BGP Techniques for Internet Service Providers

Philip Smith <pfs@cisco.com>

APRICOT 2011

Hong Kong, SAR, China

15 - 25 February 2011

Presentation Slides

- Will be available on

  ftp://ftp-eng.cisco.com

  /pfs/seminars/APRICOT2011-BGP-Techniques.pdf

  And on the APRICOT 2011 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

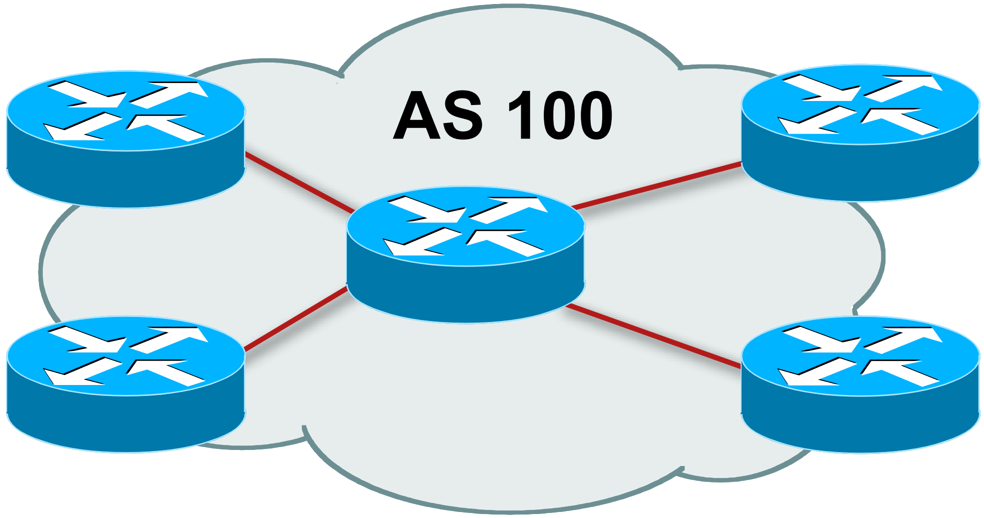
BGP Basics

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)

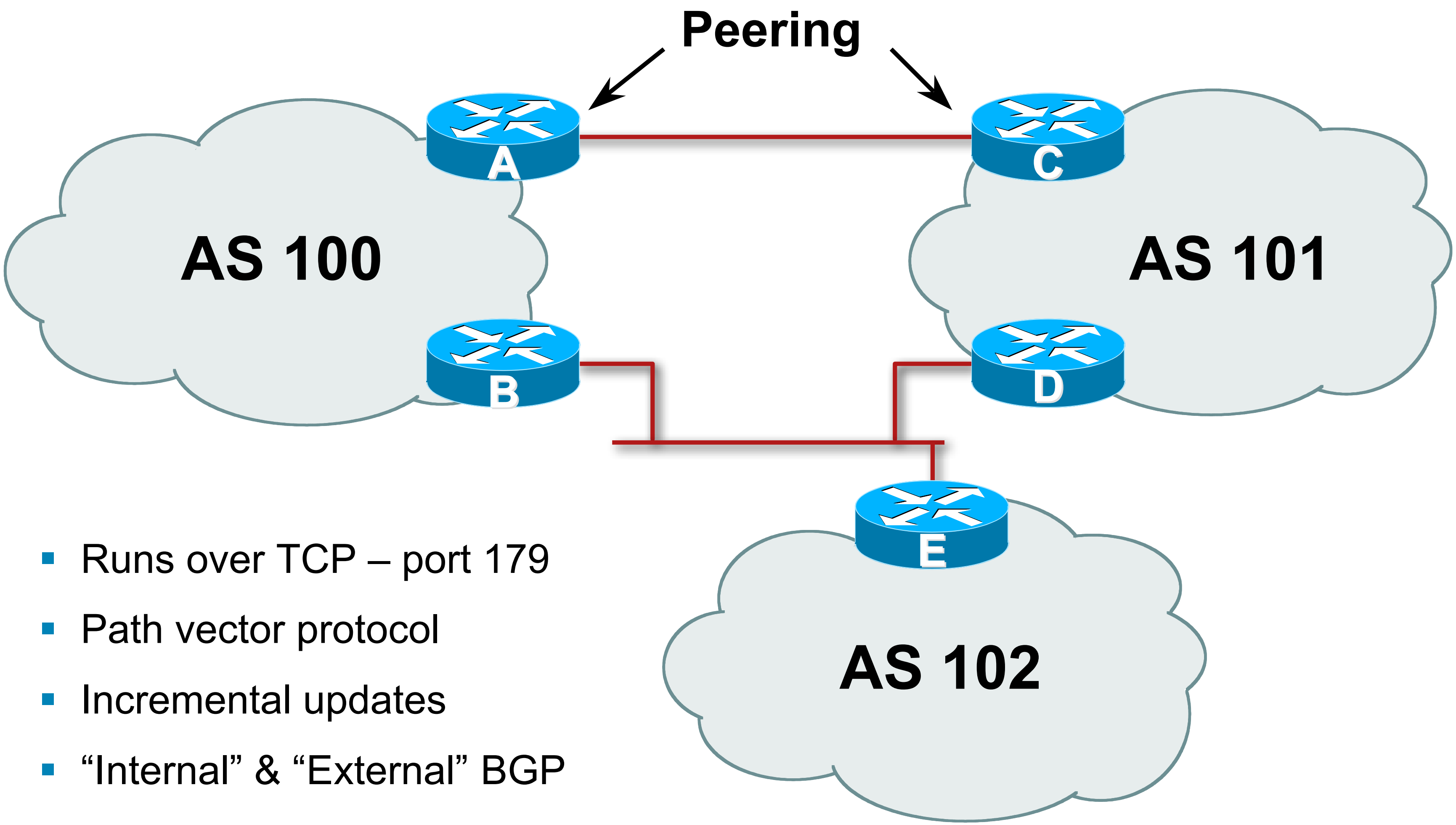
- 32-bit range representation specified in RFC5396

  Defines “asplain” (traditional format) as standard notation

Autonomous System Number (ASN)

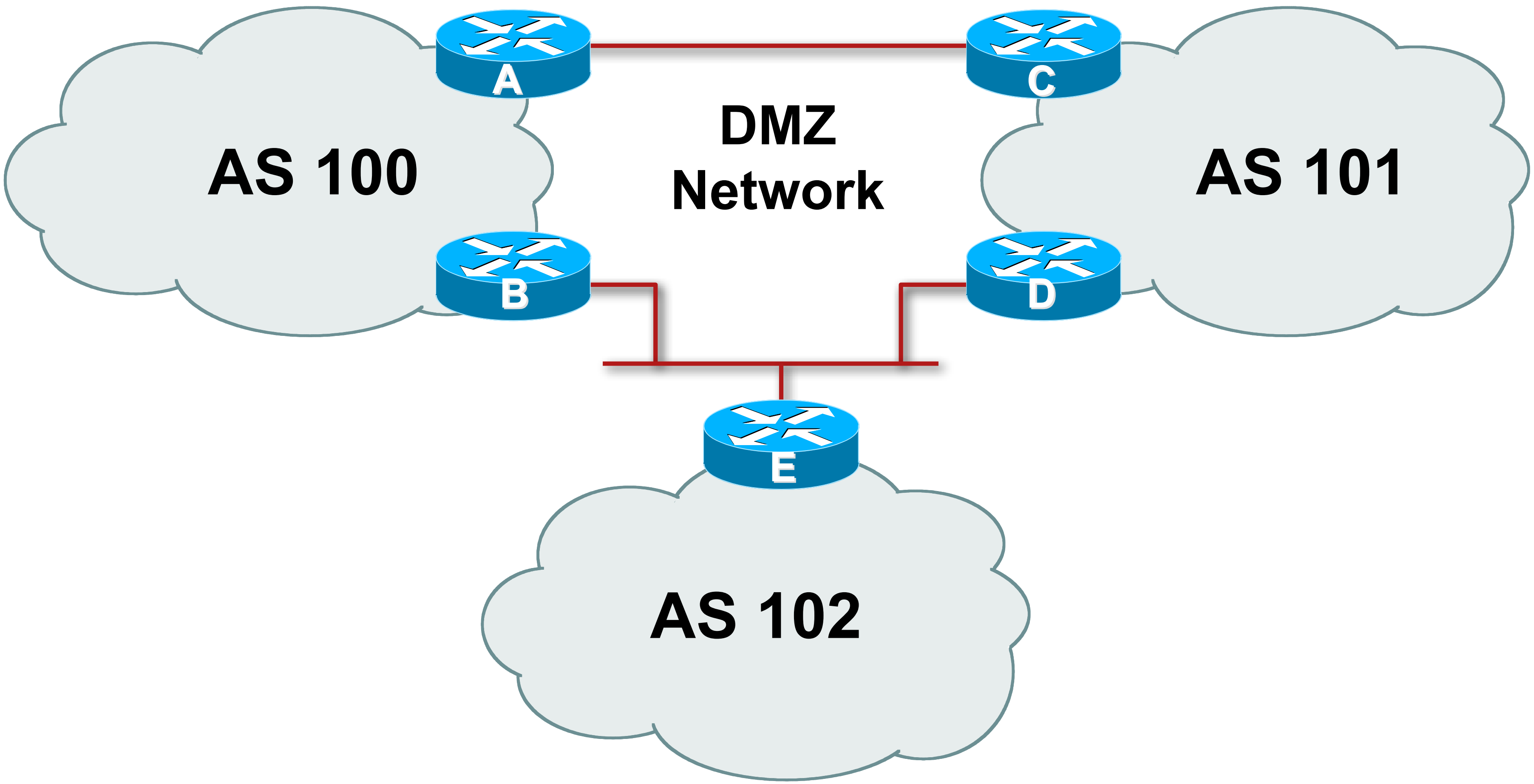
- ASNs are distributed by the Regional Internet Registries
  - They are also available from upstream ISPs who are members of one of the RIRs
- Current 16-bit ASN allocations up to 58367 have been made to the RIRs
  - Around 3600 are visible on the Internet
- Each RIR has also received a block of 32-bit ASNs
  - Out of 1063 assignments, around 600 are visible on the Internet
- 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

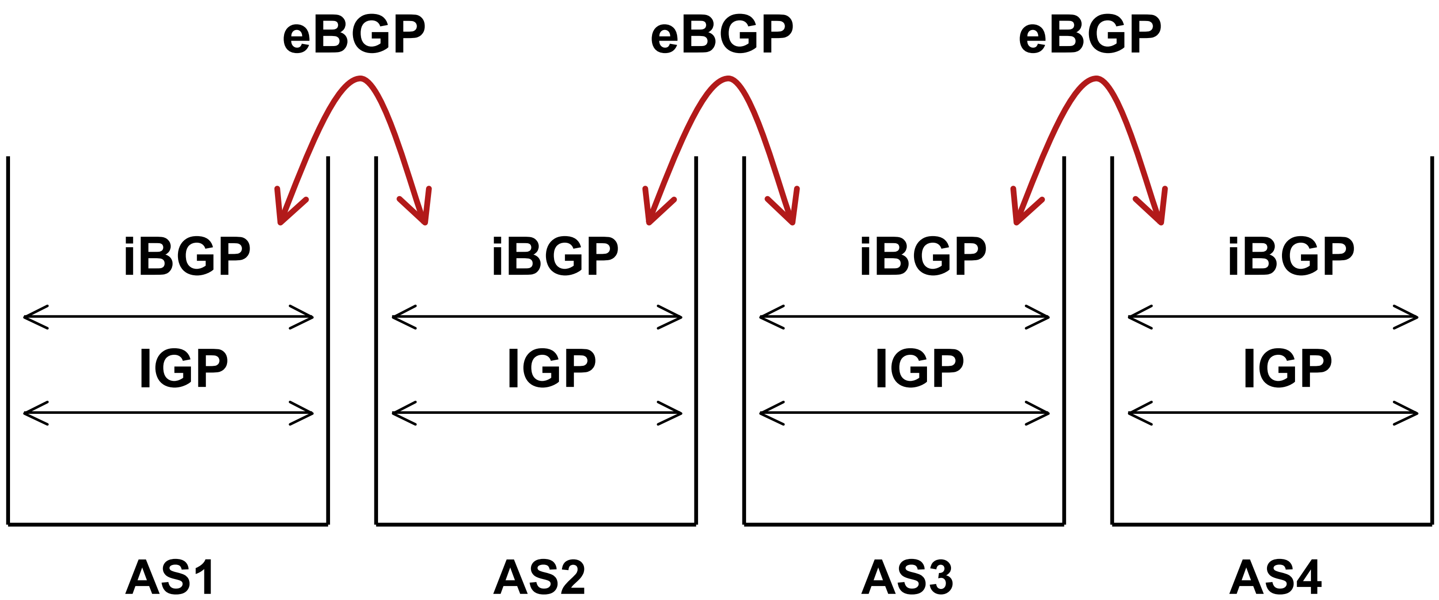
- Learns multiple paths via internal and external BGP speakers
- Picks the best path and installs in the forwarding table
- Best path is sent to external BGP neighbours
- Policies are applied by influencing the best path selection

eBGP & iBGP

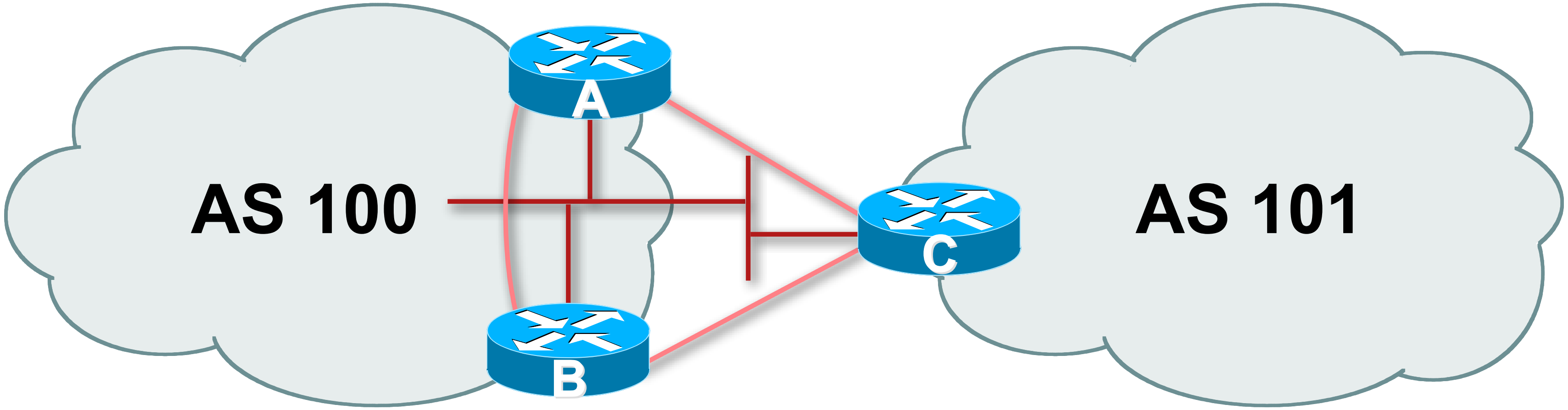
- BGP used internally (iBGP) and externally (eBGP)
- iBGP used to carry
  - Some/all Internet prefixes across ISP backbone
  - ISP’s customer prefixes
- eBGP used to
  - Exchange prefixes with other ASes
  - Implement routing policy

BGP/IGP model used in ISP networks

- Model representation



External BGP Peering (eBGP)

