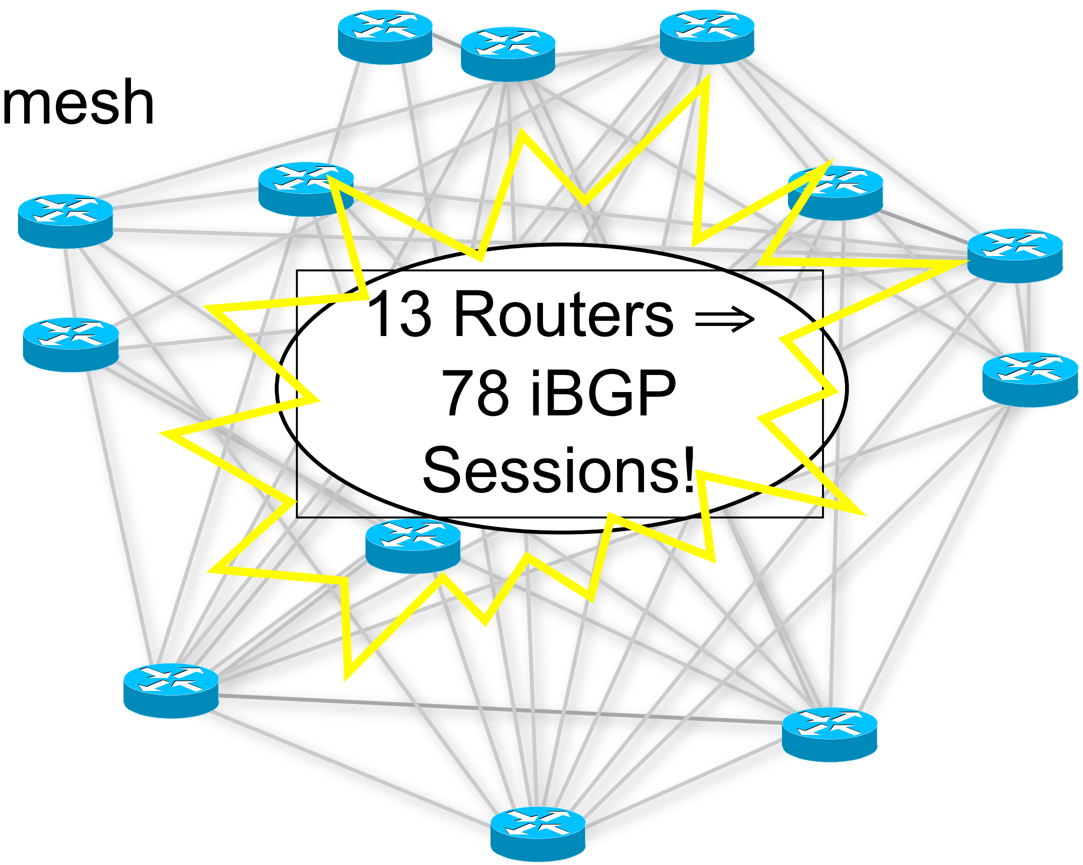
Route Reflectors

Scaling the iBGP mesh

Scaling iBGP mesh

- Avoid ½n(n-1) iBGP mesh

n=1000 ⇒ nearly half a million ibgp sessions!