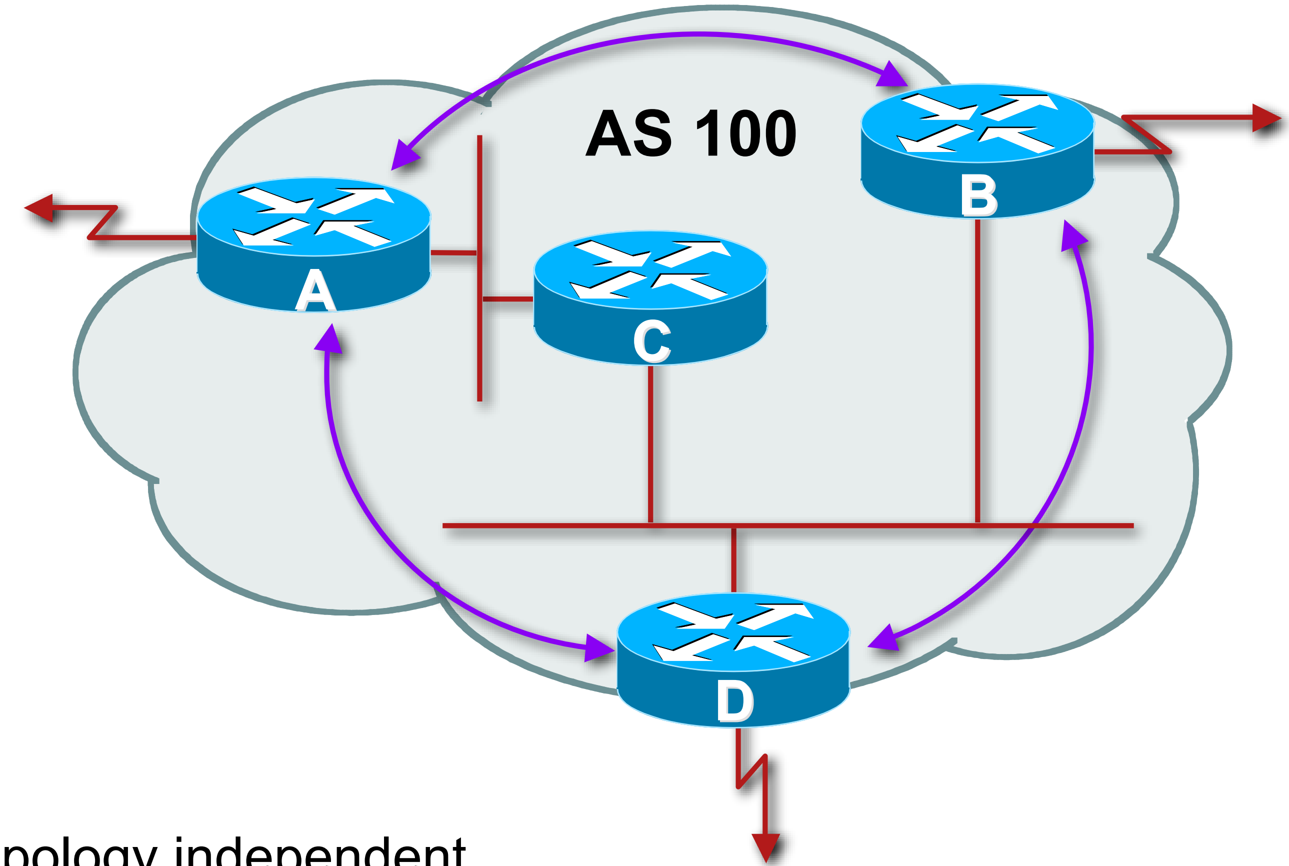


- Between BGP speakers in different AS
- Should be directly connected
- **Never** run an IGP between eBGP peers

Internal BGP (iBGP)

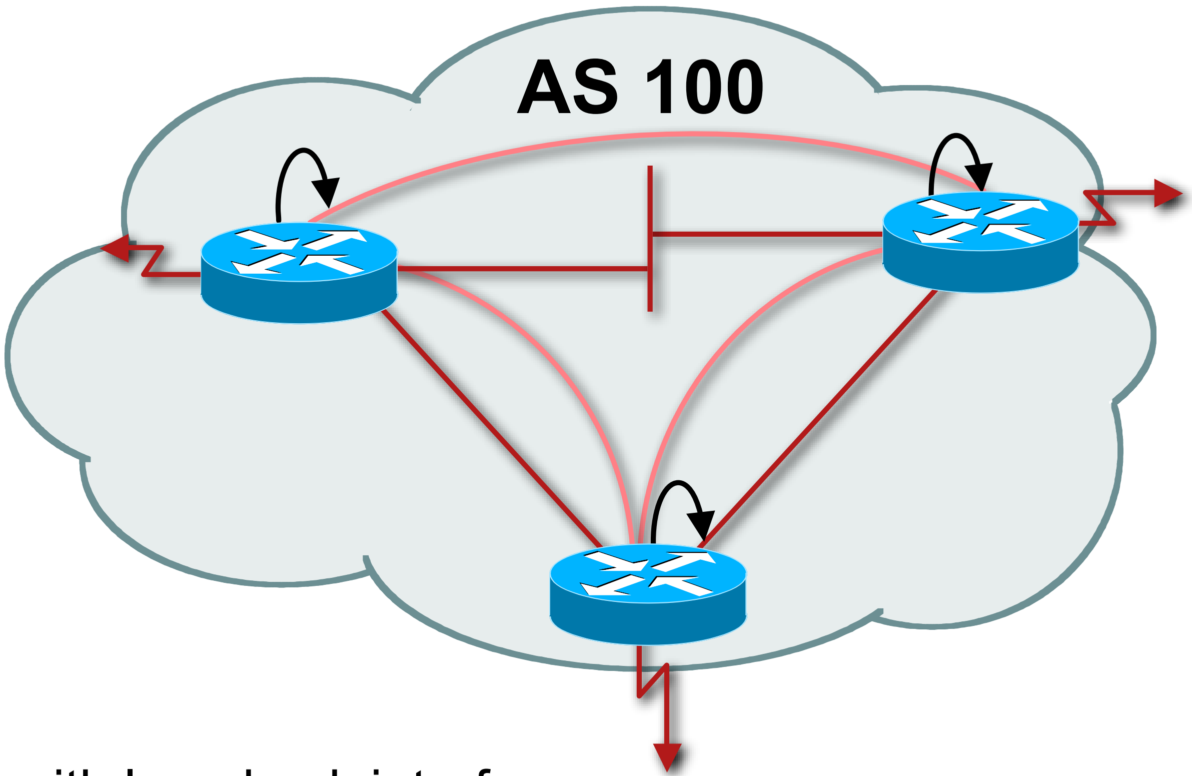
- BGP peer within the same AS
- Not required to be directly connected
  - IGP takes care of inter-BGP speaker connectivity
- iBGP speakers must to be fully meshed:
  - They originate connected networks
  - They pass on prefixes learned from outside the ASN
  - They do **not** pass on prefixes learned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

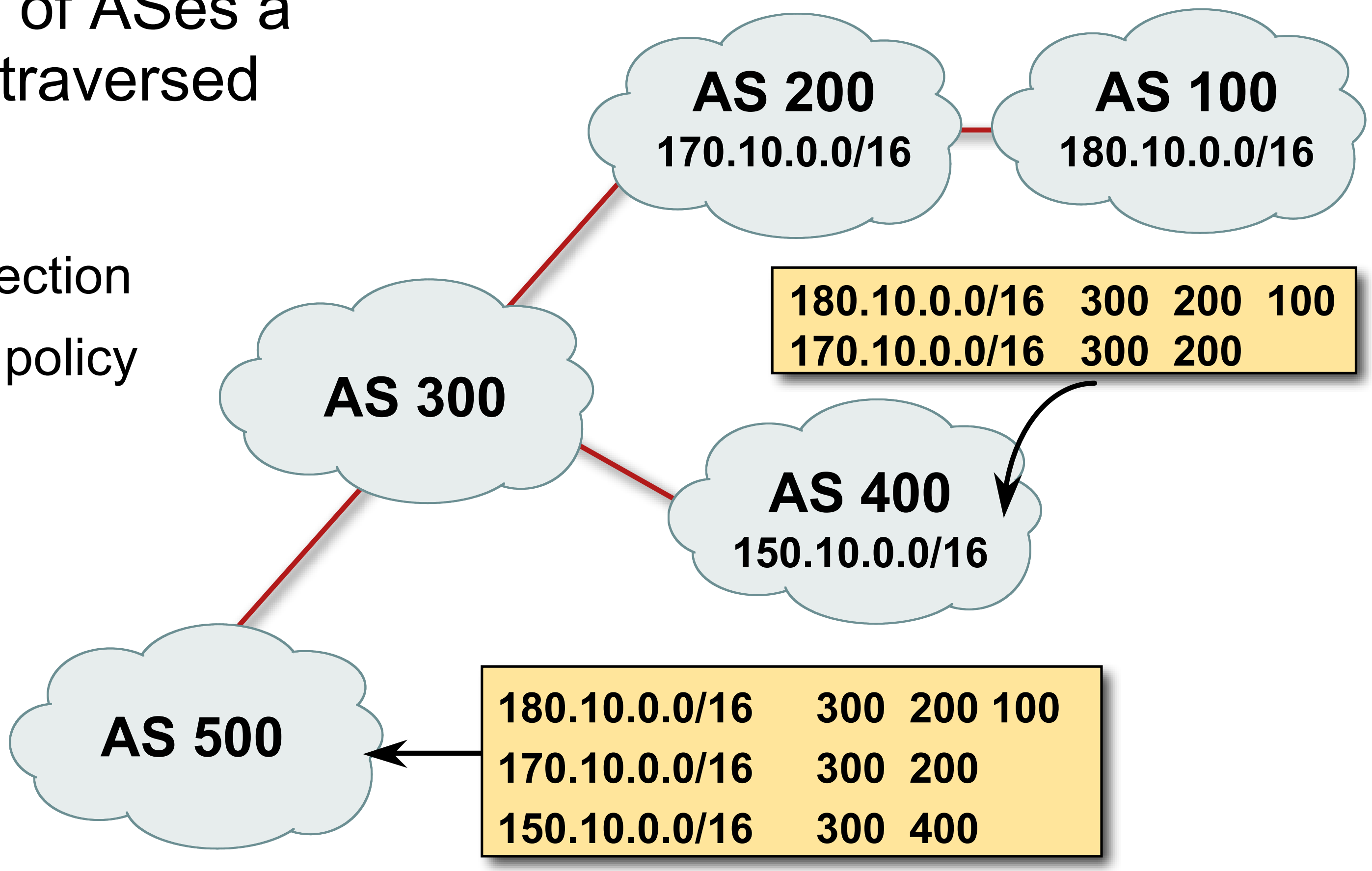
Information about BGP

AS-Path

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

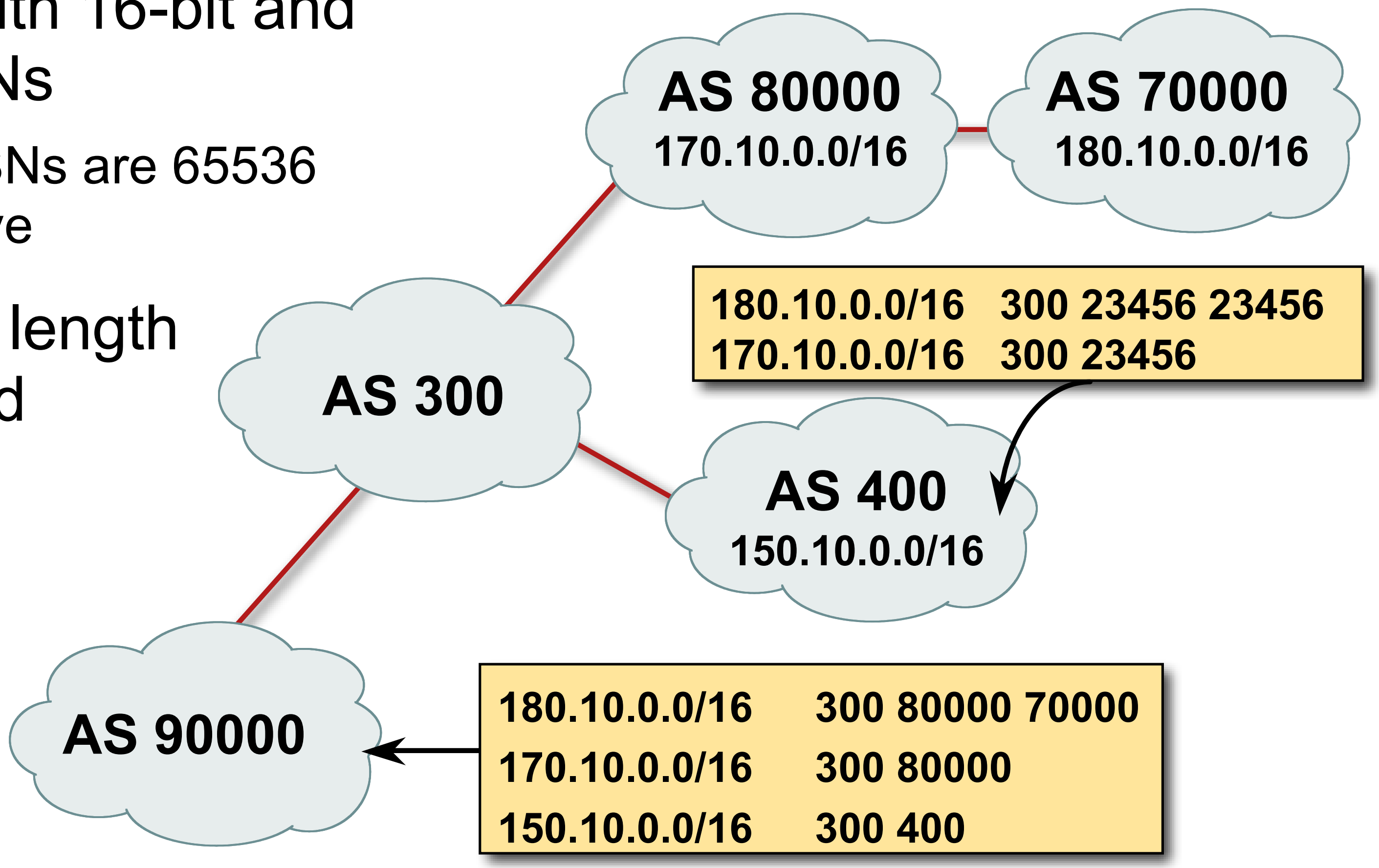
  Loop detection

  Applying policy

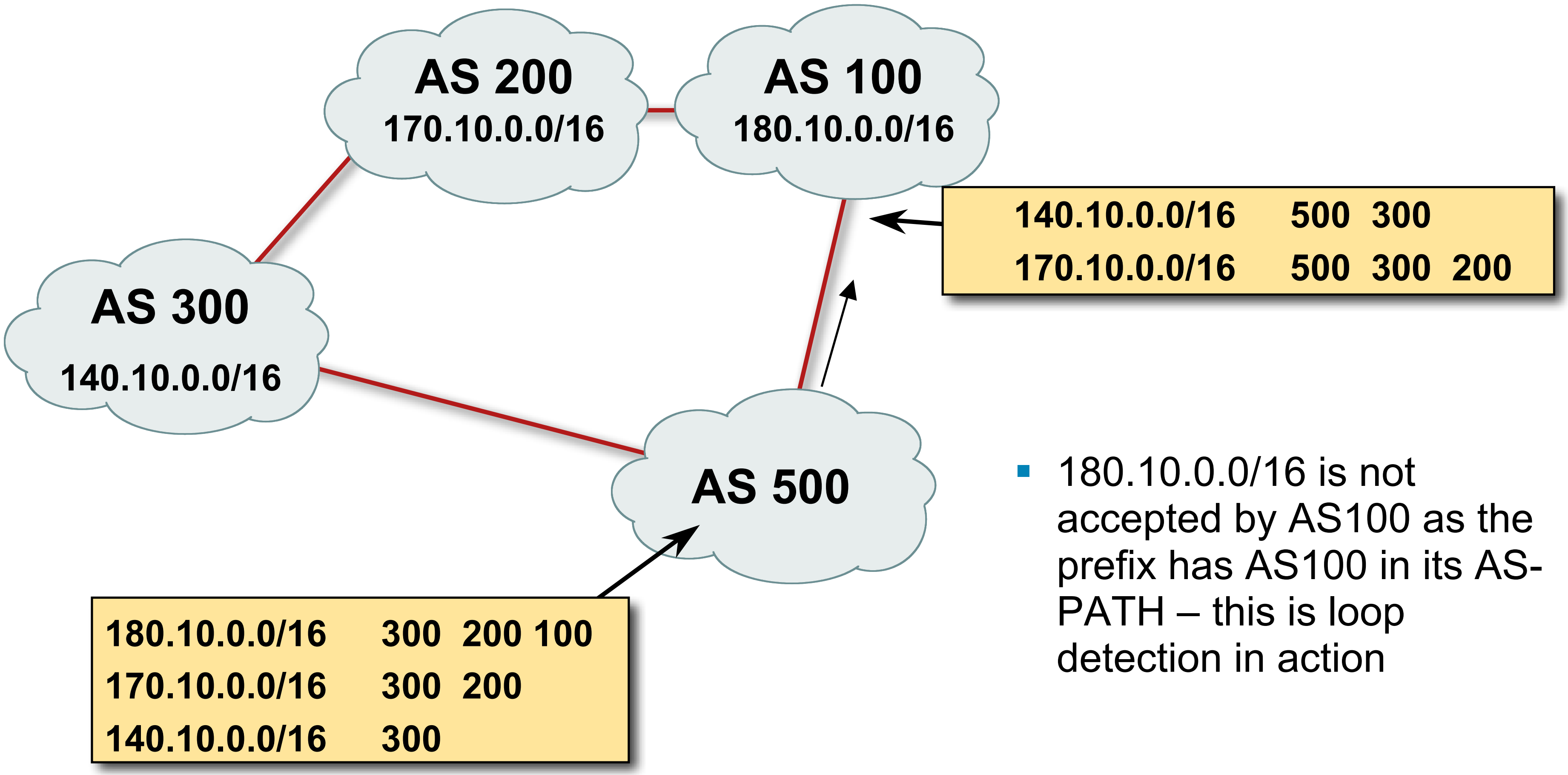


AS-Path (with 16 and 32-bit ASNs)

- Internet with 16-bit and 32-bit ASNs
  - 32-bit ASNs are 65536 and above
- AS-PATH length maintained