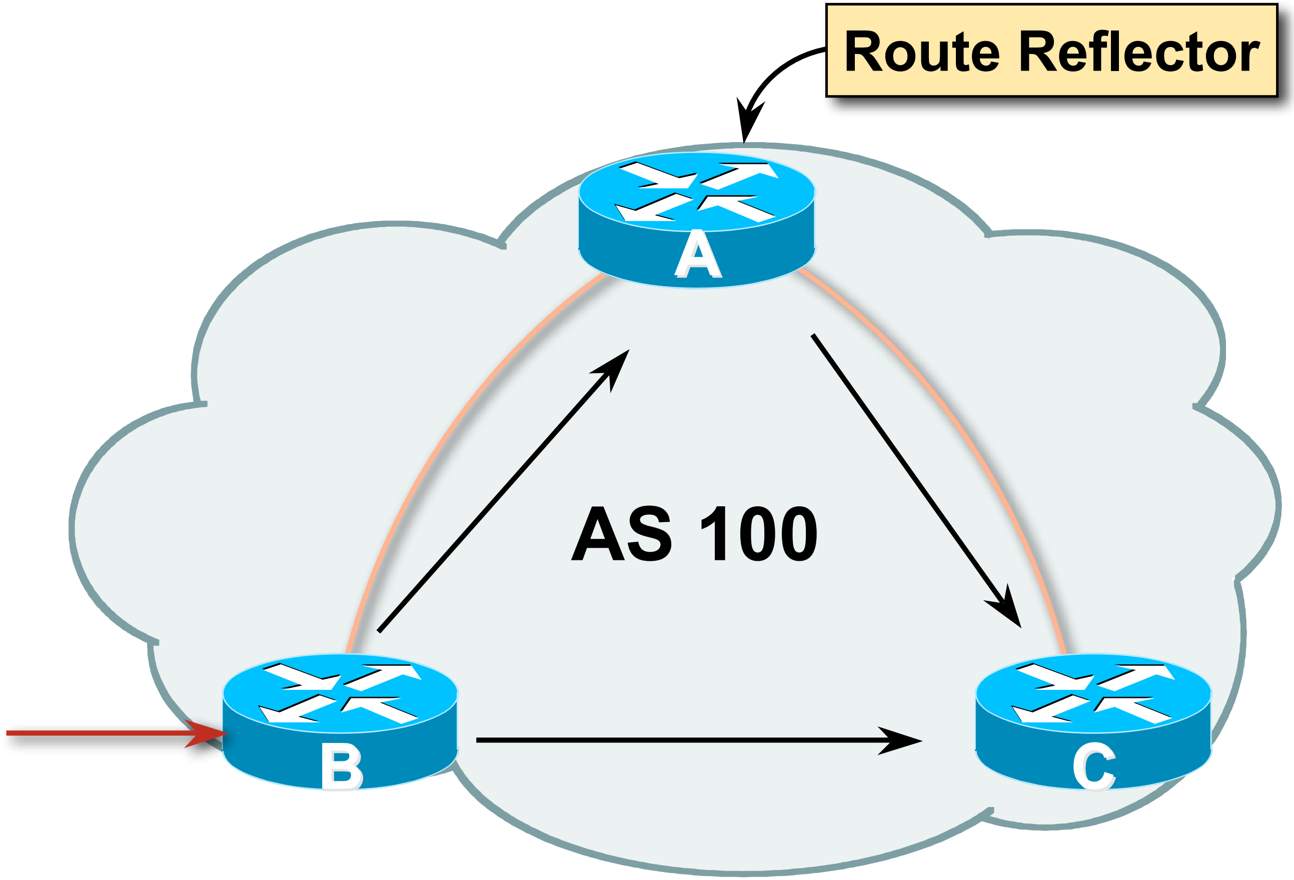


- Two solutions

  Route reflector – simpler to deploy and run

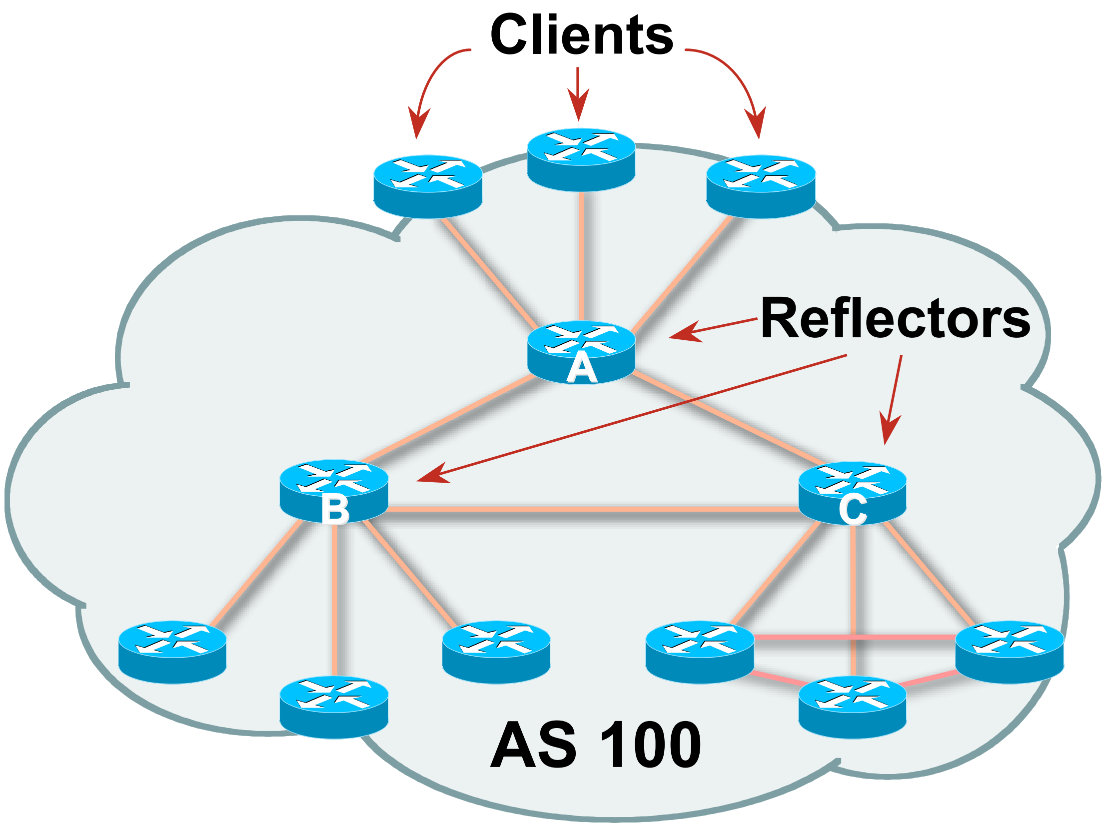
  Confederation – more complex, has corner case advantages

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!

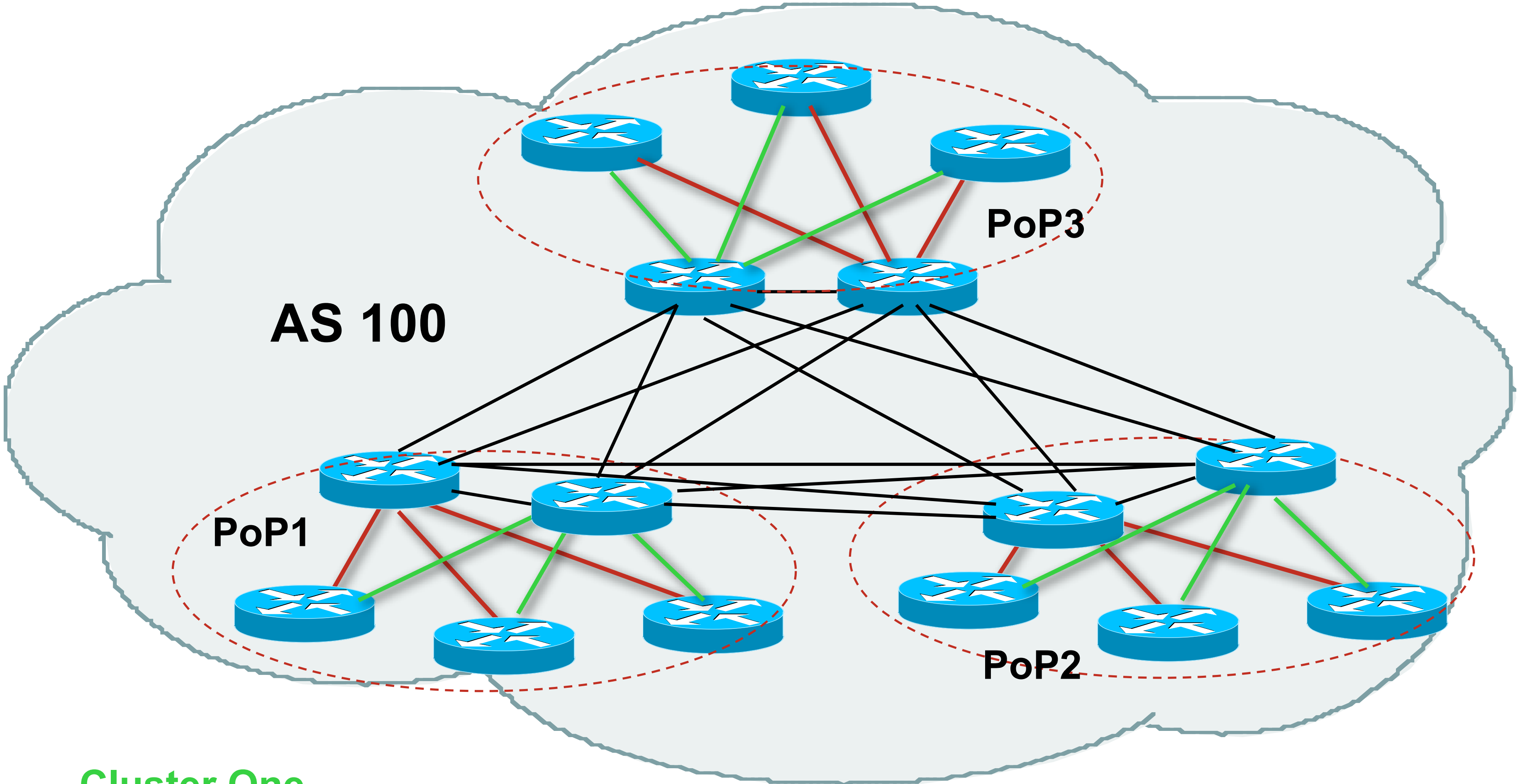
  All RRs in the cluster must have the same cluster-id (otherwise it is a different cluster)

- 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

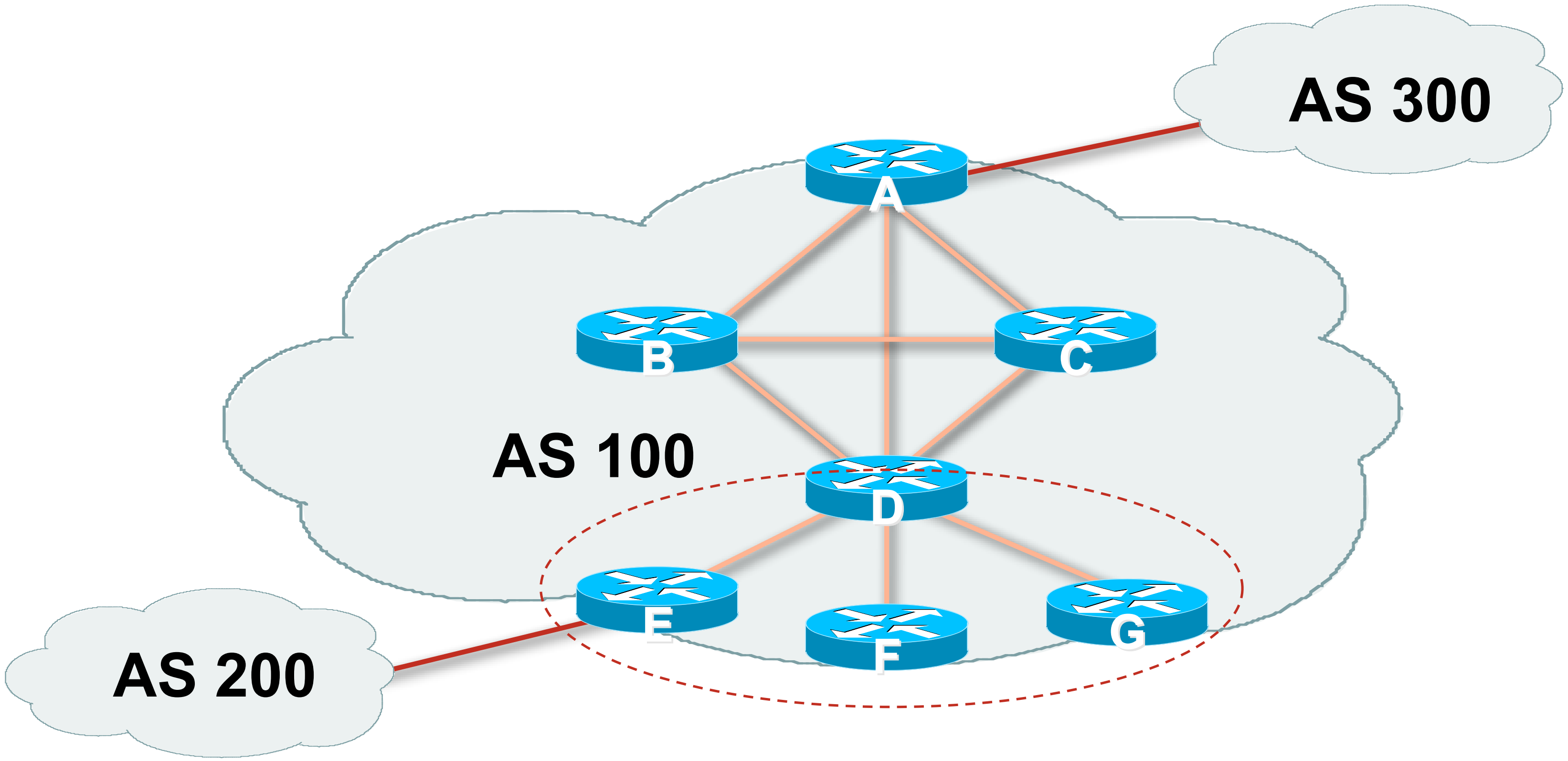
Route Reflector: Deployment

- 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"

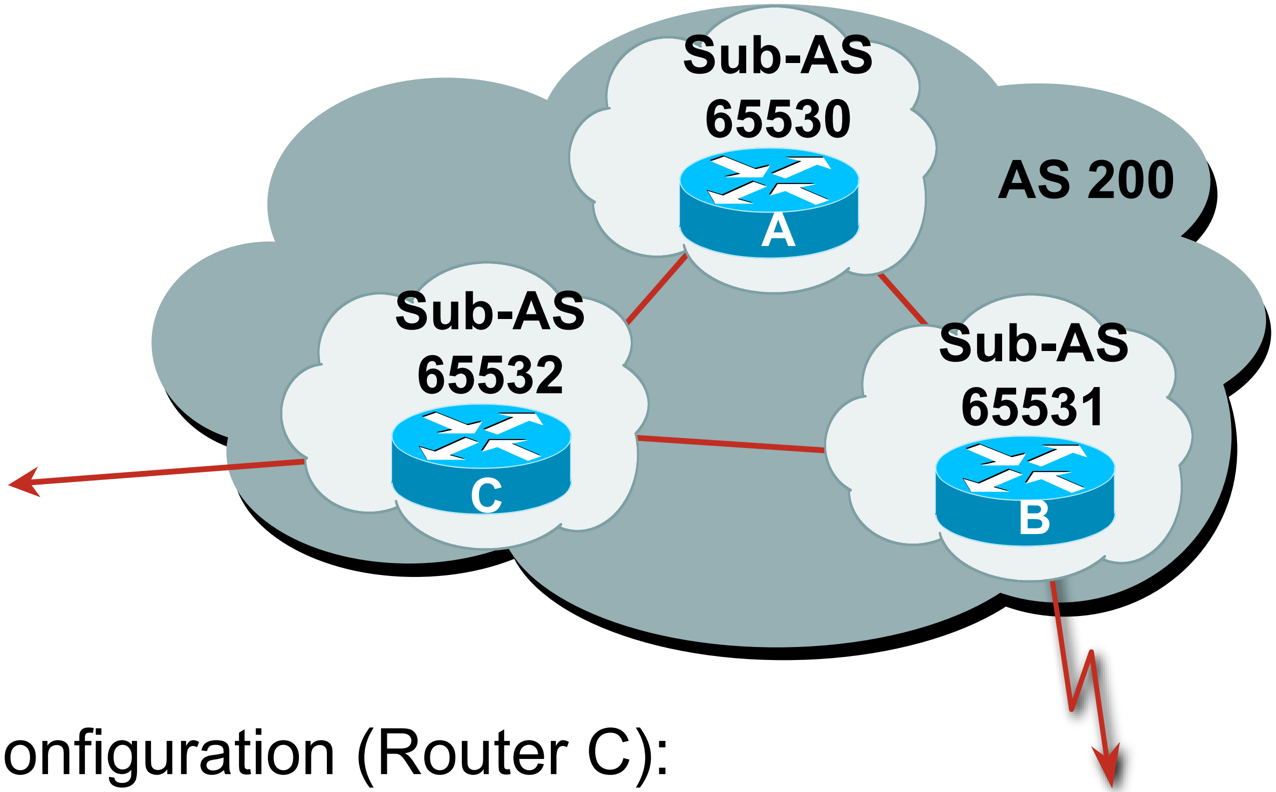
  Each sub-AS uses a number from the private AS range (64512-65534)