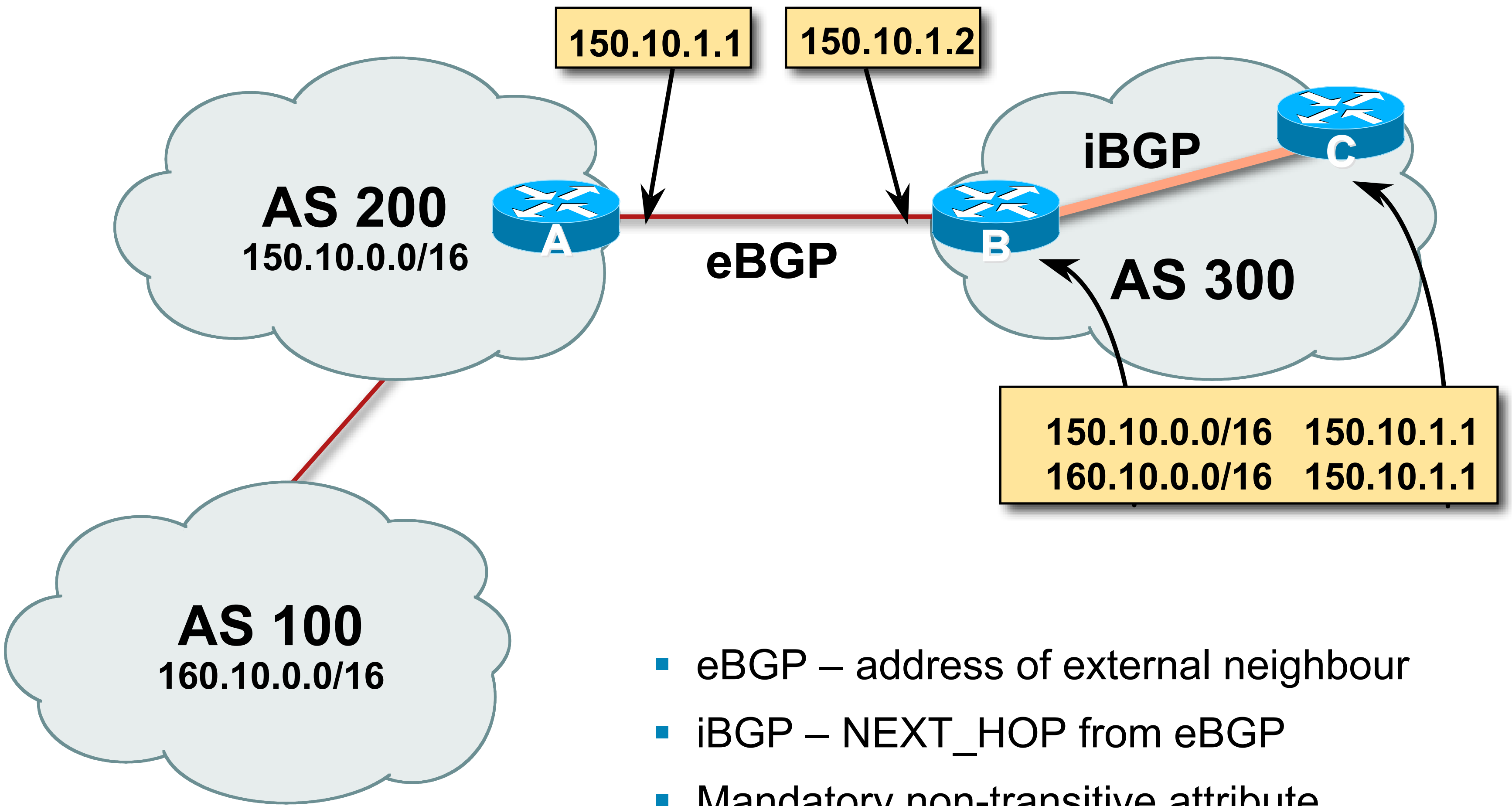


AS-Path loop detection



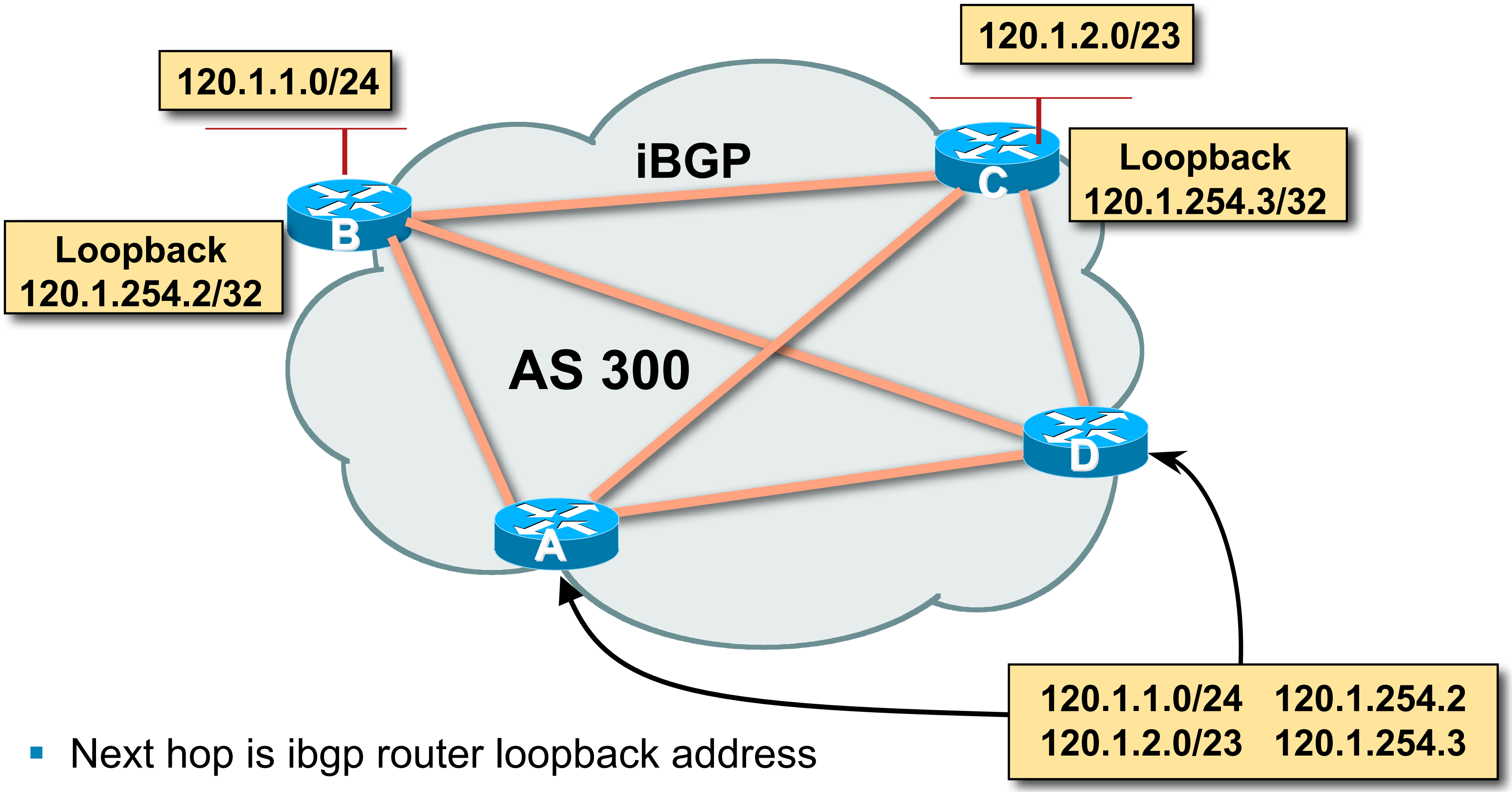
- 180.10.0.0/16 is not accepted by AS100 as the prefix has AS100 in its AS-PATH – this is loop detection in action

Next Hop



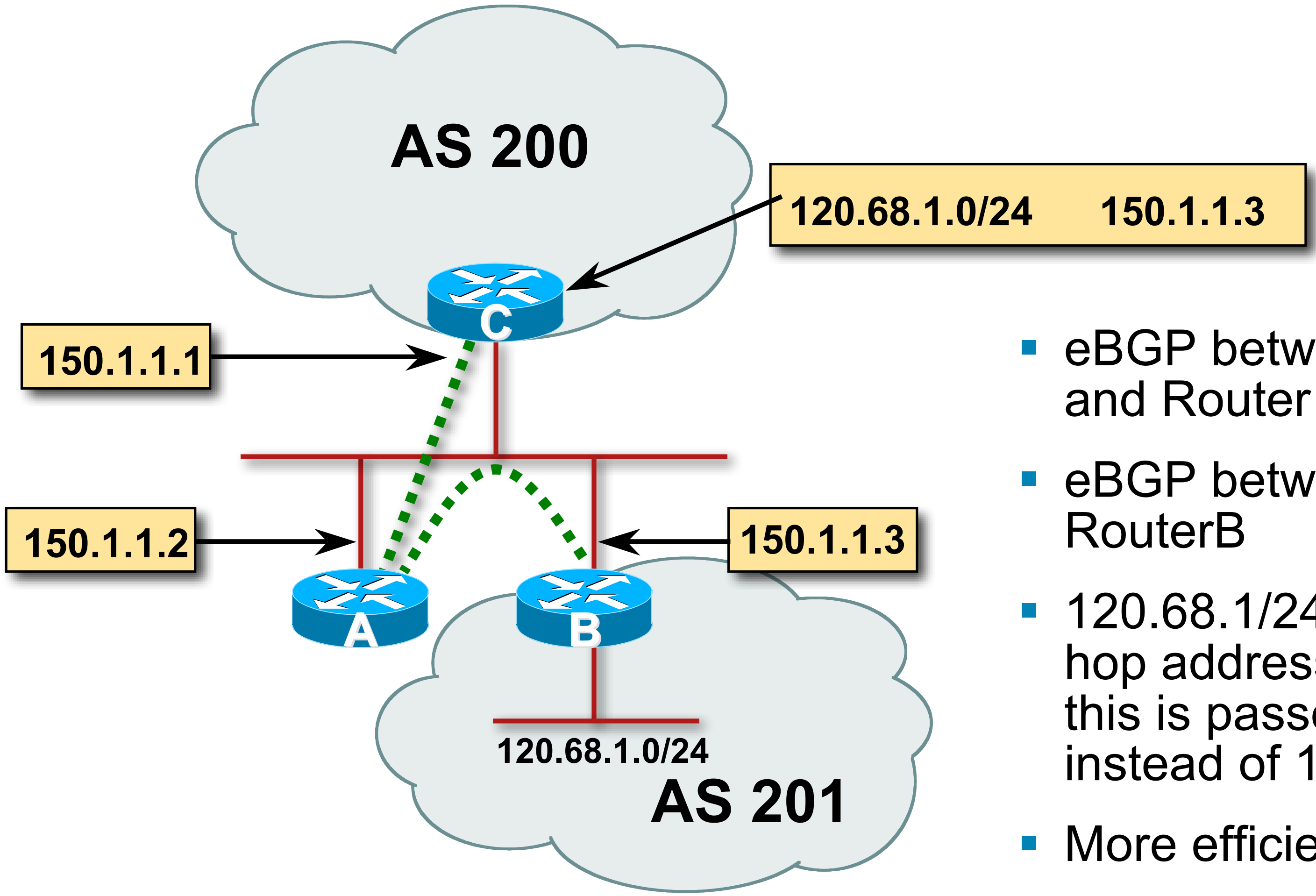
- eBGP – address of external neighbour
- iBGP – NEXT_HOP from eBGP
- Mandatory non-transitive attribute

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

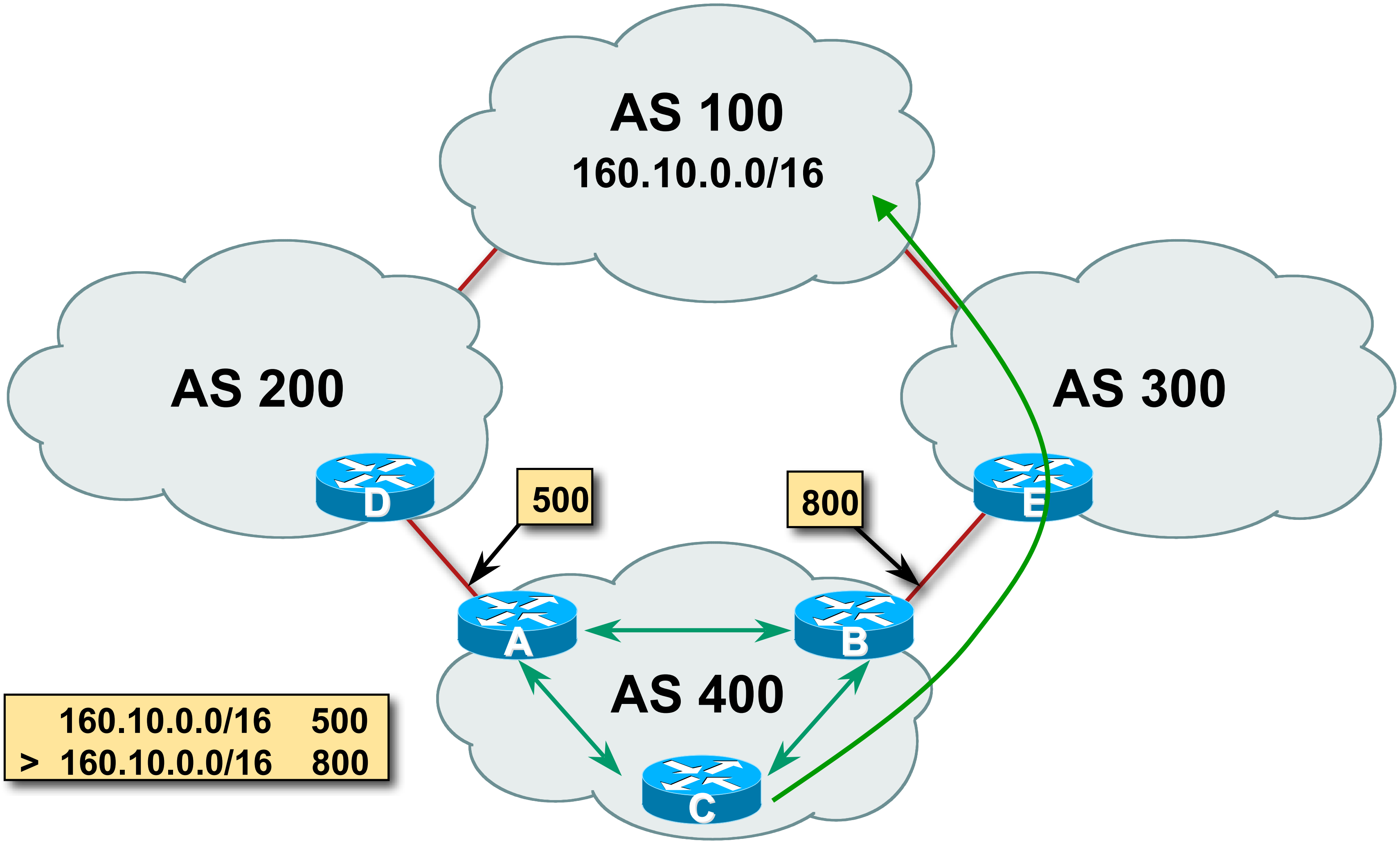
Origin

- 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



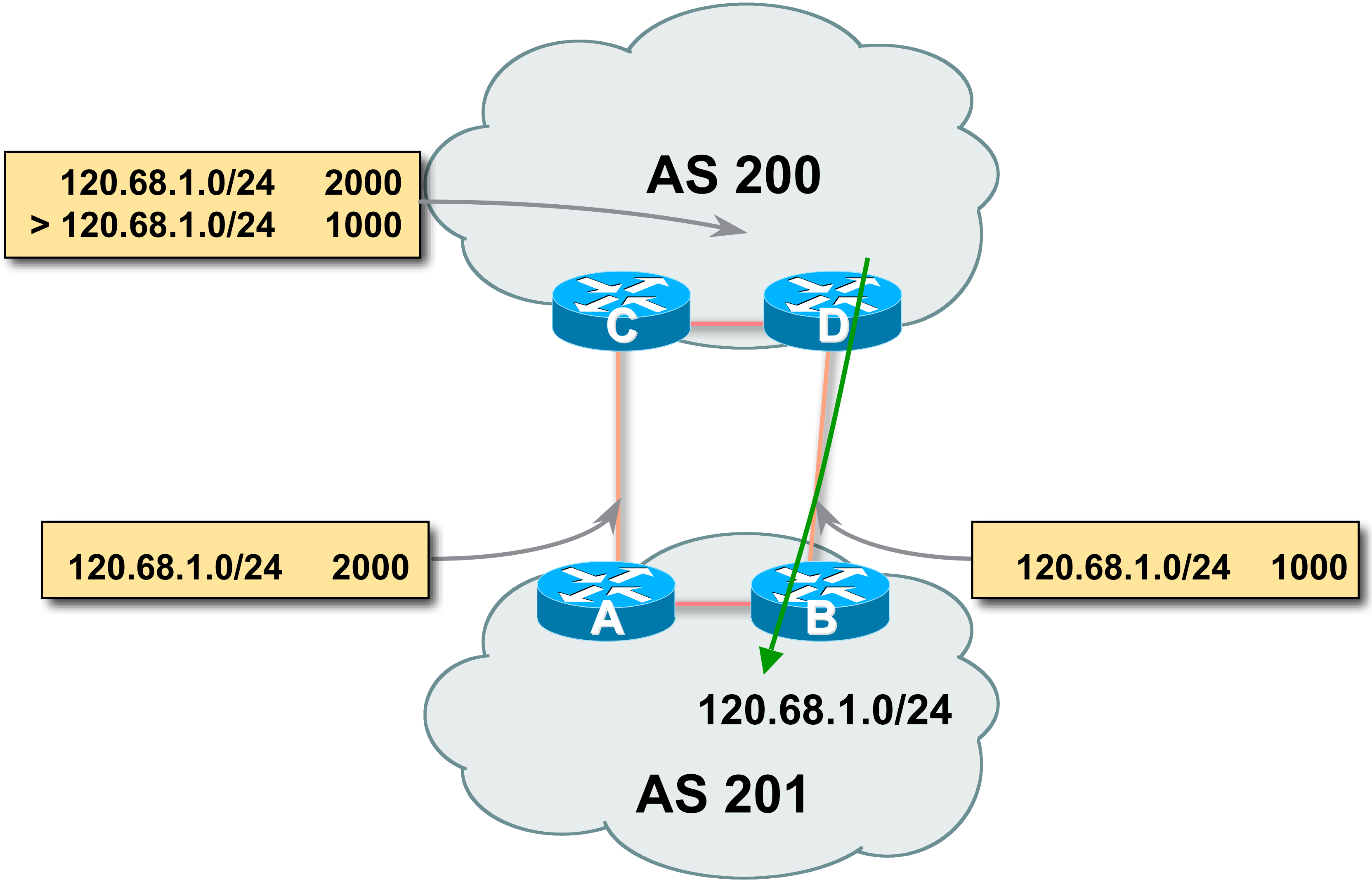
Local Preference

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

  Default local preference is 100 (Cisco 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)