- 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

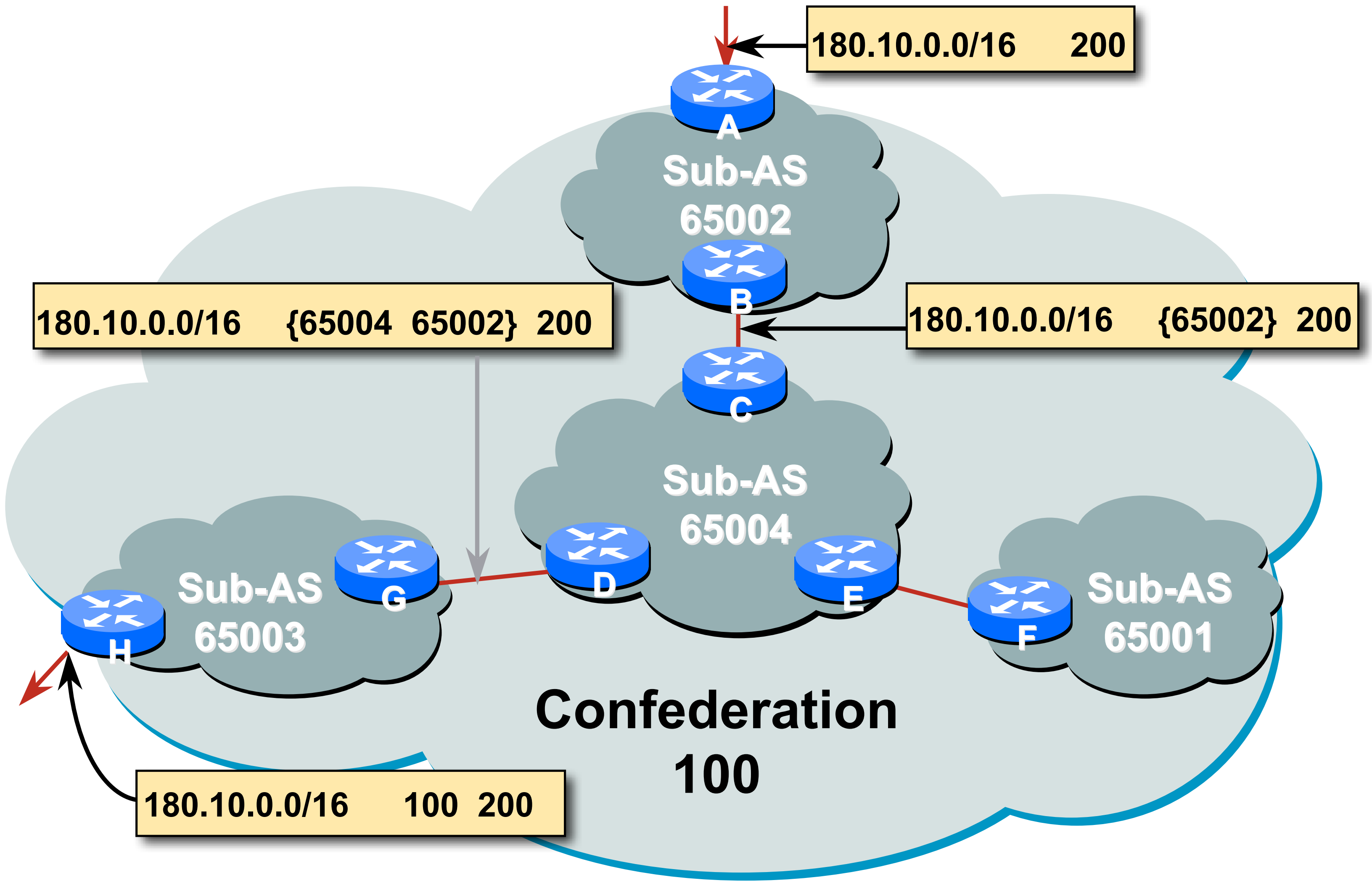
Confederations



- 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

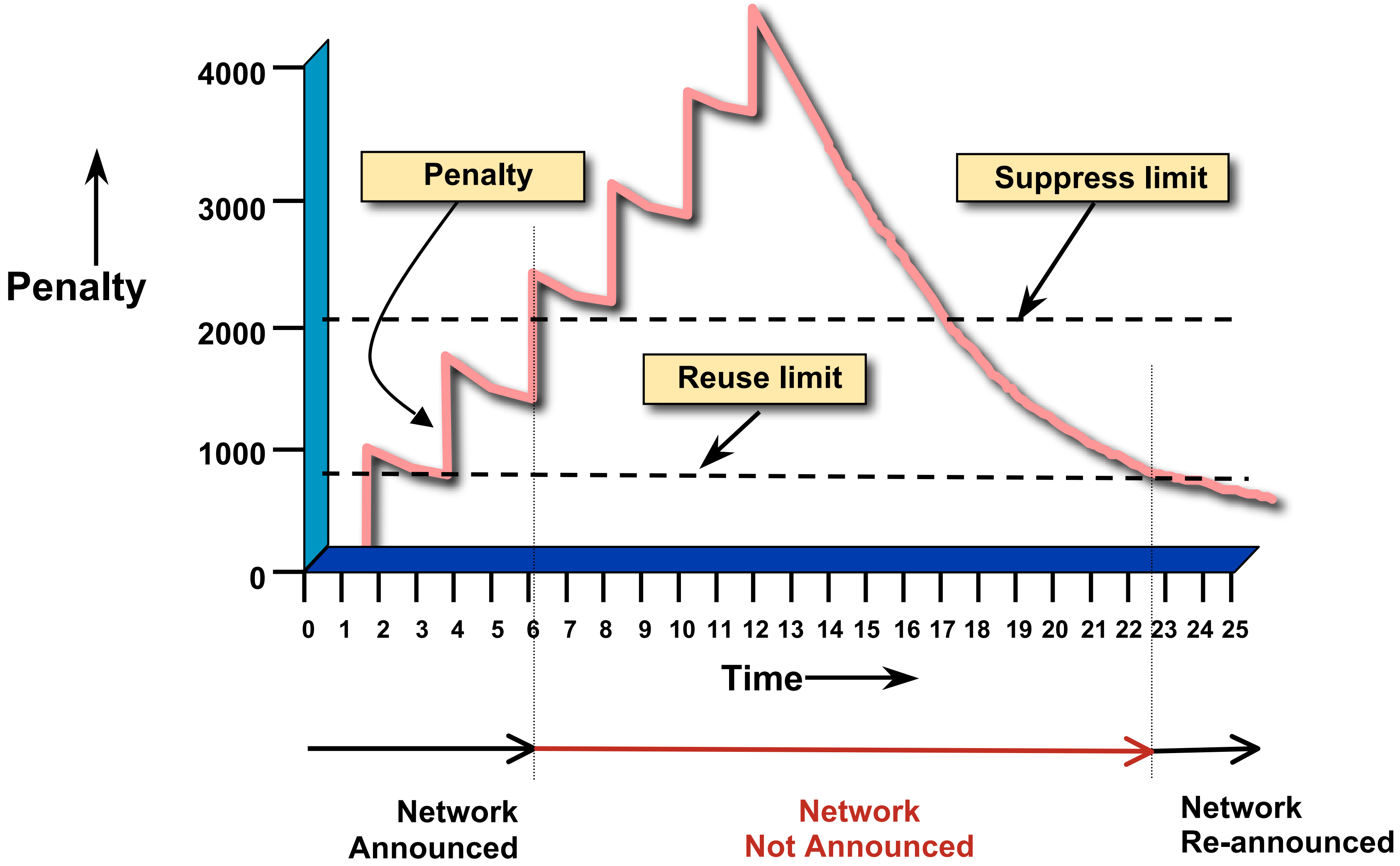
Route Flap Damping (continued)

- 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



Penalty
4000
3000
2000
1000
0
Penalty
Suppress limit
Reuse limit
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Time
Network Announced
Network Not Announced
Network Re-announced

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]

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

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

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 notifiying Sprint of your change in announcments.

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
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
```

And many many more!

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

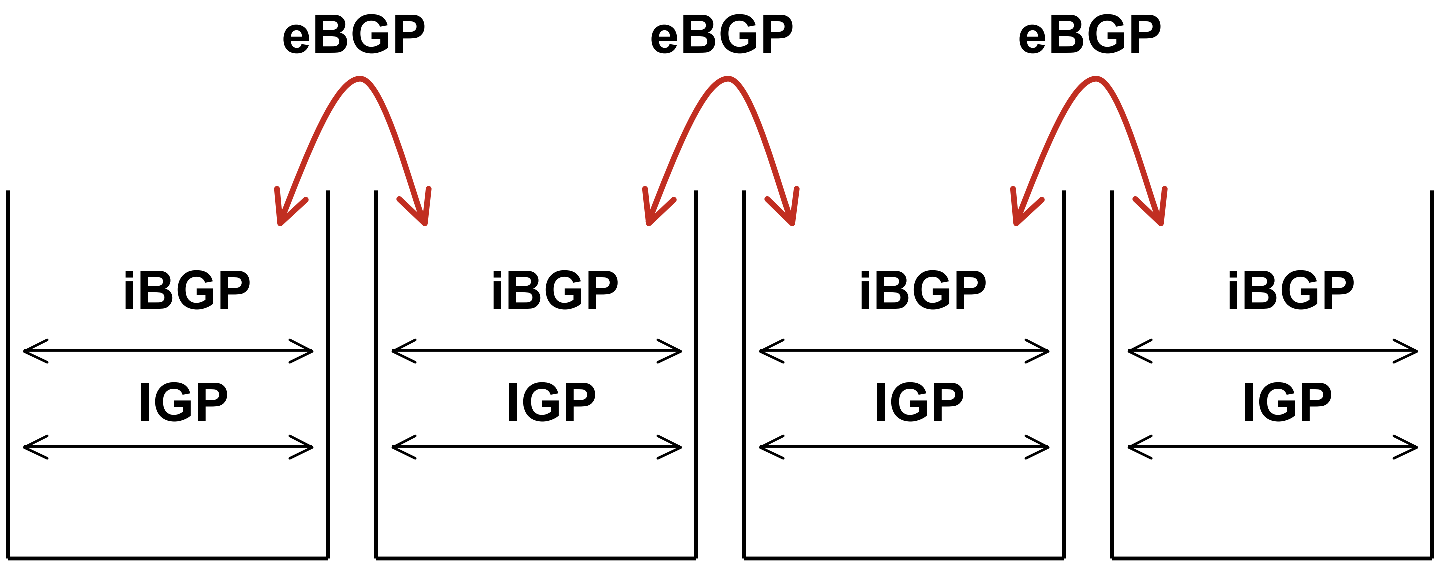
- Internal Routing Protocols (IGPs)
  - examples are ISIS and OSPF
  - used for carrying infrastructure addresses
  - **NOT** used for carrying Internet prefixes or customer prefixes
  - design goal is to minimise number of prefixes in IGP to aid scalability and rapid convergence

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

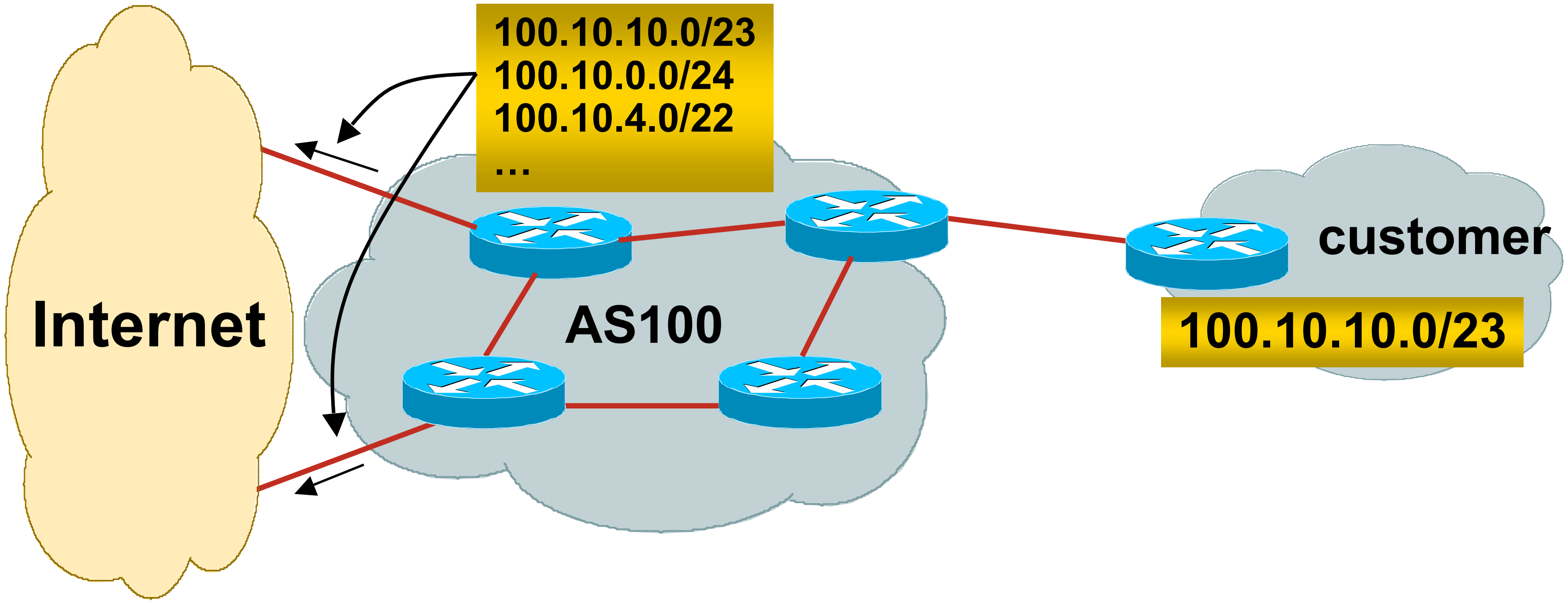
- Address block should be announced to the Internet as an aggregate
- Subprefixes of address block should NOT be announced to Internet unless special circumstances (more later)
- Aggregate should be generated internally

  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 >141000 /24s!

Aggregation – Example

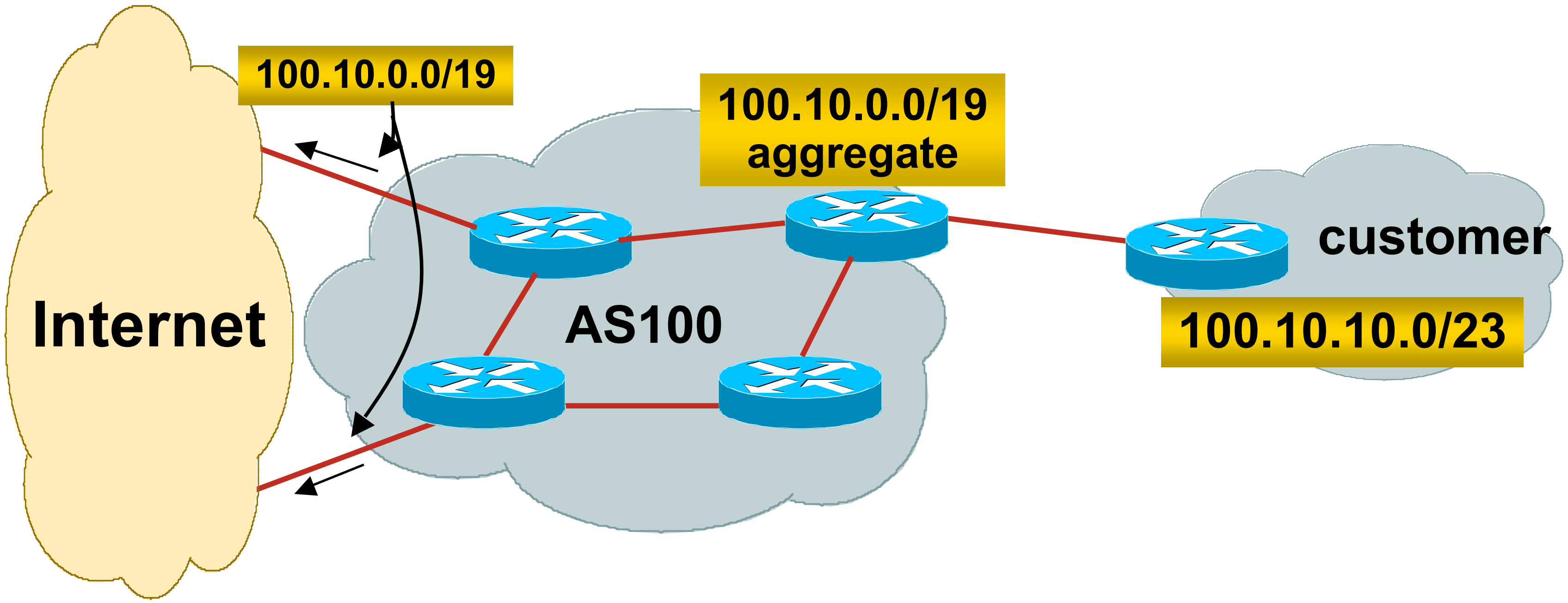


- Customer has /23 network assigned from AS100’s /19 address block
- AS100 announces customers’ individual networks to the Internet

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 (October 2008)

- Current Internet Routing Table Statistics

BGP Routing Table Entries	270153
Prefixes after maximum aggregation	130372
Unique prefixes in Internet	131760
Prefixes smaller than registry alloc	132678
/24s announced	141064
only 5753 /24s are from 192.0.0.0/8	
ASes in use	29392

“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 2008

RIR blocks contribute 219688 prefixes or 89% of the Internet Routing Table

Block	Networks	Block	Networks	Block	Networks	Block	Networks
24/8	3103	79/8	588	118/8	649	201/8	3632
41/8	1087	80/8	2162	119/8	469	202/8	10934
58/8	1479	81/8	1724	120/8	0	203/8	11000
59/8	1317	82/8	1641	121/8	1054	204/8	5601
60/8	853	83/8	1215	122/8	1600	205/8	3008
61/8	2653	84/8	1290	123/8	1225	206/8	3863
62/8	2303	85/8	2316	124/8	1787	207/8	4285
63/8	3069	86/8	768	125/8	2217	208/8	5444
64/8	5953	87/8	1484	126/8	46	209/8	5590
65/8	4012	88/8	900	173/8	0	210/8	4931
66/8	7172	89/8	2824	174/8	0	211/8	2875
67/8	2652	90/8	220	186/8	2	212/8	3015
68/8	2858	91/8	2227	187/8	6	213/8	3310
69/8	4203	96/8	255	189/8	1475	216/8	7129
70/8	1798	97/8	162	190/8	3203	217/8	2666
71/8	1186	98/8	389	192/8	6929	218/8	1375
72/8	3543	99/8	282	193/8	6220	219/8	1320
73/8	254	112/8	0	194/8	4926	220/8	2153
74/8	3002	113/8	0	195/8	4480	221/8	969
75/8	1086	114/8	4	196/8	1769	222/8	1268
76/8	1029	115/8	4	198/8	4799		
77/8	1515	116/8	1011	199/8	4116		
78/8	1169	117/8	960	200/8	8626		

"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 88 /8s will eventually cause 440000 prefix announcements
- Food for thought:
  - Remaining 39 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

CIDR Report

http://www.cidr-report.org/as2.0/

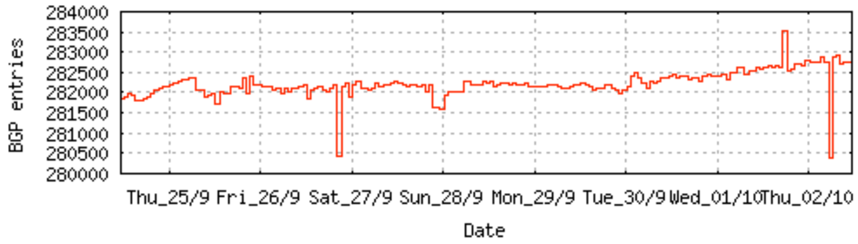
Radio ▾ Philip ▾ ADSL ▾ Networking ▾ Internet ▾ Cisco ▾ Miscellaneous ▾ TinyURL!

CIDR Report

Status Summary

Table History

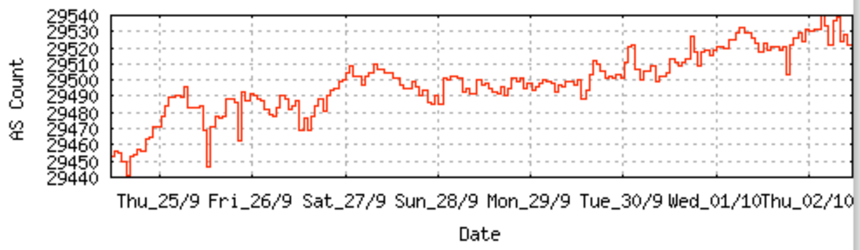
Date	Prefixes	CIDR Aggregated
25-09-08	282130	173067
26-09-08	282212	172840
27-09-08	281895	173376
28-09-08	281607	173846
29-09-08	282138	174099
30-09-08	282044	173861
01-10-08	282391	174307
02-10-08	282791	171834



Plot: BGP Table Size

AS Summary

29528	Number of ASes in routing system
12509	Number of ASes announcing only one prefix
5033	Largest number of prefixes announced by an AS AS4538: ERX-CERNET-BKB China Education and Research Network Center
88349184	Largest address span announced by an AS (/32s) AS721: DISA-ASNBLK - 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)

CIDR Report

http://www.cidr-report.org/as2.0/

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').

--- 02Oct08 ---

ASnum	NetsNow	NetsAggr	NetGain	% Gain	Description
Table	282810	171877	110933	39.2%	All ASes
AS4538	5033	880	4153	82.5%	ERX-CERNET-BKB China Education and Research Network Center
AS6389	4300	351	3949	91.8%	BELLSOUTH-NET-BLK - BellSouth.net Inc.
AS209	2948	1333	1615	54.8%	ASN-QWEST - Qwest
AS1785	1670	161	1509	90.4%	AS-PAETEC-NET - PaeTec Communications, Inc.
AS6298	2010	717	1293	64.3%	COX-PHX - Cox Communications Inc.
AS4755	1455	272	1183	81.3%	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
AS17488	1393	300	1093	78.5%	HATHWAY-NET-AP Hathway IP Over Cable Internet
AS4323	1531	586	945	61.7%	TWTC - tw telecom holdings, inc.
AS8151	1410	543	867	61.5%	Uninet S.A. de C.V.
AS22773	991	190	801	80.8%	CCINET-2 - Cox Communications Inc.
AS19262	953	174	779	81.7%	VZGNI-TRANSIT - Verizon Internet Services Inc.
AS11492	1215	443	772	63.5%	CABLEONE - CABLE ONE
AS18566	1055	322	733	69.5%	COVAD - Covad Communications Co.
AS18101	782	91	691	88.4%	RIL-IDC Reliance Infocom Ltd Internet Data Centre,
AS2386	1560	916	644	41.3%	INS-AS - AT&T Data Communications Services
AS9498	678	71	607	89.5%	BBIL-AP BHARTI Airtel Ltd.
AS6478	1195	593	602	50.4%	ATT-INTERNET3 - AT&T WorldNet Services
AS3356	1033	541	492	47.6%	LEVEL3 Level 3 Communications
AS855	596	120	476	79.9%	CANET-ASN-4 - Bell Aliant
AS4766	903	427	476	52.7%	KIXS-AS-KR Korea Telecom
AS4808	616	145	471	76.5%	CHINA169-BJ CNCGROUP IP network China169 Beijing Province Network
AS20115	1806	1336	470	26.0%	CHARTER-NET-HKY-NC - Charter Communications
AS17676	524	64	460	87.8%	GIGAINFRA BB TECHNOLOGY Corp.
AS9443	524	77	447	85.3%	INTERNETPRIMUS-AS-AP Primus Telecommunications
AS7011	913	476	437	47.9%	FRONTIER-AND-CITIZENS - Frontier Communications of America, Inc.

CIDR Report

http://www.cidr-report.org/as2.0/

Radio Philip ADSL Networking Internet Cisco Miscellaneous TinyURL!

CIDR Report

Top 20 Added Routes this week per Originating AS

Prefixes	ASnum	AS Description
194	AS17908	TCISL Tata Communications
128	AS20115	CHARTER-NET-HKY-NC - Charter Communications
69	AS10620	TV Cable S.A.
37	AS11139	CWRIN CW BARBADOS
35	AS37054	DTS
32	AS47992	MARYANEWEB-AS Mary & Anne TRADING SRL
32	AS47966	IG-AS I & G 2000 IMPEX SRL
28	AS4847	CNIX-AP China Networks Inter-Exchange
22	AS31793	BROADSTAR - BroadStar
21	AS6478	ATT-INTERNET3 - AT&T WorldNet Services
19	AS747	TAEGU-AS - DoD Network Information Center
18	AS21769	AS-COLOAM - Colocation America Corporation
18	AS18101	RIL-IDC Reliance Infocom Ltd Internet Data Centre,
17	AS6298	COX-PHX - Cox Communications Inc.
17	AS17524	DSN DS Networks
17	AS27855	AXESAT S.A
16	AS4323	TWTC - tw telecom holdings, inc.
16	AS14576	RHNL-NET - Righthosting.com
16	AS47931	ALENETWORK A.L.E. COM NETWORK S.R.L
15	AS16712	Soft Seven Informática Ltda.

Top 20 Withdrawn Routes this week per Originating AS

Prefixes	ASnum	AS Description
-202	AS4755	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
-91	AS10507	SPCS - Sprint Personal Communications Systems
-56	AS7029	WINDSTREAM - Windstream Communications Inc
-46	AS2706	HKSUPER-HK-AP Pacific Internet (Hong Kong) Limited
-42	AS17964	DXTNET Beijing Dian-Xin-Tong Network Technologies Co., Ltd.
-33	AS38107	CDNETWORKS-AS-KR CDNetworks
-28	AS15611	Iranian Research Organization for Science & Technology
-26	AS15582	COMCORTV-AS COMCOR-TV Autonomous System
-23	AS2920	LACOE - Los Angeles County Office of Education
-21	AS5511	OPENTRANSIT France Telecom - Orange
-20	AS6006	DDN-ASNBLK - DoD Network Information Center

CIDR Report

http://www.cidr-report.org/as2.0/

More Specifics

A list of route advertisements that appear to be more specfic 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
4954	5033	AS4538	ERX-CERNET-BKB China Education and Research Network Center
4152	4300	AS6389	BELLSOUTH-NET-BLK - BellSouth.net Inc.
2742	2948	AS209	ASN-QWEST - Qwest
2004	2010	AS6298	COX-PHX - Cox Communications Inc.
1767	1806	AS20115	CHARTER-NET-HKY-NC - Charter Communications
1587	1670	AS1785	AS-PAETEC-NET - PaeTec Communications, Inc.
1460	1560	AS2386	INS-AS - AT&T Data Communications Services
1434	1455	AS4755	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
1403	1410	AS8151	Uninet S.A. de C.V.
1393	1393	AS17488	HATHWAY-NET-AP Hathway IP Over Cable Internet
1328	1531	AS4323	TWTC - tw telecom holdings, inc.
1320	1322	AS1803	ICMNET-5 - Sprint
1200	1215	AS11492	CABLEONE - CABLE ONE
1195	1195	AS6478	ATT-INTERNET3 - AT&T WorldNet Services
1156	1416	AS7018	ATT-INTERNET4 - AT&T WorldNet Services
1107	1108	AS9583	SIFY-AS-IN Sify Limited
1045	1055	AS18566	COVAD - Covad Communications Co.
973	973	AS23577	ATM-MPLS-AS-KR Korea Telecom
955	991	AS22773	CCINET-2 - Cox Communications Inc.
915	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)

AS Report

http://www.cidr-report.org/cgi-bin/as-report?as=AS4755&view=2.0

Radio Philip ADSL Networking Internet Cisco Miscellaneous TinyURL!

AS Report

Announced Prefixes

Rank	AS	Type	Originate Addr Space	(pfx)	Transit Addr space	(pfx)	Description
101	AS4755	ORG+TRN	Originate: 3728384	/10.17	Transit: 3726592	/10.17	TATACOMM-AS TATA Communications formerly VSN

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	1455	1245	62	272	1183	81.31%

Prefix	AS Path	Aggregation Suggestion
59.151.144.0/22	4777 2516 4755	
59.160.0.0/14	4777 2516 4755	
59.160.0.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.4.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.5.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.8.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.12.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.15.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.16.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.24.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.24.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.28.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.32.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.38.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.40.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.44.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.48.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.48.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.56.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.64.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.71.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.72.0/21	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.73.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.81.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.82.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.83.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.88.0/22	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.88.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.89.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.96.0/20	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755
59.160.97.0/24	4777 2516 4755	- Withdrawn - matching aggregate 59.160.0.0/14 4777 2516 4755

AS Report

http://www.cidr-report.org/cgi-bin/as-report?as=AS18566&view=2.0

Announced Prefixes

Rank	AS	Type	Originate Addr Space	(pfx)	Transit Addr space	(pfx)	Description
144	AS18566	ORIGIN	Originate:	2348288 /10.84	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	Wthdw	Aggte	Annce	Redctn	%
14	AS18566	COVAD - Covad Communications Co.	1055	895	162	322	733	69.48%