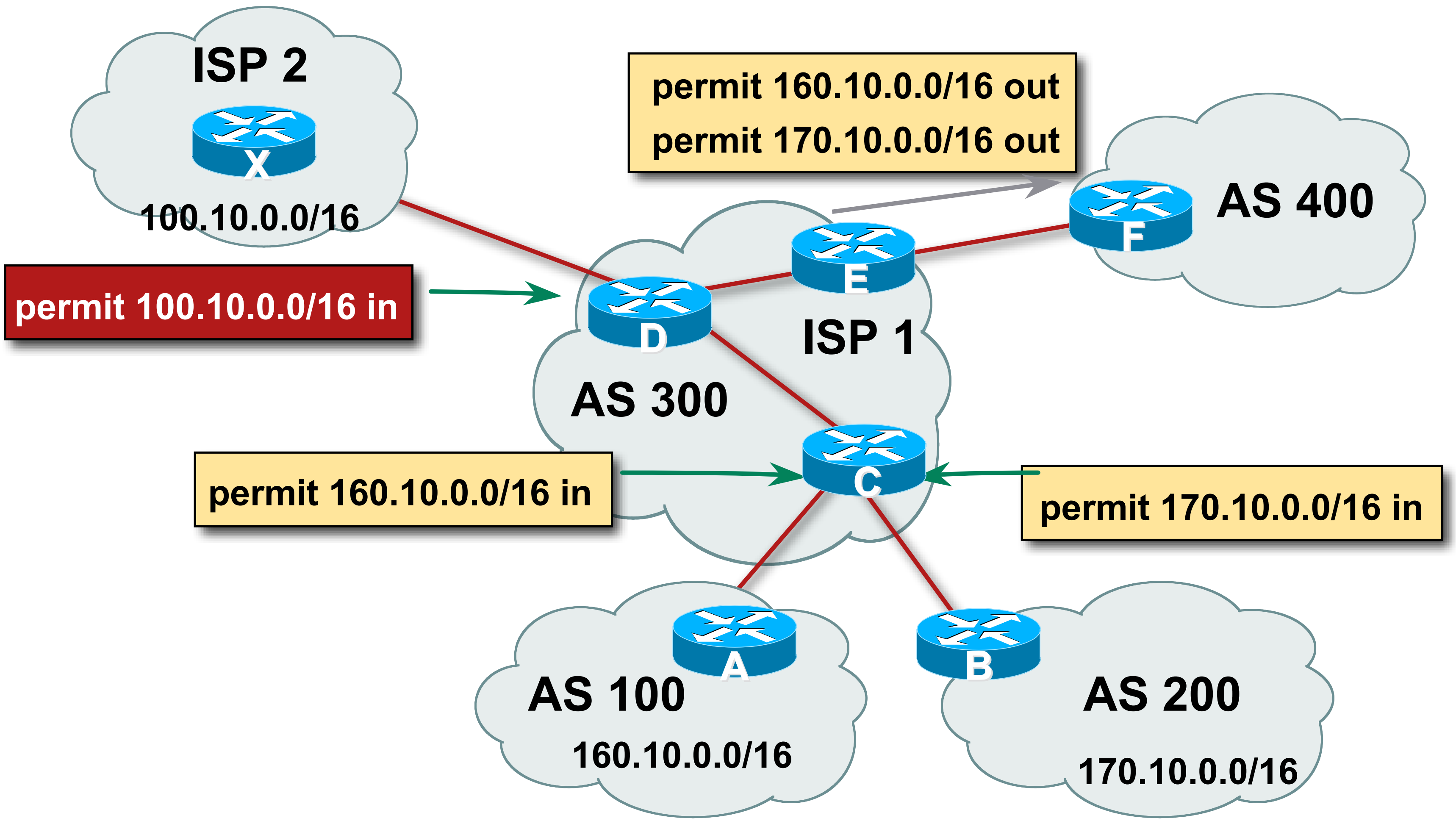
Multi-Exit Discriminator
"metric confusion"

- 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 handled absence of metric as meaning a metric of 0
  - Other implementations handled the absence of metric as meaning a metric of 2^{32}-1 (highest possible) or 2^{32}-2
  - Potential for "metric confusion"

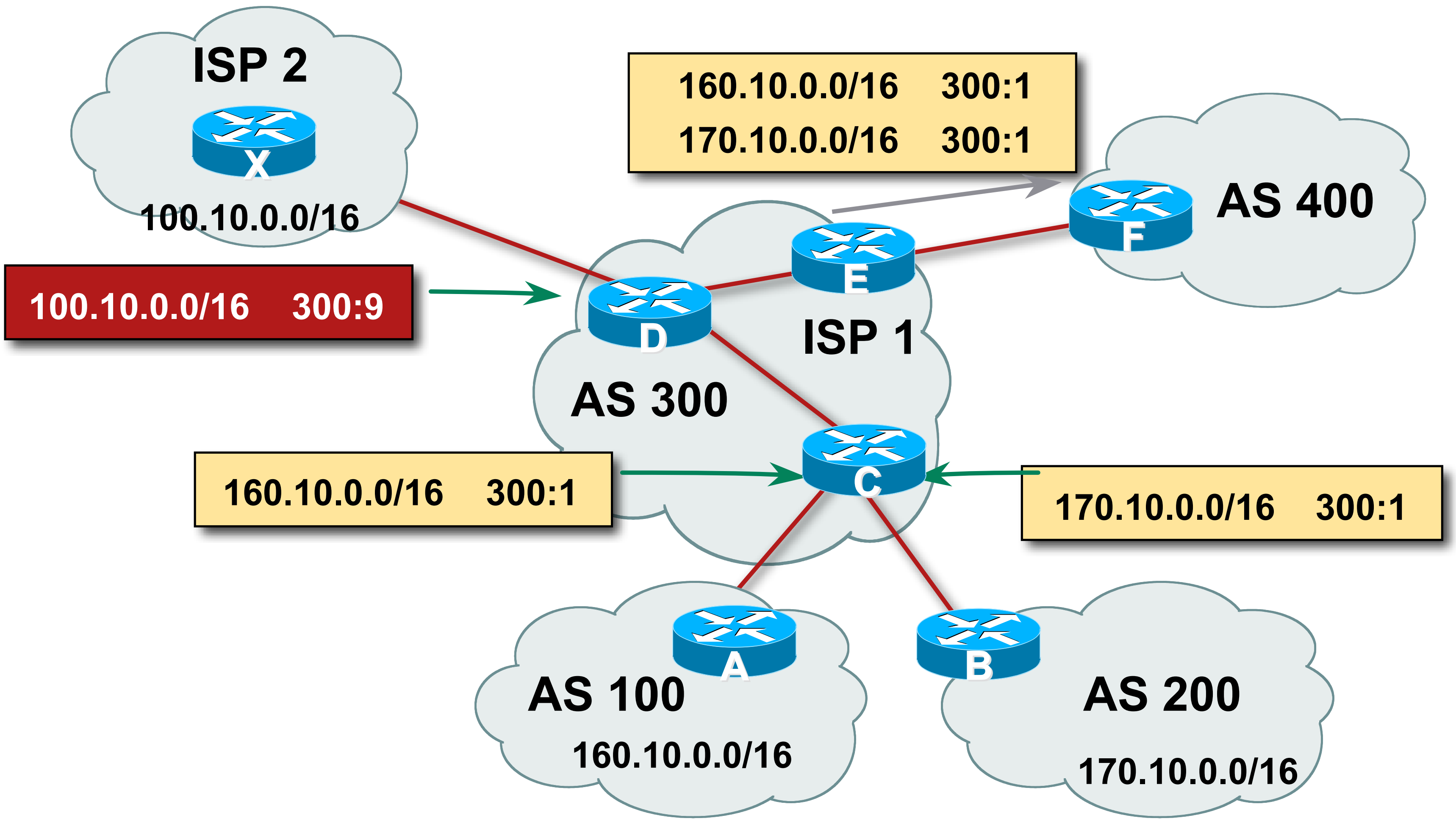
Community

- 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

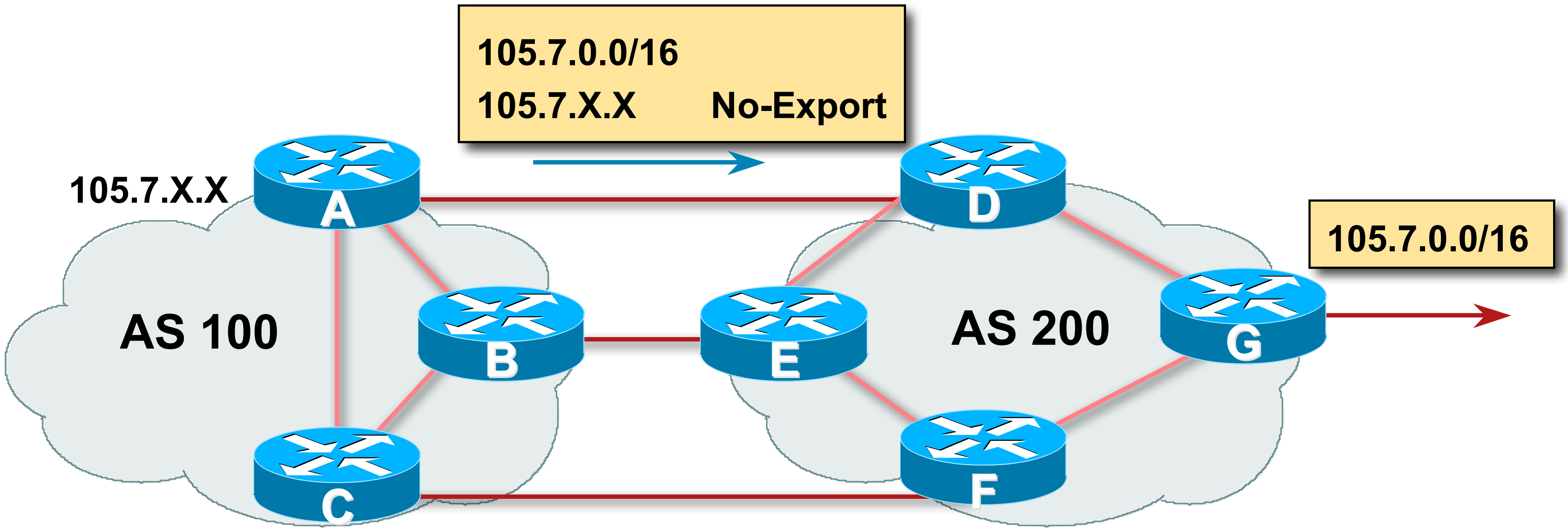
- no-export-subconfed 65535:65283

  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

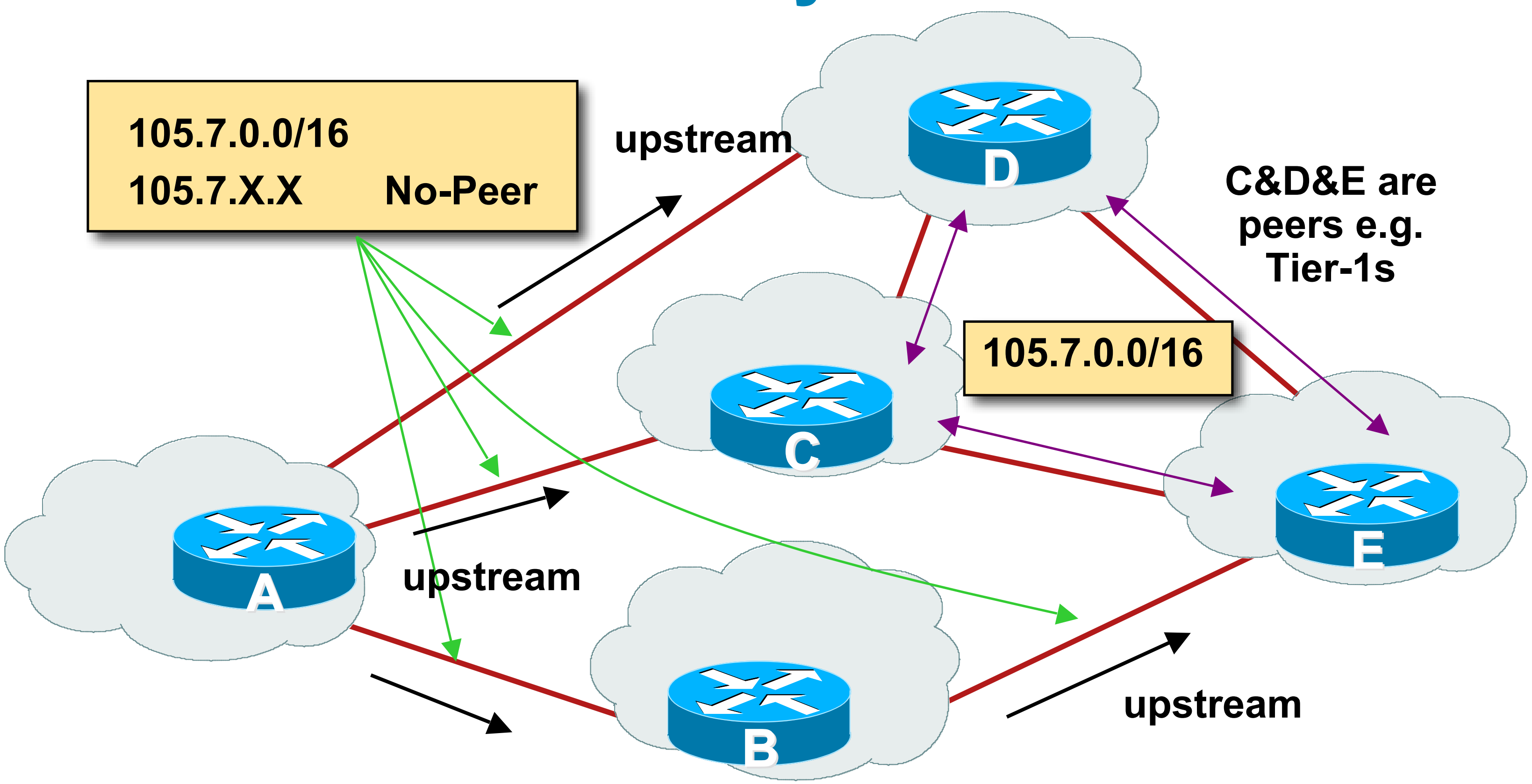


- AS100 announces aggregate and subprefixes

    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

What about 4-byte ASNs?