```
Prefix              AS Path                         Aggregation Suggestion
64.105.0.0/16       12654 7018 2828 18566
64.105.0.0/23       12654 3257 2828 18566
64.105.4.0/22       12654 3257 2828 18566 + Announce - aggregate of 64.105.4.0/23 (12654 3257 2828 18566) and 64.105.6.0/23 (12
64.105.4.0/23       12654 3257 2828 18566 - Withdrawn - aggregated with 64.105.6.0/23 (12654 3257 2828 18566)
64.105.6.0/23       12654 3257 2828 18566 - Withdrawn - aggregated with 64.105.4.0/23 (12654 3257 2828 18566)
64.105.8.0/23       12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
64.105.10.0/23      12654 3257 2828 18566
64.105.14.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
64.105.16.0/23      12654 3257 2828 18566 + Announce - aggregate of 64.105.16.0/24 (12654 3257 2828 18566) and 64.105.17.0/24 (
64.105.16.0/24      12654 3257 2828 18566 - Withdrawn - aggregated with 64.105.17.0/24 (12654 3257 2828 18566)
64.105.17.0/24      12654 3257 2828 18566 - Withdrawn - aggregated with 64.105.16.0/24 (12654 3257 2828 18566)
64.105.18.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
64.105.20.0/22      12654 3257 2828 18566 + Announce - aggregate of 64.105.20.0/23 (12654 3257 2828 18566) and 64.105.22.0/23 (
64.105.20.0/23      12654 3257 2828 18566 - Withdrawn - aggregated with 64.105.22.0/23 (12654 3257 2828 18566)
64.105.22.0/23      12654 3257 2828 18566 - Withdrawn - aggregated with 64.105.20.0/23 (12654 3257 2828 18566)
64.105.24.0/21      12654 3257 2828 18566
64.105.32.0/21      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
64.105.40.0/22      12654 3257 2828 18566 + Announce - aggregate of 64.105.40.0/23 (12654 3257 2828 18566) and 64.105.42.0/23 (
64.105.40.0/23      12654 3257 2828 18566 - Withdrawn - aggregated with 64.105.42.0/23 (12654 3257 2828 18566)
64.105.42.0/23      12654 3257 2828 18566 - Withdrawn - aggregated with 64.105.40.0/23 (12654 3257 2828 18566)
64.105.44.0/23      12654 3257 2828 18566
64.105.46.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
64.105.48.0/22      12654 3257 2828 18566 + Announce - aggregate of 64.105.48.0/23 (12654 3257 2828 18566) and 64.105.50.0/23 (
64.105.48.0/23      12654 3257 2828 18566 - Withdrawn - aggregated with 64.105.50.0/23 (12654 3257 2828 18566)
64.105.50.0/23      12654 3257 2828 18566 - Withdrawn - aggregated with 64.105.48.0/23 (12654 3257 2828 18566)
64.105.52.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
64.105.54.0/23      12654 3257 2828 18566
64.105.56.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
64.105.58.0/23      12654 3257 2828 18566
64.105.60.0/22      12654 3257 2828 18566 + Announce - aggregate of 64.105.60.0/23 (12654 3257 2828 18566) and 64.105.62.0/23 (
64.105.60.0/23      12654 3257 2828 18566 - Withdrawn - aggregated with 64.105.62.0/23 (12654 3257 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