- Communities are widely used for encoding ISP routing policy
  - 32 bit attribute
- RFC1998 format is now "standard" practice
  - ***ASN:number***
- Fine for 2-byte ASNs, but 4-byte ASNs cannot be encoded
- Solutions:
  - Use "private ASN" for the first 16 bits
  - Wait for www.ietf.org/internet-drafts/draft-ietf-idr-as4octet-extcomm-generic-subtype-02.txt to be implemented

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]
5	Extended Next Hop Encoding	[RFC5549]
64	Graceful Restart Capability	[RFC4724]
65	Support for 4 octet ASNs	[RFC4893]
66	Deprecated	
67	Support for Dynamic Capability	[ID]
68	Multisession BGP	[ID]
69	Add Path Capability	[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

- Original BGP specification and implementation was fine for the Internet of the early 1990s
  - But didn't scale
- Issues as the Internet grew included:
  - Scaling 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

- Current Best Practice Scaling Techniques
  - Route Refresh
  - Peer-groups
  - Route Reflectors (and Confederations)
- Deploying 4-byte ASNs
- Deprecated Scaling Techniques
  - Route Flap Damping

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

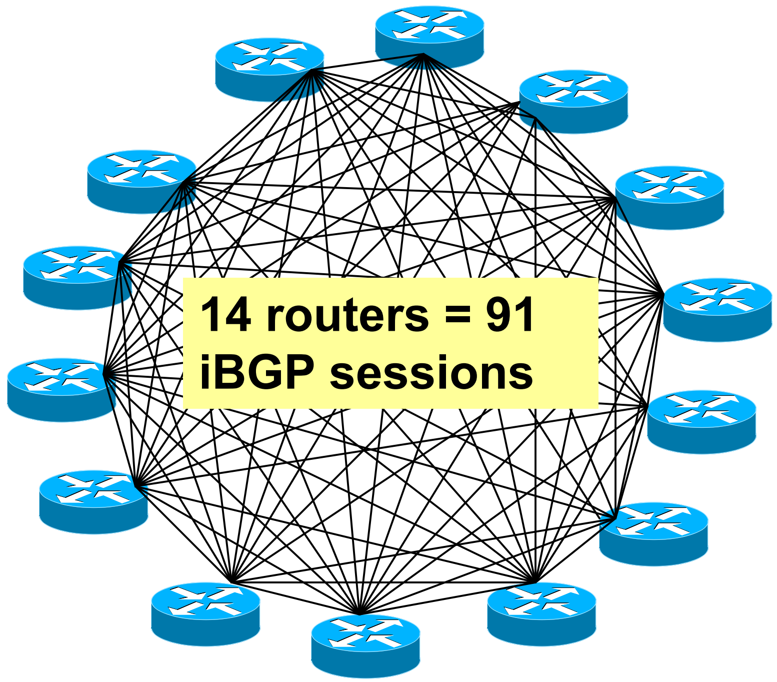
Route Reflectors

Scaling the iBGP mesh

Scaling iBGP mesh

- Avoid ½n(n-1) iBGP mesh

n=1000 ⇒ nearly half a million ibgp sessions!

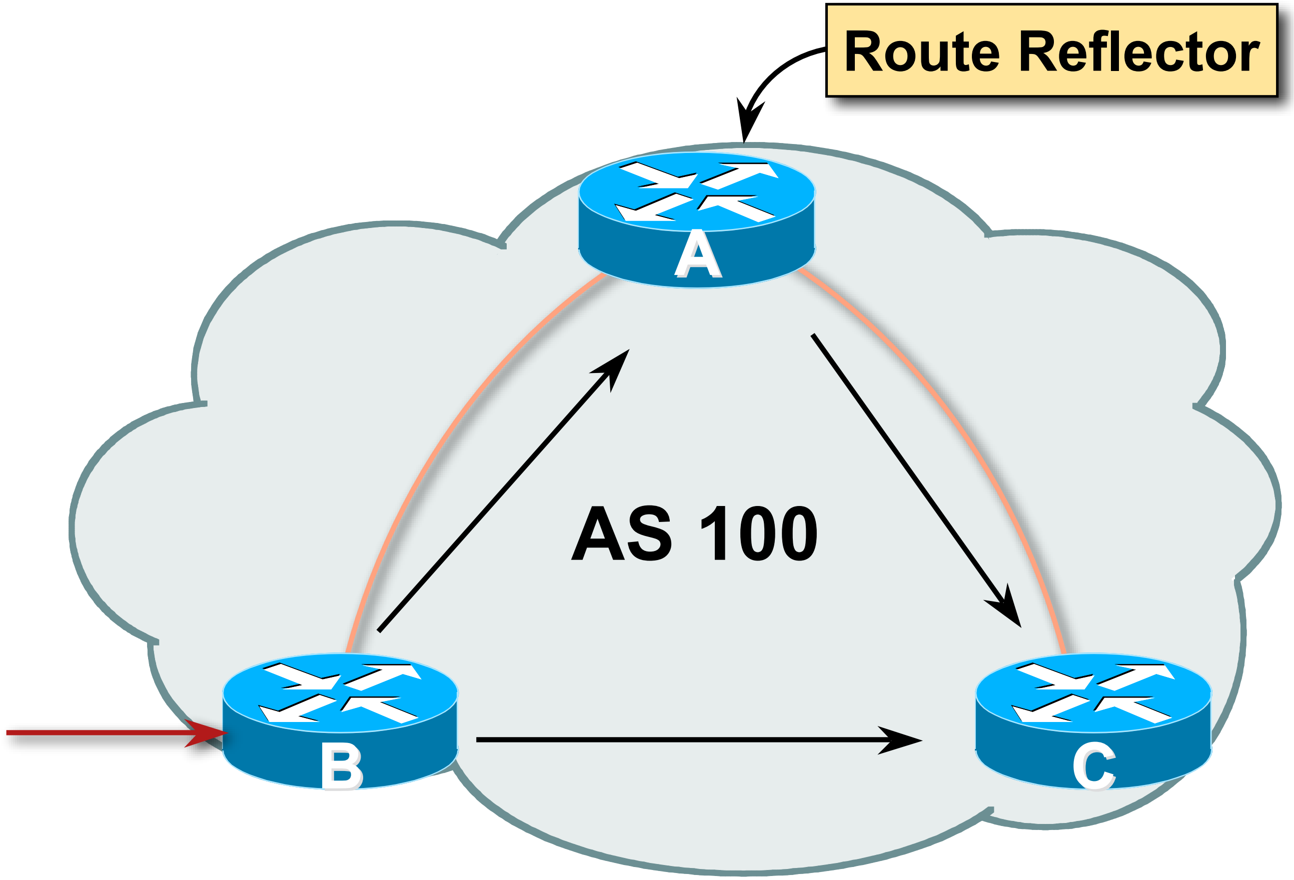


- Two solutions

  Route reflector – simpler to deploy and run

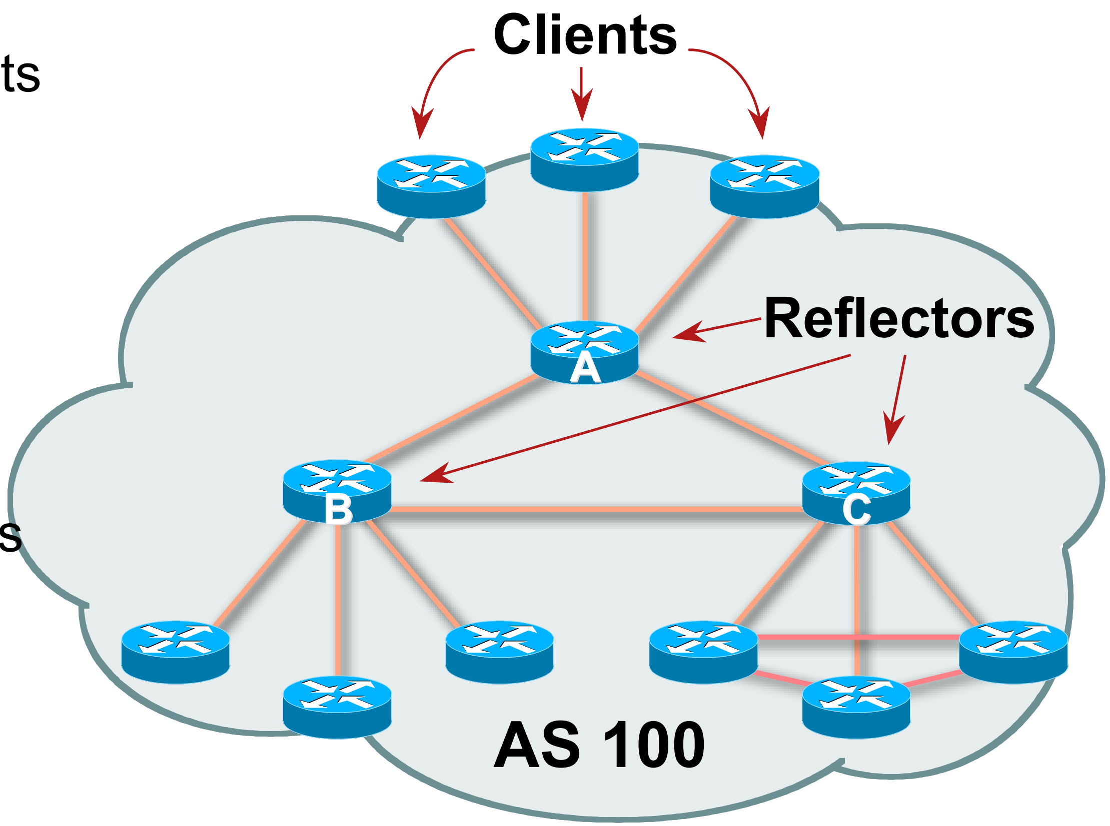
  Confederation – more complex, has corner case advantages

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!

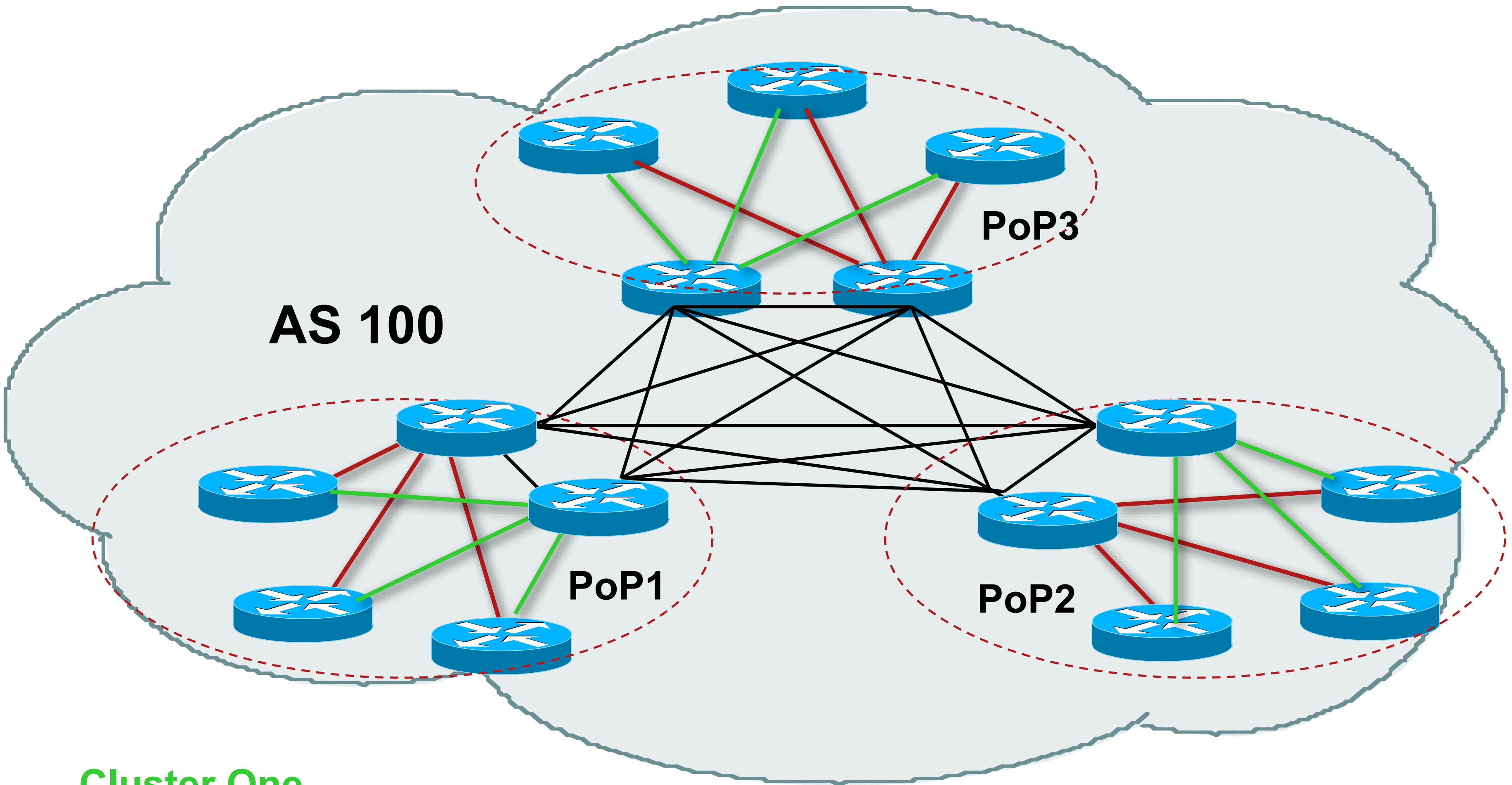
    All RRs in the cluster must have the same cluster-id (otherwise it is a different cluster)

- 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 Reflectors: 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

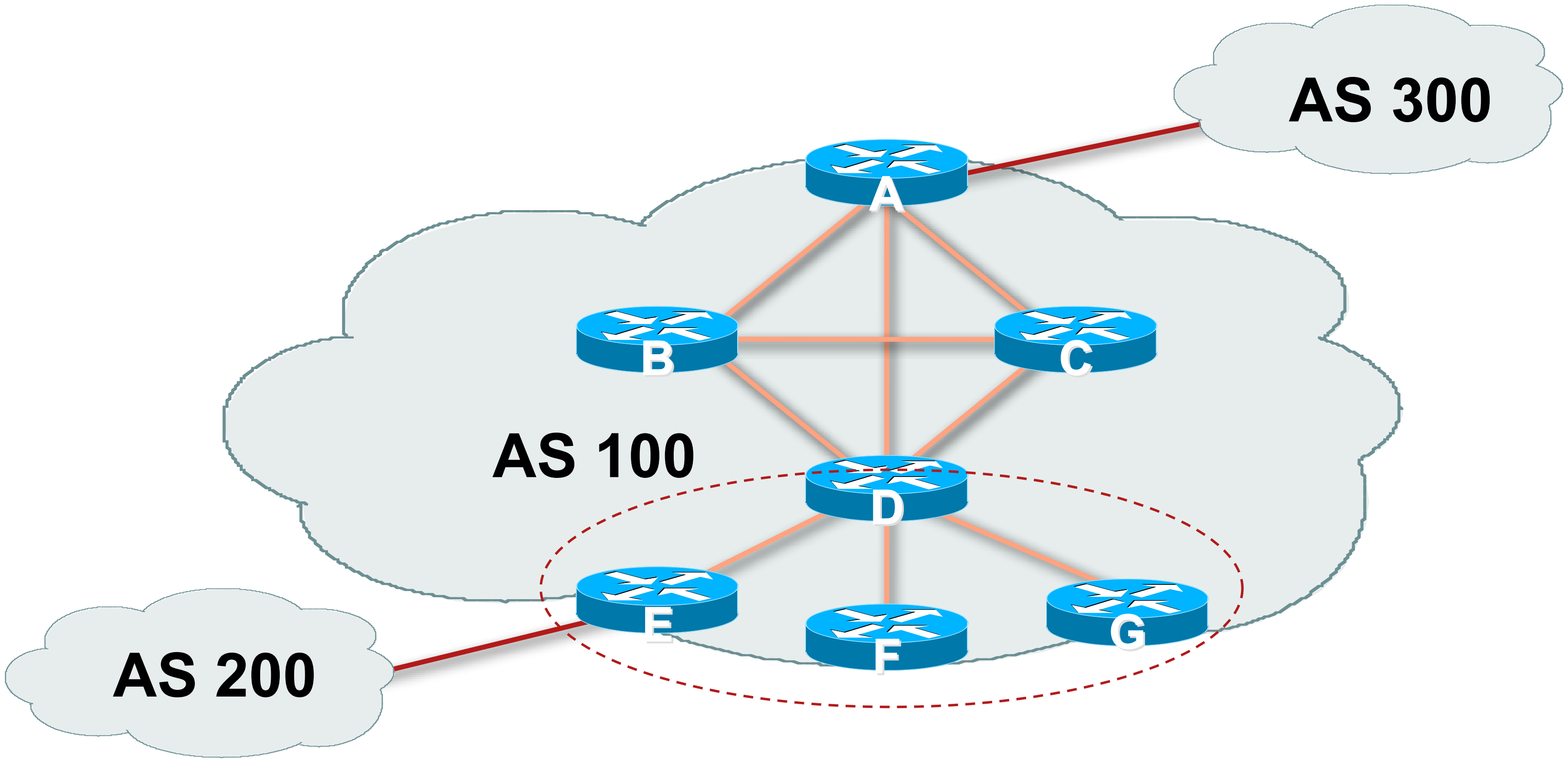
Route Reflector: Deployment

- 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

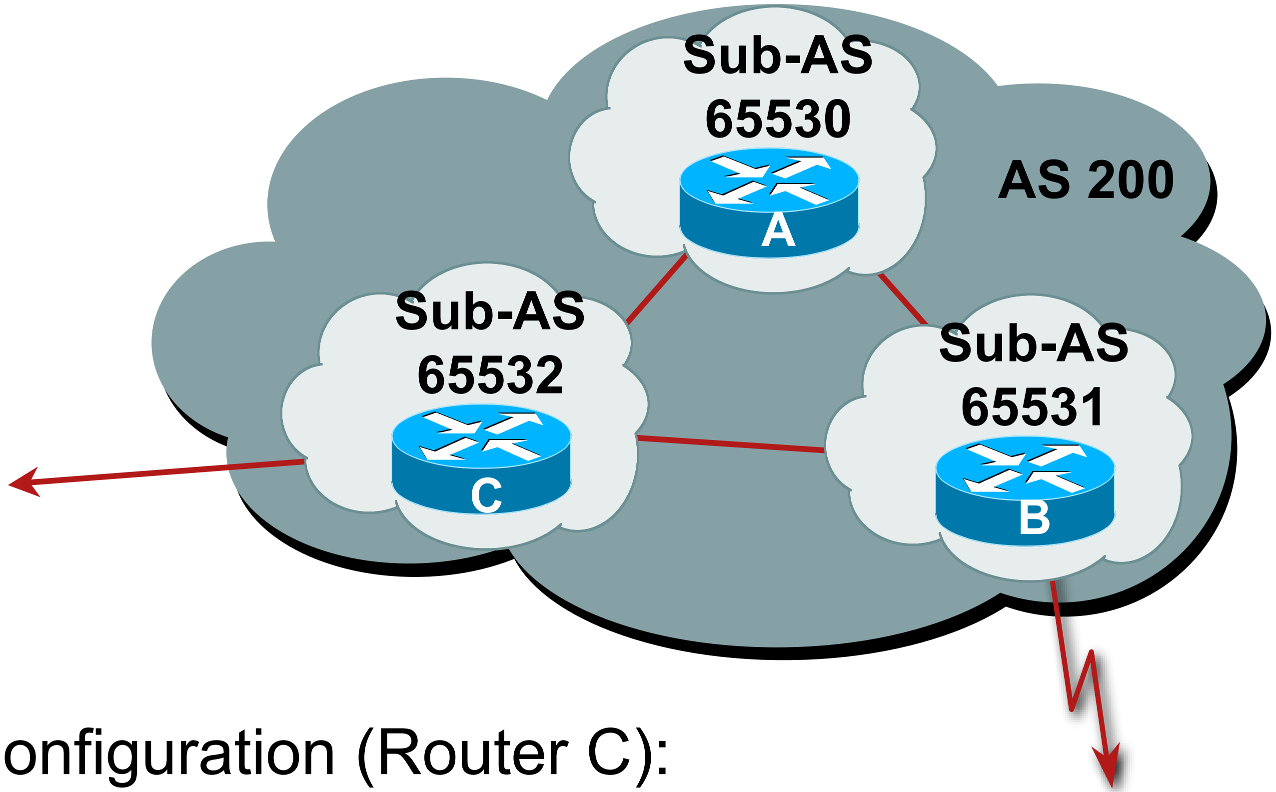
Confederations

- 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

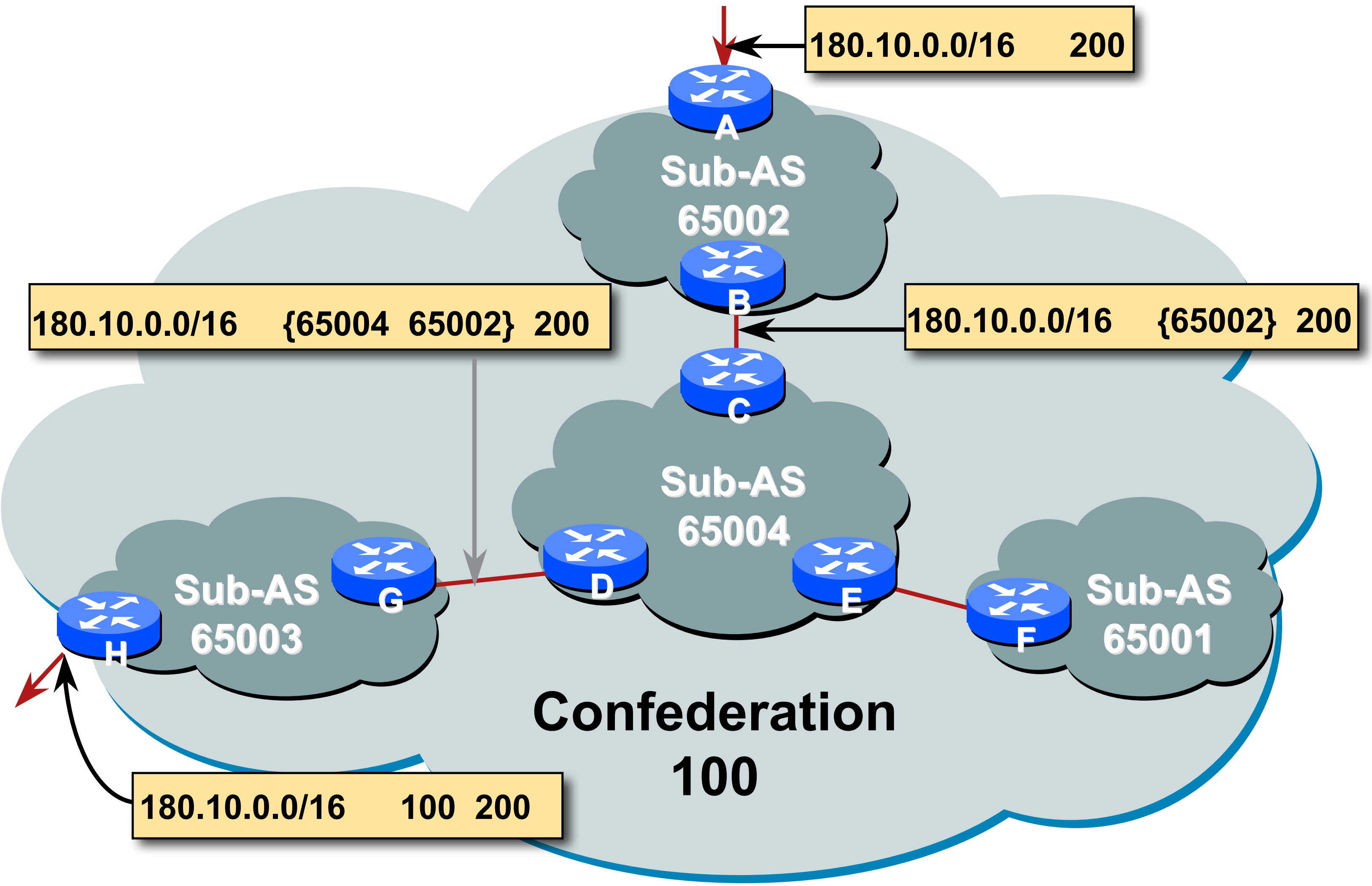
Confederations



- 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

Deploying 32-bit ASNs

How to support customers using the extended ASN range

32-bit ASNs

- Standards documents

  Description of 32-bit ASNs

  www.rfc-editor.org/rfc/rfc4893.txt

  Textual representation

  www.rfc-editor.org/rfc/rfc5396.txt

  New extended community

  www.rfc-editor.org/rfc/rfc5668.txt

- AS 23456 is reserved as interface between 16-bit and 32-bit ASN world

32-bit ASNs – terminology

- 16-bit ASNs

  Refers to the range 0 to 65535

- 32-bit ASNs

  Refers to the range 65536 to 4294967295

  (or the extended range)

- 32-bit ASN pool

  Refers to the range 0 to 4294967295

Getting a 32-bit ASN

- Sample RIR policy

  www.apnic.net/docs/policy/asn-policy.html

- From 1st January 2007

  32-bit ASNs were available on request

- From 1st January 2009

  32-bit ASNs were assigned by default

  16-bit ASNs were only available on request

- From 1st January 2010

  No distinction – ASNs assigned from the 32-bit pool

Representation

- Representation of 0-4294967295 ASN range
  - Most operators favour traditional format (asplain)
  - A few prefer dot notation (X.Y):
    - asdot for 65536-4294967295, e.g 2.4
    - asdot+ for 0-4294967295, e.g 0.64513
  - **But regular expressions will have to be completely rewritten for asdot and asdot+ !!!**
- For example:
  - ^[0-9]+$ matches any ASN (16-bit and asplain)
  - This and equivalents extensively used in BGP multihoming configurations for traffic engineering
- Equivalent regexp for asdot is: ^([0-9]+)|([0-9]+\.[0-9]+)$
- Equivalent regexp for asdot+ is: ^[0-9]+\.[0-9]+$

Changes

- 32-bit ASNs are backward compatible with 16-bit ASNs
- **There is no flag day**
- You do NOT need to:

  Throw out your old routers

  Replace your 16-bit ASN with a 32-bit ASN
- You do need to be aware that:

  Your customers will come with 32-bit ASNs

  ASN 23456 is not a bogon!

  You will need a router supporting 32-bit ASNs to use a 32-bit ASN locally
- If you have a proper BGP implementation, 32-bit ASNs will be transported silently across your network

How does it work?

- If local router and remote router supports configuration of 32-bit ASNs

  BGP peering is configured as normal using the 32-bit ASN

- If local router and remote router does not support configuration of 32-bit ASNs

  BGP peering can only use a 16-bit ASN

- If local router only supports 16-bit ASN and remote router/network has a 32-bit ASN

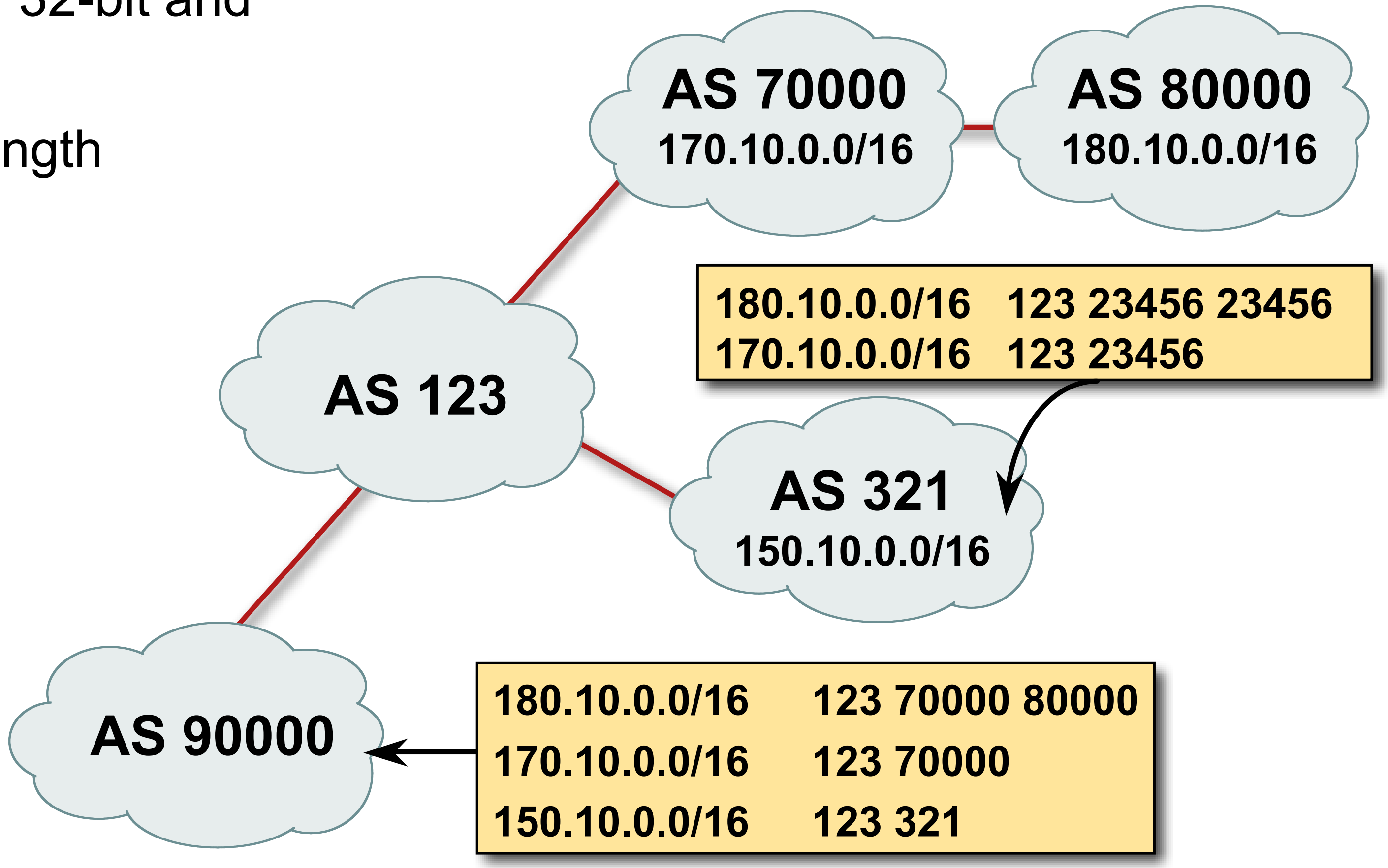
  Compatibility mode is initiated…

Compatibility Mode:

- Local router only supports 16-bit ASN and remote router uses 32-bit ASN
- BGP peering initiated:
  - Remote asks local if 32-bit supported (BGP capability negotiation)
  - When local says “no”, remote then presents AS23456
  - Local needs to be configured to peer with remote using AS23456
- BGP peering initiated (cont):
  - BGP session established using AS23456
  - 32-bit ASN included in a new BGP attribute called AS4_PATH
    - (as opposed to AS_PATH for 16-bit ASNs)
- Result:
  - 16-bit ASN world sees 16-bit ASNs and 23456 standing in for 32-bit ASNs
  - 32-bit ASN world sees 16 and 32-bit ASNs

Example:

- Internet with 32-bit and 16-bit ASNs
- AS-PATH length maintained



What has changed?

- Two new BGP attributes:
  - AS4_PATH
    - Carries 32-bit ASN path info
  - AS4_AGGREGATOR
    - Carries 32-bit ASN aggregator info
  - Well-behaved BGP implementations will simply pass these along if they don't understand them
- AS23456 (AS_TRANS)

What do they look like?

- IPv4 prefix originated by AS196613

asplain format

```
as4-7200#sh ip bgp 145.125.0.0/20
BGP routing table entry for 145.125.0.0/20, version 58734
Paths: (1 available, best #1, table default)
  131072 12654 196613
    204.69.200.25 from 204.69.200.25 (204.69.200.25)
      Origin IGP, localpref 100, valid, internal, best
```

- IPv4 prefix originated by AS3.5

asdot format

```
as4-7200#sh ip bgp 145.125.0.0/20
BGP routing table entry for 145.125.0.0/20, version 58734
Paths: (1 available, best #1, table default)
  2.0 12654 3.5
    204.69.200.25 from 204.69.200.25 (204.69.200.25)
      Origin IGP, localpref 100, valid, internal, best
```

What do they look like?