http://bgpupdates.potaroo.net/instability/bgpupd.html

Radio Philip ADSL Networking Internet Cisco Miscellaneous TinyURL!

The BGP Instability Report

The BGP Instability Report is updated daily. This report was generated on 02 October 2008 03:56 (UTC+1000)

50 Most active ASes for the past 31 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	9583	275795	3.14%	1235	223.32	SIFY-AS-IN Sify Limited
2	1803	112630	1.28%	1357	83.00	ICMNET-5 - Sprint
3	4538	104412	1.19%	5036	20.73	ERX-CERNET-BKB China Education and Research Network Center
4	5691	78864	0.90%	13	6066.46	MITRE-AS-5 - The MITRE Corporation
5	8151	73547	0.84%	2447	30.06	Uninet S.A. de C.V.
6	6389	68007	0.77%	4353	15.62	BELLSOUTH-NET-BLK - BellSouth.net Inc.
7	9051	62029	0.71%	159	390.12	IDM Autonomous System
8	4184	53618	0.61%	2	26809.00	STORTEK-WHQ - Storage Technology Corporation
9	14593	51965	0.59%	1	51965.00	BRAND-INSTITUTE - Brand Instiute, Inc.
10	10396	49963	0.57%	55	908.42	COQUI-NET - DATACOM CARIBE, INC.
11	20255	48680	0.55%	24	2028.33	Tecnowind S.A.
12	4274	46547	0.53%	68	684.51	ERX-AU-NET Assumption University
13	209	45939	0.52%	3011	15.26	ASN-QWEST - Qwest
14	11971	43557	0.50%	7	6222.43	PFIZERNET-GROTON - PFIZER INC.
15	30890	40681	0.46%	1357	29.98	EVOLVA Evolva Telecom
16	20115	38378	0.44%	1997	19.22	CHARTER-NET-HKY-NC - Charter Communications
17	7018	38105	0.43%	1477	25.80	ATT-INTERNET4 - AT&T WorldNet Services
18	18231	36236	0.41%	249	145.53	EXATT-AS-AP IOL NETCOM LTD
19	17488	34829	0.40%	1492	23.34	HATHWAY-NET-AP Hathway IP Over Cable Internet
20	8866	34332	0.39%	332	103.41	BTC-AS Bulgarian Telecommunication Company Plc.
21	6458	34250	0.39%	341	100.44	Telgua
22	33783	34036	0.39%	142	239.69	EEPAD
23	30969	32153	0.37%	8	4019.12	TAN-NET TransAfrica Networks