- IPv4 prefix originated by AS196613

  But 16-bit AS world view:

```
BGP-view1>sh ip bgp 145.125.0.0/20
BGP routing table entry for 145.125.0.0/20, version 113382
Paths: (1 available, best #1, table Default-IP-Routing-Table)
23456 12654 23456
    204.69.200.25 from 204.69.200.25 (204.69.200.25)
      Origin IGP, localpref 100, valid, external, best
```

Transition AS

If 32-bit ASN not supported:

- Inability to distinguish between peer ASes using 32-bit ASNs
  - They will all be represented by AS23456
  - Could be problematic for transit provider's policy
- Inability to distinguish prefix's origin AS
  - How to tell whether origin is real or fake?
  - The real and fake both represented by AS23456
  - (There should be a better solution here!)
- Incorrect NetFlow summaries:
  - Prefixes from 32-bit ASNs will all be summarised under AS23456
  - Traffic statistics need to be measured per prefix and aggregated
  - Makes it hard to determine peerability of a neighbouring network

Implementations (Feb 2011)

- Cisco IOS-XR 3.4 onwards
- Cisco IOS-XE 2.3 onwards
- Cisco IOS 12.0(32)S12, 12.4(24)T, 12.2SRE, 12.2(33)SXI1 onwards
- Cisco NX-OS 4.0(1) onwards
- Quagga 0.99.10 (patches for 0.99.6)
- OpenBGPd 4.2 (patches for 3.9 & 4.0)
- Juniper JunOSe 4.1.0 & JunOS 9.1 onwards
- Redback SEOS
- Force10 FTOS7.7.1 onwards

http://as4.cluepon.net/index.php/Software_Support for a complete list

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

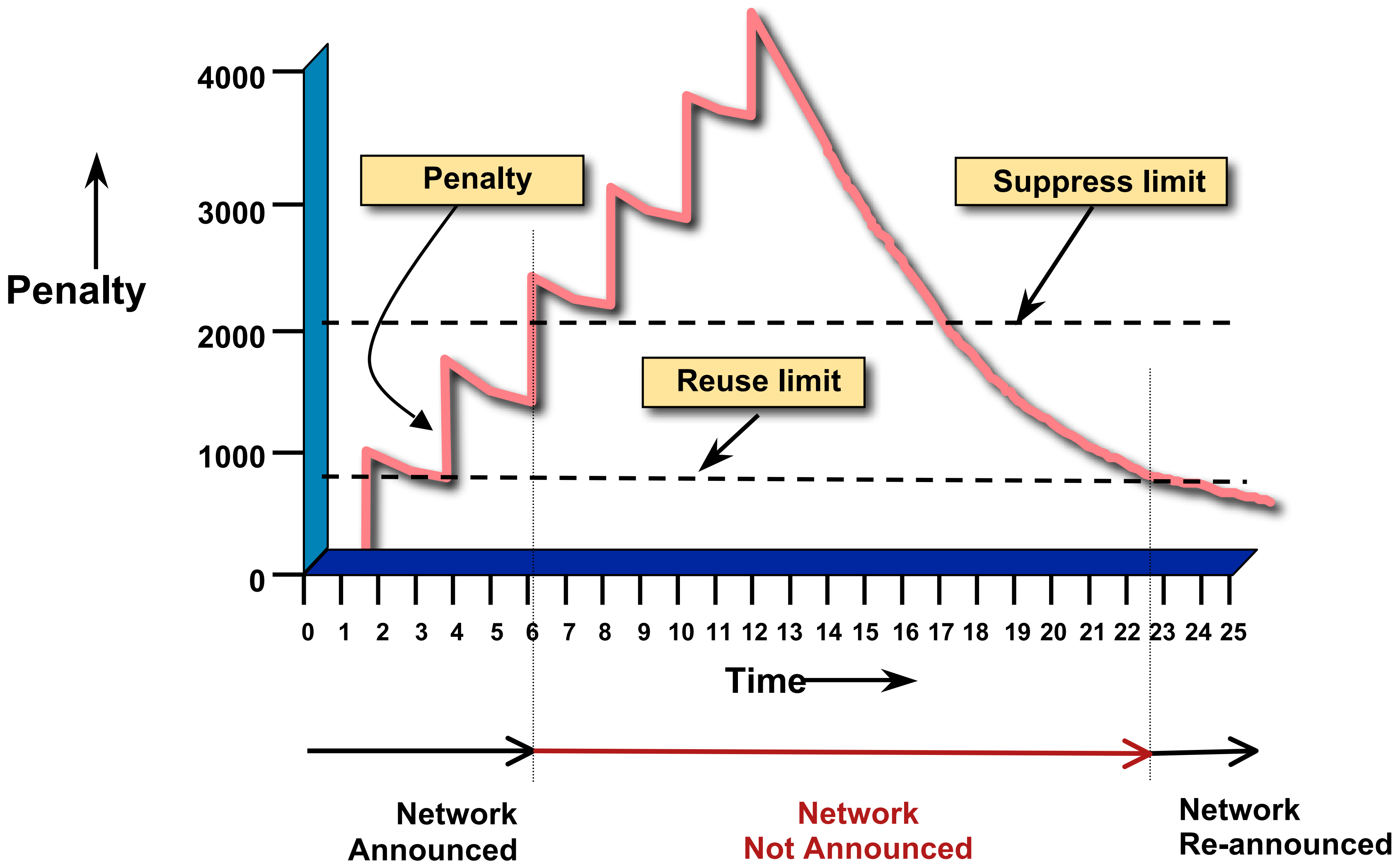
Route Flap Damping (continued)

- 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

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 difference 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

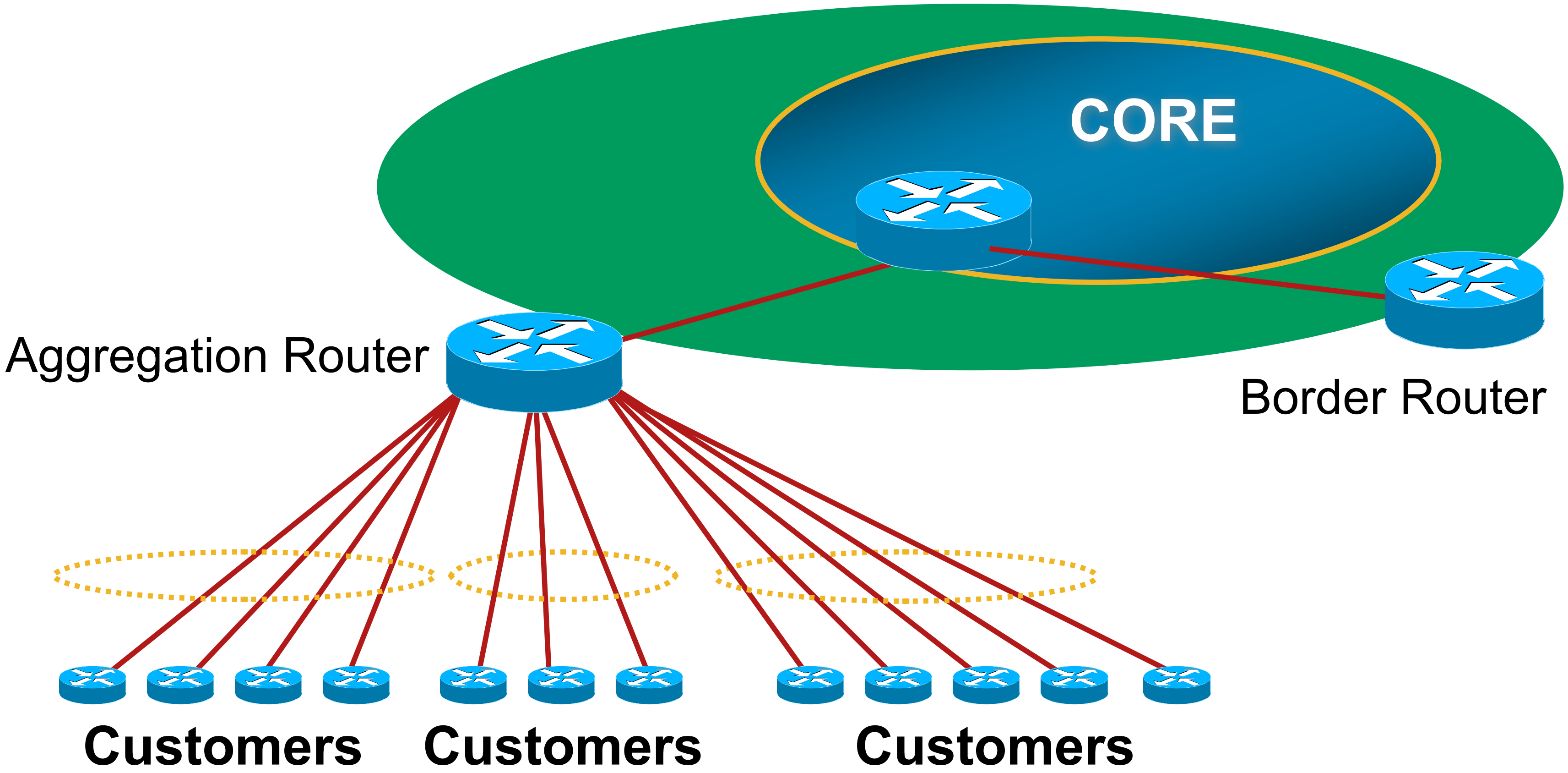
- Community assignments:

  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

ISP Examples: Sprint

IP/MPLS Products from Sprint

https://www.sprint.net/index.php?p=policy_bgp

within 3 business days of receipt of the request.

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.

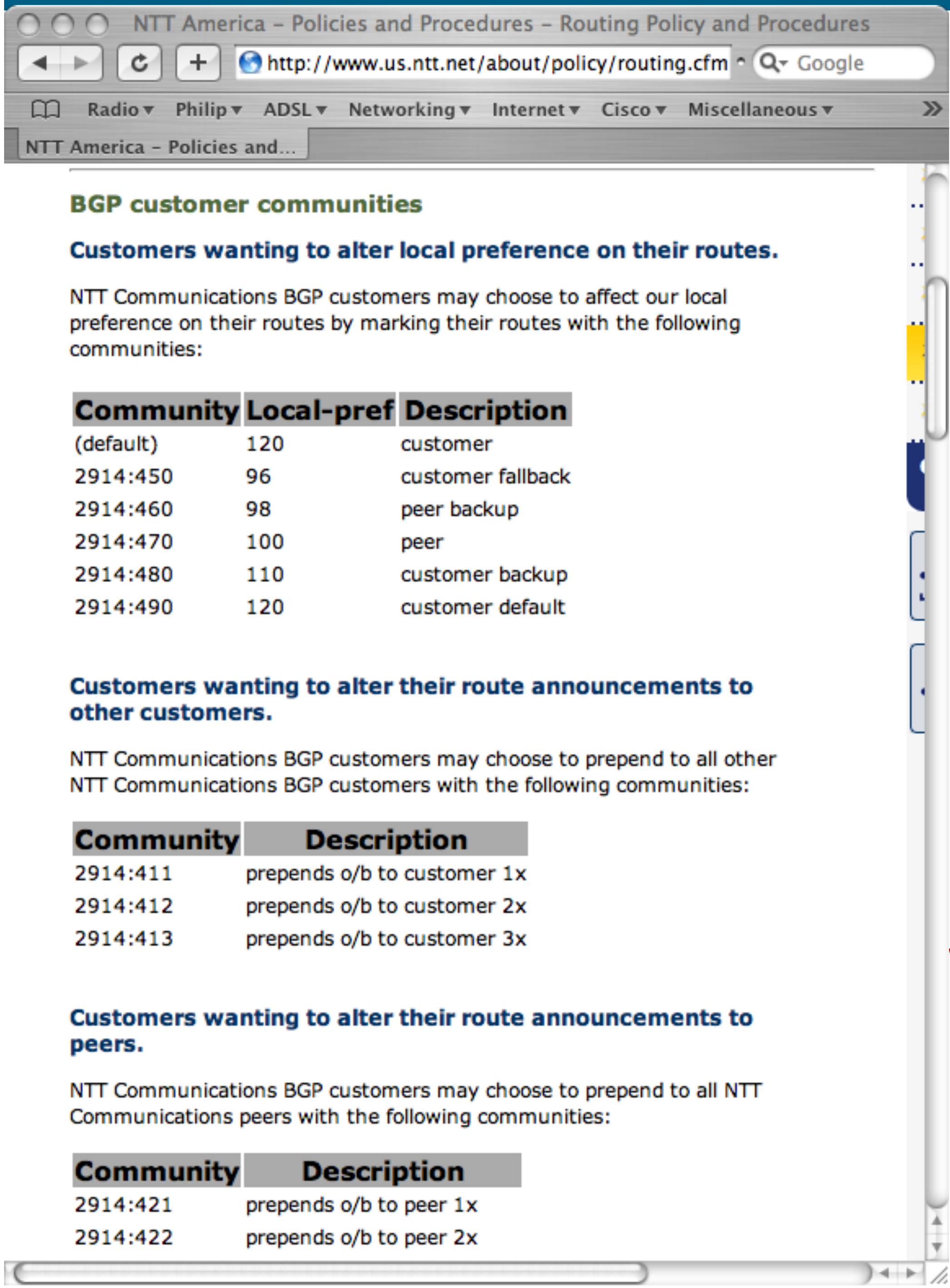
String	Resulting AS Path to ASXXX
65000:XXX	Do not advertise to ASXXX
65001:XXX	1239 (default) ...
65002:XXX	1239 1239 ...
65003:XXX	1239 1239 1239 ...
65004:XXX	1239 1239 1239 1239 ...

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 ...
65054:XXX	1239 1239 1239 1239 ...

More info at
https://www.sprint.net/index.php?p=policy_bgp

Some ISP Examples: NTT



NTT America – Policies and Procedures – Routing Policy and Procedures

http://www.us.ntt.net/about/policy/routing.cfm

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

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)
```

ISP Examples: Verizon Business 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