The BGP Instability Report

http://bgpupdates.potaroo.net/instability/bgpupd.html

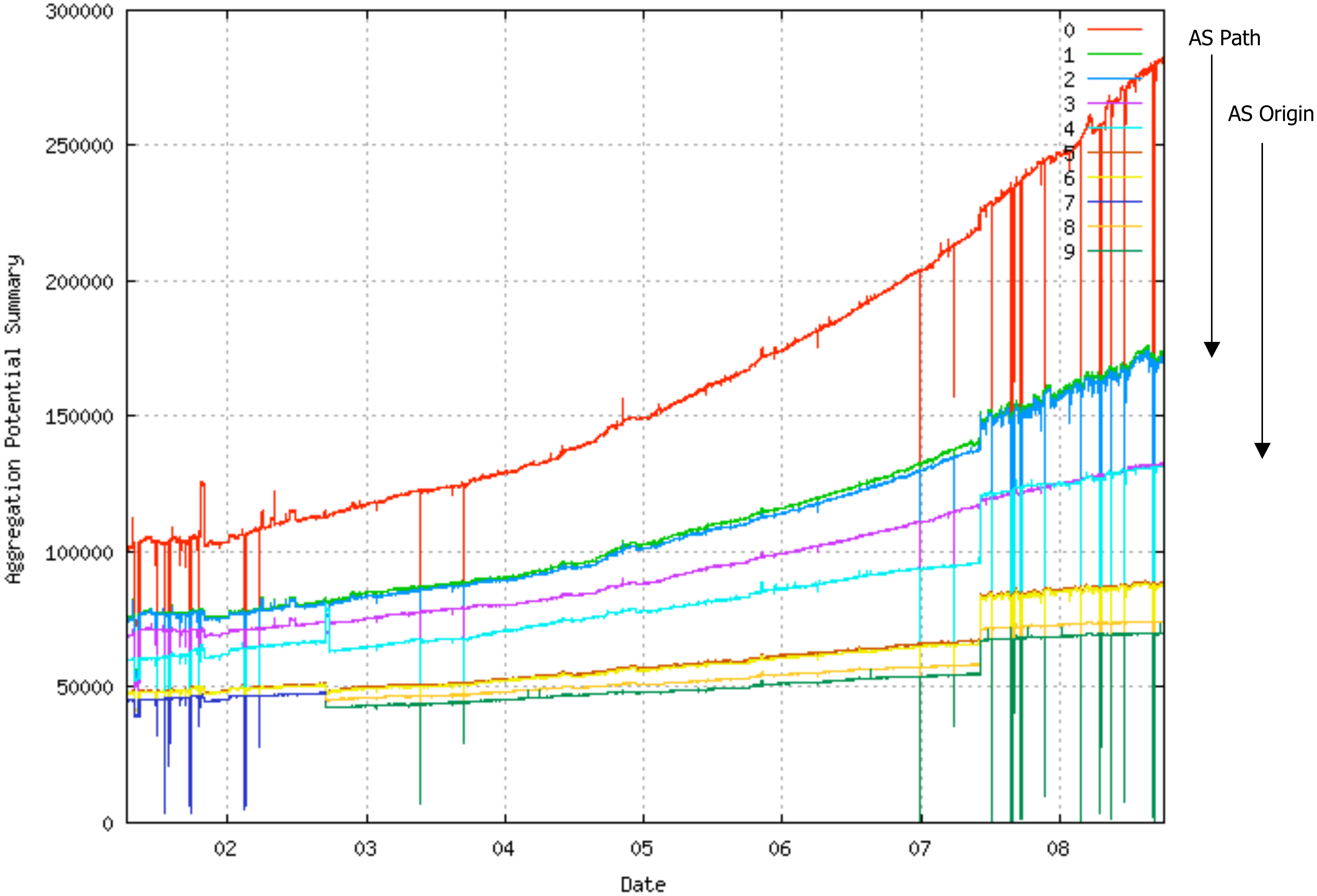
Radio Philip ADSL Networking Internet Cisco Miscellaneous TinyURL!

The BGP Instability Report

50 Most active Prefixes for the past 31 days

RANK	PREFIX	UPDs	%	Origin AS -- AS NAME
1	192.12.120.0/24	78753	0.84%	5691 -- MITRE-AS-5 - The MITRE Corporation
2	210.214.151.0/24	61905	0.66%	9583 -- SIFY-AS-IN Sify Limited
3	221.134.222.0/24	58307	0.62%	9583 -- SIFY-AS-IN Sify Limited
4	194.126.143.0/24	52762	0.56%	9051 -- IDM Autonomous System
5	12.8.7.0/24	51965	0.56%	14593 -- BRAND-INSTITUTE - Brand Instiute, Inc.
6	221.135.80.0/24	48043	0.51%	9583 -- SIFY-AS-IN Sify Limited
7	210.210.112.0/24	47034	0.50%	9583 -- SIFY-AS-IN Sify Limited
8	12.18.36.0/24	43289	0.46%	11971 -- PFIZERNET-GROTON - PFIZER INC.
9	221.135.251.0/24	34665	0.37%	9583 -- SIFY-AS-IN Sify Limited
10	221.128.192.0/18	28066	0.30%	18231 -- EXATT-AS-AP IOL NETCOM LTD
11	199.117.144.0/22	26810	0.29%	4184 -- STORTEK-WHQ - Storage Technology Corporation
12	129.80.0.0/16	26808	0.29%	4184 -- STORTEK-WHQ - Storage Technology Corporation
13	200.108.200.0/24	24612	0.26%	20255 -- Tecnowind S.A.
14	72.50.96.0/20	24525	0.26%	10396 -- COQUI-NET - DATACOM CARIBE, INC.
15	196.42.0.0/20	24506	0.26%	10396 -- COQUI-NET - DATACOM CARIBE, INC.
16	200.108.220.0/24	23626	0.25%	20255 -- Tecnowind S.A.
17	83.228.71.0/24	23266	0.25%	8866 -- BTC-AS Bulgarian Telecommunication Company Plc.
18	193.93.148.0/22	18591	0.20%	8266 -- NEXUSTEL Nexus Telecommunications
19	196.27.108.0/22	15866	0.17%	30969 -- TAN-NET TransAfrica Networks
20	196.27.104.0/21	15848	0.17%	30969 -- TAN-NET TransAfrica Networks
21	89.4.131.0/24	13760	0.15%	24731 -- ASN-NESMA National Engineering Services and Marketing Company Ltd. (NESMA)
22	205.162.132.0/23	12644	0.14%	23541 -- Scarlet B.V.
23	64.162.116.0/24	10820	0.12%	5033 -- ISW - Internet Specialties West Inc.
24	89.38.98.0/23	10655	0.11%	6663 -- EUROWEBRO Euroweb Romania SA
25	86.105.182.0/24	10643	0.11%	6663 -- EUROWEBRO Euroweb Romania SA
26	203.63.26.0/24	10132	0.11%	9747 -- EZINTERNET-AS-AP EZInternet Pty Ltd
27	195.251.5.0/24	9519	0.10%	5408 -- GR-NET Greek Research & Technology Network, http://www.grnet.gr
28	192.221.76.0/24	7148	0.08%	10026 -- ANC Asia Netcom 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:       davids@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

Configuration Tips

Of passwords, tricks and templates

iBGP and IGPs
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"

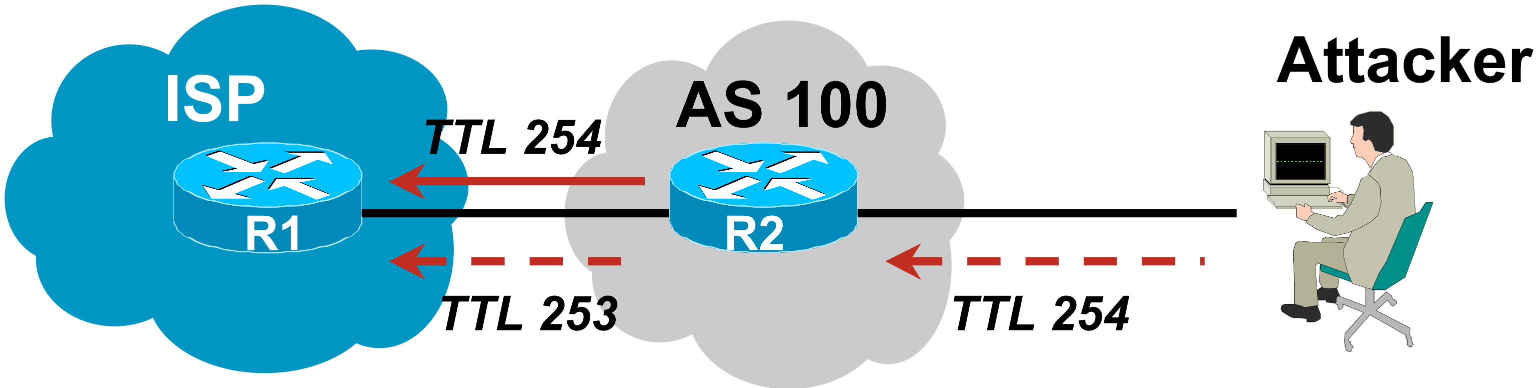
- Implement RFC5082 on BGP peerings

  (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

End of Tutorial!