```
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

```
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

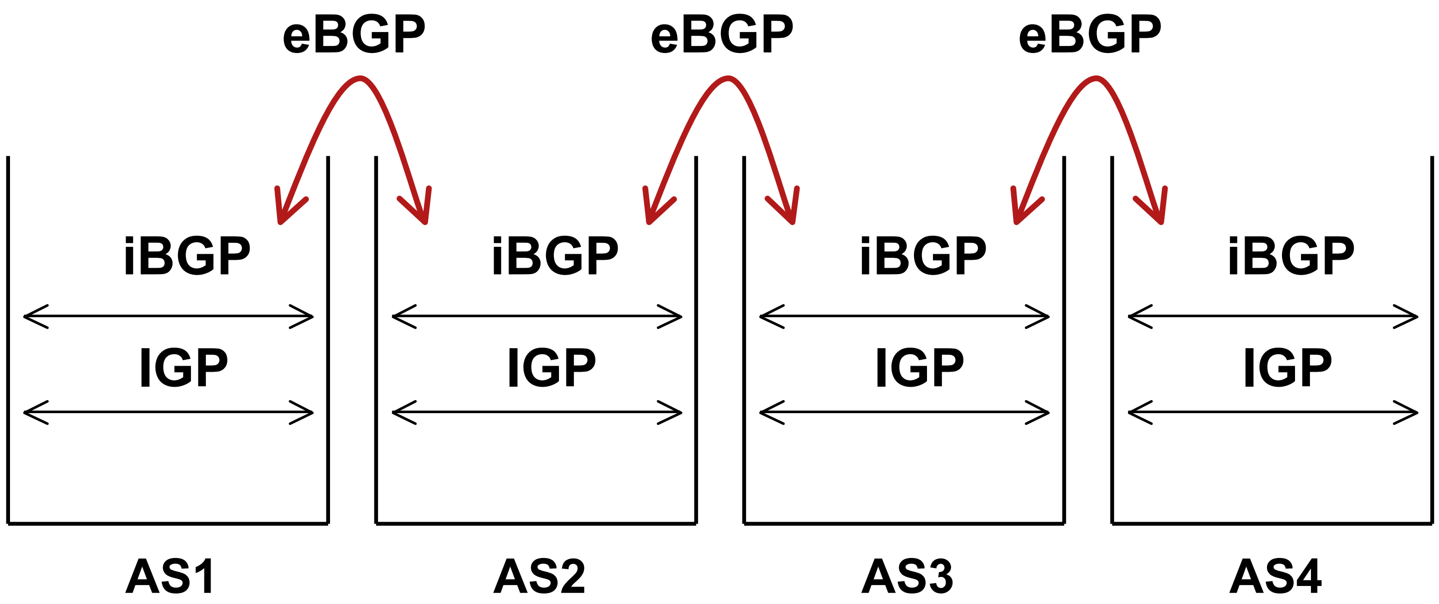
  design goal is to minimise number of prefixes in IGP to aid scalability and rapid convergence

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 for traffic engineering purposes

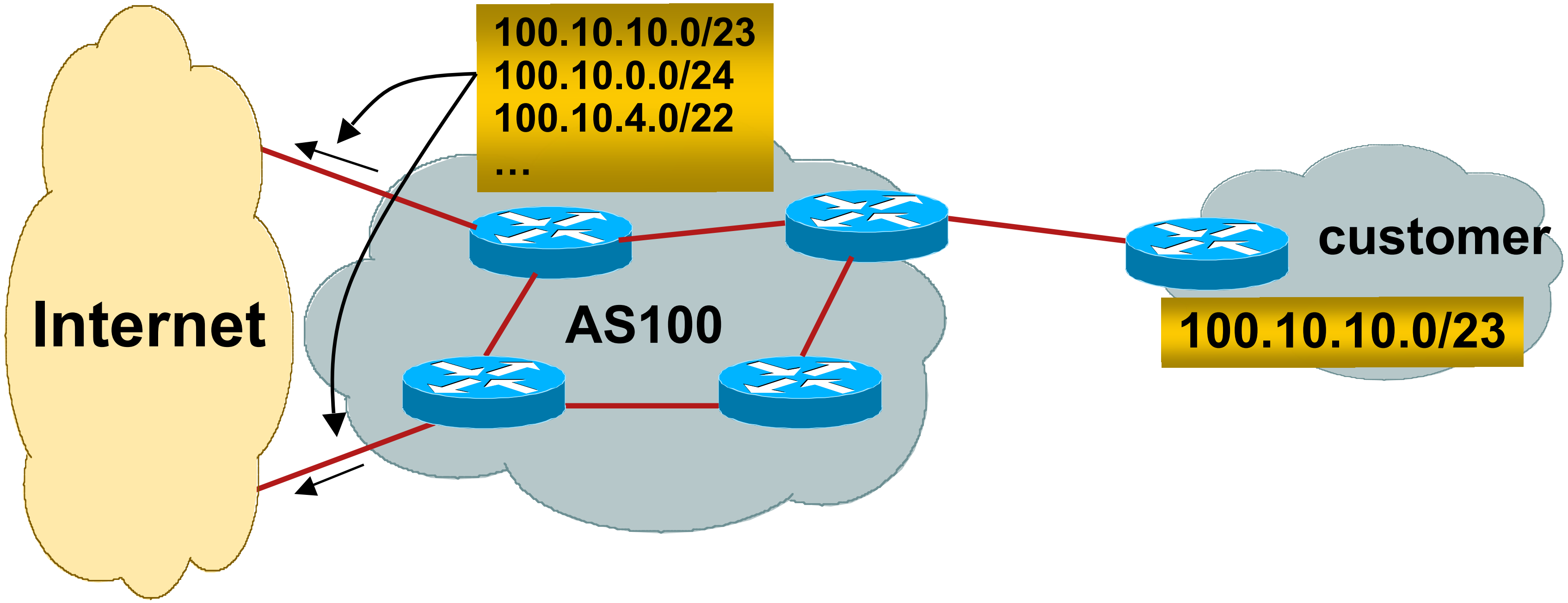
  (see BGP Multihoming Tutorial)
- Aggregate should be generated internally

  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 >180000 /24s!
- But: APNIC changed (Oct 2010) its minimum allocation size on all blocks to /24
  - IPv4 run-out is starting to have an impact

Aggregation – Example

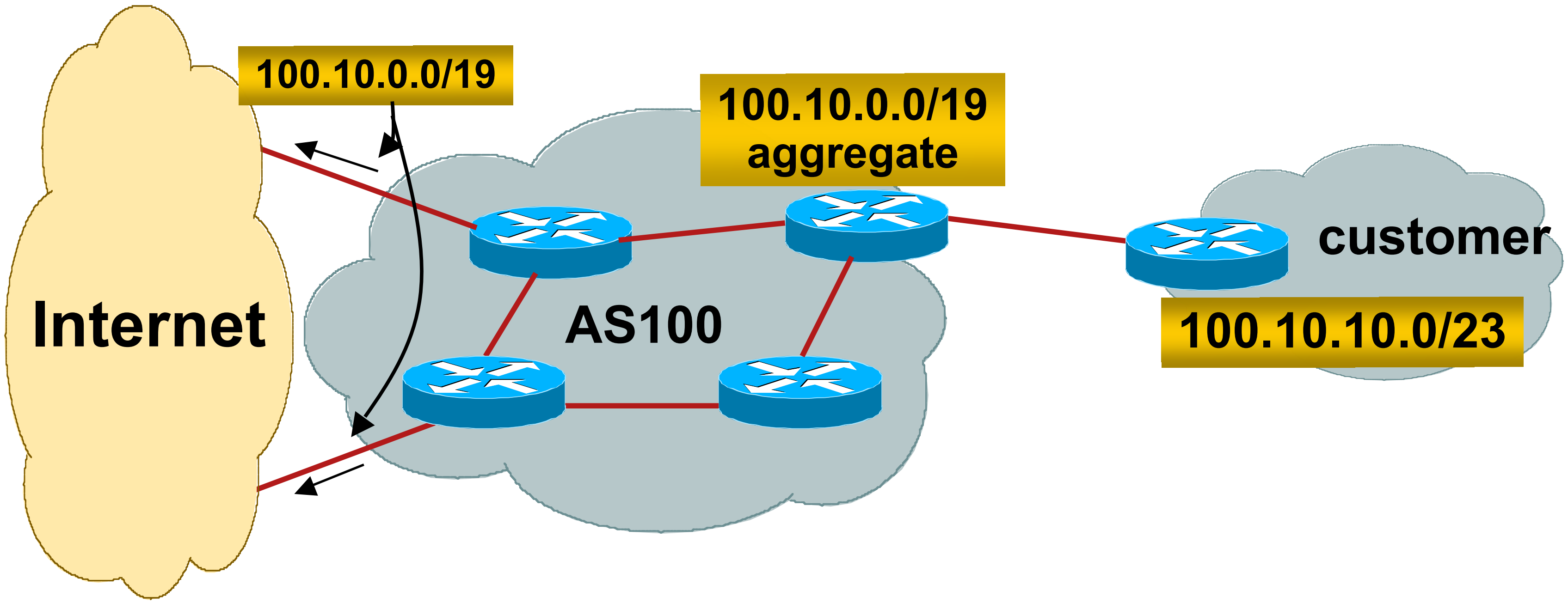


- Customer has /23 network assigned from AS100’s /19 address block
- AS100 announces customers’ individual networks to the Internet

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?

Separation of iBGP and eBGP

- Many ISPs do not understand the importance of separating iBGP and eBGP
  - iBGP is where all customer prefixes are carried
  - eBGP is used for announcing aggregate to Internet and for Traffic Engineering
- Do NOT do traffic engineering with customer originated iBGP prefixes
  - Leads to instability similar to that mentioned in the earlier bad example
  - Even though aggregate is announced, a flapping subprefix will lead to instability for the customer concerned
- **Generate traffic engineering prefixes on the Border Router**

The Internet Today (February 2011)

- Current Internet Routing Table Statistics

BGP Routing Table Entries	345357
Prefixes after maximum aggregation	155769
Unique prefixes in Internet	170883
Prefixes smaller than registry alloc	142680
/24s announced	180715
ASes in use	35825

"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 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 2010

RIR blocks contribute about 87% of the Internet Routing Table

Block	Networks	Block	Networks	Block	Networks	Block	Networks
24/8	3328	79/8	1119	118/8	1349	201/8	4136
41/8	3448	80/8	2335	119/8	1694	202/8	11354
58/8	1675	81/8	1709	120/8	531	203/8	11677
59/8	1575	82/8	1358	121/8	1756	204/8	5744
60/8	888	83/8	1357	122/8	2687	205/8	3037
61/8	2890	84/8	1341	123/8	2400	206/8	3951
62/8	2418	85/8	2492	124/8	2259	207/8	4635
63/8	3114	86/8	780	125/8	2514	208/8	6498
64/8	*6601*	87/8	1466	126/8	106	209/8	5536
65/8	3966	88/8	1068	173/8	1994	210/8	4977
66/8	*7782*	89/8	3168	174/8	1089	211/8	3130
67/8	3771	90/8	377	186/8	1223	212/8	3550
68/8	3221	91/8	4555	187/8	1501	213/8	3442
69/8	5280	96/8	778	189/8	3063	216/8	7645
70/8	2008	97/8	725	190/8	6945	217/8	3136
71/8	1327	98/8	1312	192/8	6952	218/8	1512
72/8	4050	99/8	288	193/8	6820	219/8	1303
73/8	4	112/8	883	194/8	5177	220/8	2108
74/8	*5074*	113/8	890	195/8	5325	221/8	980
75/8	1164	114/8	996	196/8	1857	222/8	1058
76/8	1034	115/8	1616	198/8	4504		
77/8	1964	116/8	1755	199/8	4372		
78/8	1397	117/8	1611	200/8	8884		

"The New Swamp" Summary

- RIR space shows creeping deaggregation

  It seems that an RIR /8 block averages around 5000 prefixes (and upwards) once fully allocated

- Food for thought:

  The 120 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

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

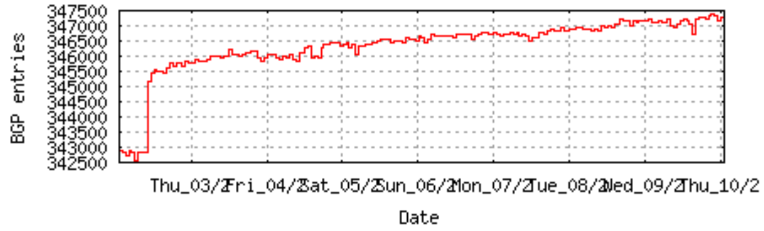
http://www.cidr-report.org/as2.0/

Radio Philip Networking Cisco Miscellaneous Smart Bookmarks TinyURL!

Status Summary

Table History

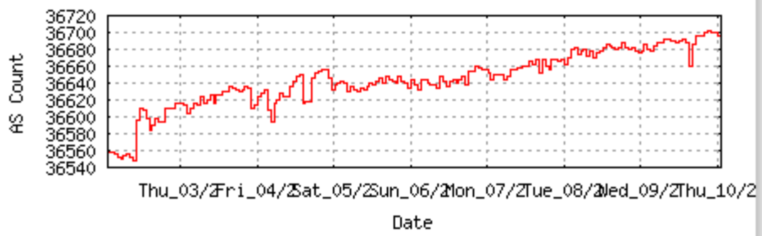
Date	Prefixes	CIDR Aggregated
03-02-11	345793	203290
04-02-11	345917	203345
05-02-11	346361	203582
06-02-11	346524	203630
07-02-11	346746	203680
08-02-11	346840	203697
09-02-11	347143	203702
10-02-11	347139	203784



Plot: BGP Table Size

AS Summary

36701	Number of ASes in routing system
15546	Number of ASes announcing only one prefix
3714	Largest number of prefixes announced by an AS AS6389: BELLSOUTH-NET-BLK - BellSouth.net Inc.
106679808	Largest address span announced by an AS (/32s) AS4134: CHINANET-BACKBONE No.31,Jin-rong Street



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

CIDR Report

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

Radio Philip Networking Cisco Miscellaneous Smart Bookmarks TinyURL!

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').

--- 19Feb11 ---

ASnum	NetsNow	NetsAggr	NetGain	% Gain	Description
Table	348860	204803	144057	41.3%	All ASes
AS6389	3694	266	3428	92.8%	BELLSOUTH-NET-BLK - BellSouth.net Inc.
AS4323	2614	410	2204	84.3%	TWTC - tw telecom holdings, inc.
AS19262	1843	283	1560	84.6%	VZGNI-TRANSIT - Verizon Online LLC
AS4766	2307	837	1470	63.7%	KIXS-AS-KR Korea Telecom
AS22773	1268	88	1180	93.1%	ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
AS4755	1419	341	1078	76.0%	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
AS6478	1510	468	1042	69.0%	ATT-INTERNET3 - AT&T Services, Inc.
AS1785	1787	763	1024	57.3%	AS-PAETEC-NET - PaeTec Communications, Inc.
AS28573	1232	309	923	74.9%	NET Servicos de Comunicao S.A.
AS7545	1628	720	908	55.8%	TPG-INTERNET-AP TPG Internet Pty Ltd
AS10620	1370	481	889	64.9%	Telmex Colombia S.A.
AS18566	1435	589	846	59.0%	COVAD - Covad Communications Co.
AS18101	930	154	776	83.4%	RELIANCE-COMMUNICATIONS-IN Reliance Communications Ltd.DAKC MUMBAI
AS7303	900	152	748	83.1%	Telecom Argentina S.A.
AS24560	1113	377	736	66.1%	AIRTELBROADBAND-AS-AP Bharti Airtel Ltd., Telemedia Services
AS4808	1046	323	723	69.1%	CHINA169-BJ CNCGROUP IP network China169 Beijing Province Network
AS6503	1167	445	722	61.9%	Axtel, S.A.B. de C.V.
AS3356	1190	490	700	58.8%	LEVEL3 Level 3 Communications
AS11492	1280	601	679	53.0%	CABLEONE - CABLE ONE, INC.
AS17488	950	275	675	71.1%	HATHWAY-NET-AP Hathway IP Over Cable Internet
AS8151	1350	676	674	49.9%	Uninet S.A. de C.V.
AS9498	764	129	635	83.1%	BBIL-AP BHARTI Airtel Ltd.
AS17676	651	70	581	89.2%	GIGAINFRA Softbank BB Corp.
AS855	633	58	575	90.8%	CANET-ASN-4 - Bell Aliant Regional Communications, Inc.
AS7552	663	110	553	83.4%	VIETEL-AS-AP Vietel Corporation
AS4780	767	226	541	70.5%	SEEDNET Digital United Inc.
AS14420	[illegible]	[illegible]	[illegible]	[illegible]	CORPORACION NACIONAL DE TELECOMUNICACIONES - CNT EP

CIDR Report

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

Radio Philip Networking Cisco Miscellaneous Smart Bookmarks TinyURL!

Top 20 Added Routes this week per Originating AS

Prefixes	ASnum	AS Description
336	AS4766	KIXS-AS-KR Korea Telecom
307	AS15475	NOL
153	AS18566	COVAD - Covad Communications Co.
128	AS8452	TE-AS TE-AS
78	AS35908	VPLSNET - VPLS Inc. d/b/a Krypt Technologies
62	AS16422	NEWSKIES-NETWORKS SES WORLD SKIES ARIN AS, for routing RIPE space.
62	AS1659	ERX-TANET-ASN1 Tiawan Academic Network (TANet) Information Center
60	AS11139	CWRIN CW BARBADOS
54	AS23650	CHINANET-JS-AS-AP AS Number for CHINANET jiangsu province backbone
46	AS36992	ETISALAT-MISR
43	AS15706	Sudatel
39	AS3	MIT-GATEWAYS - Massachusetts Institute of Technology
36	AS45975	CNSCE-AS-KR CHUNGCHEONGNAMDO SEOCHEON OFFICE OF EDUCATION
35	AS43875	DATAINFO-ASN SC Data Media Info SRL
34	AS33770	KDN
34	AS3816	COLOMBIA TELECOMUNICACIONES S.A. ESP
34	AS3130	RGNET-3130 RGnet/PSGnet
33	AS55595	--No Registry Entry--
32	AS23487	CONECEL
30	AS56048	CMNET-BEIJING-AP China Mobile Communicaitons Corporation

Top 20 Withdrawn Routes this week per Originating AS

Prefixes	ASnum	AS Description
-330	AS36992	ETISALAT-MISR
-231	AS24863	LINKdotNET-AS
-125	AS50010	NAWRAS-AS Omani Qatari Telecommunications Company SAOC
-115	AS8452	TE-AS TE-AS
-79	AS9318	HANARO-AS Hanaro Telecom Inc.
-62	AS41843	ERTH-OMSK-AS CJSC "ER-Telecom Holding" Omsk branch
-61	AS15475	NOL
-56	AS17911	BRAINPK-AS-AP Brain Telecommunication Ltd.
-54	AS24835	RAYA-AS
-46	AS52026	TRUF-AS TRUF d.o.o.
-29	AS36935	Vodafone-EG
-28	AS27817	Red Nacional Académica de Tecnología Avanzada - RENATA

CIDR Report

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

Radio Philip Networking Cisco Miscellaneous Smart Bookmarks TinyURL!

Report: Withdrawn Route count per Originating AS

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
3593	3694	AS6389	BELLSOUTH-NET-BLK - BellSouth.net Inc.
2407	2614	AS4323	TWTC - tw telecom holdings, inc.
2248	2307	AS4766	KIXS-AS-KR Korea Telecom
1777	1843	AS19262	VZGNI-TRANSIT - Verizon Online LLC
1698	1787	AS1785	AS-PAETEC-NET - PaeTec Communications, Inc.
1572	1628	AS7545	TPG-INTERNET-AP TPG Internet Pty Ltd
1509	1510	AS6478	ATT-INTERNET3 - AT&T Services, Inc.
1483	1527	AS20115	CHARTER-NET-HKY-NC - Charter Communications
1423	1435	AS18566	COVAD - Covad Communications Co.
1406	1419	AS4755	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
1396	1409	AS17974	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
1368	1370	AS10620	Telmex Colombia S.A.
1344	1350	AS8151	Uninet S.A. de C.V.
1271	1280	AS11492	CABLEONE - CABLE ONE, INC.
1232	1232	AS28573	NET Servicos de Comunicao S.A.
1226	1268	AS22773	ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
1207	1299	AS2386	INS-AS - AT&T Data Communications Services
1165	1167	AS6503	Axtel, S.A.B. de C.V.
1155	1160	AS7011	FRONTIER-AND-CITIZENS - Frontier Communications of America, Inc.
1113	1370	AS7018	ATT-INTERNET4 - AT&T 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)

Possible Bogus Routes and AS Announcements

AS Report

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

Radio ▾ Philip ▾ Networking ▾ Cisco ▾ Miscellaneous ▾ Smart Bookmarks ▾ TinyURL!

Announced Prefixes

Rank	AS	Type	Originate Addr Space	(pfx)	Transit Addr space	(pfx)	Description
124	AS4755	ORG+TRN	Originate: 3601920	/10.22	Transit: 10295296	/8.70	TATACOMM-AS TATA Communications formerly VSNL is

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	%
8	AS4755	TATACOMM-AS TATA Communications formerly VSNL	1414	1124	46	336	1078	76.24%

Prefix	AS Path	Aggregation Suggestion
14.140.0.0/14	4777 2516 6453 4755	
14.140.0.0/22	4777 2516 6453 4755	- Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
14.140.16.0/22	4777 2516 6453 4755	- Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
14.140.24.0/22	4777 2516 6453 4755	- Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
14.140.254.0/23	4777 2516 6453 4755	- Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
49.32.0.0/12	4777 2516 6453 4755	
59.151.144.0/22	4608 1221 4637 6453 4755	
59.160.0.0/16	4777 2516 6453 4755	
59.160.0.0/22	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.4.0/22	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.5.0/24	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.8.0/22	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.11.0/24	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.12.0/22	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.15.0/24	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.16.0/21	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.24.0/21	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.24.0/24	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.32.0/21	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.34.0/24	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.38.0/24	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.44.0/22	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.46.0/23	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.48.0/21	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.48.0/24	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.56.0/21	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.64.0/21	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.71.0/24	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.72.0/21	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.73.0/24	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.81.0/24	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.83.0/24	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
59.160.88.0/22	4777 2516 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755

AS Report

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

Radio Philip Networking Cisco Miscellaneous Smart Bookmarks TinyURL!

Announced Prefixes

```
Rank  AS       Type    Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
168   AS18566        ORG+TRN Originate:    2395136 /10.81  Transit:       1024 /22.00 COVAD - Covad Communications Co.
```

Aggregation Suggestions

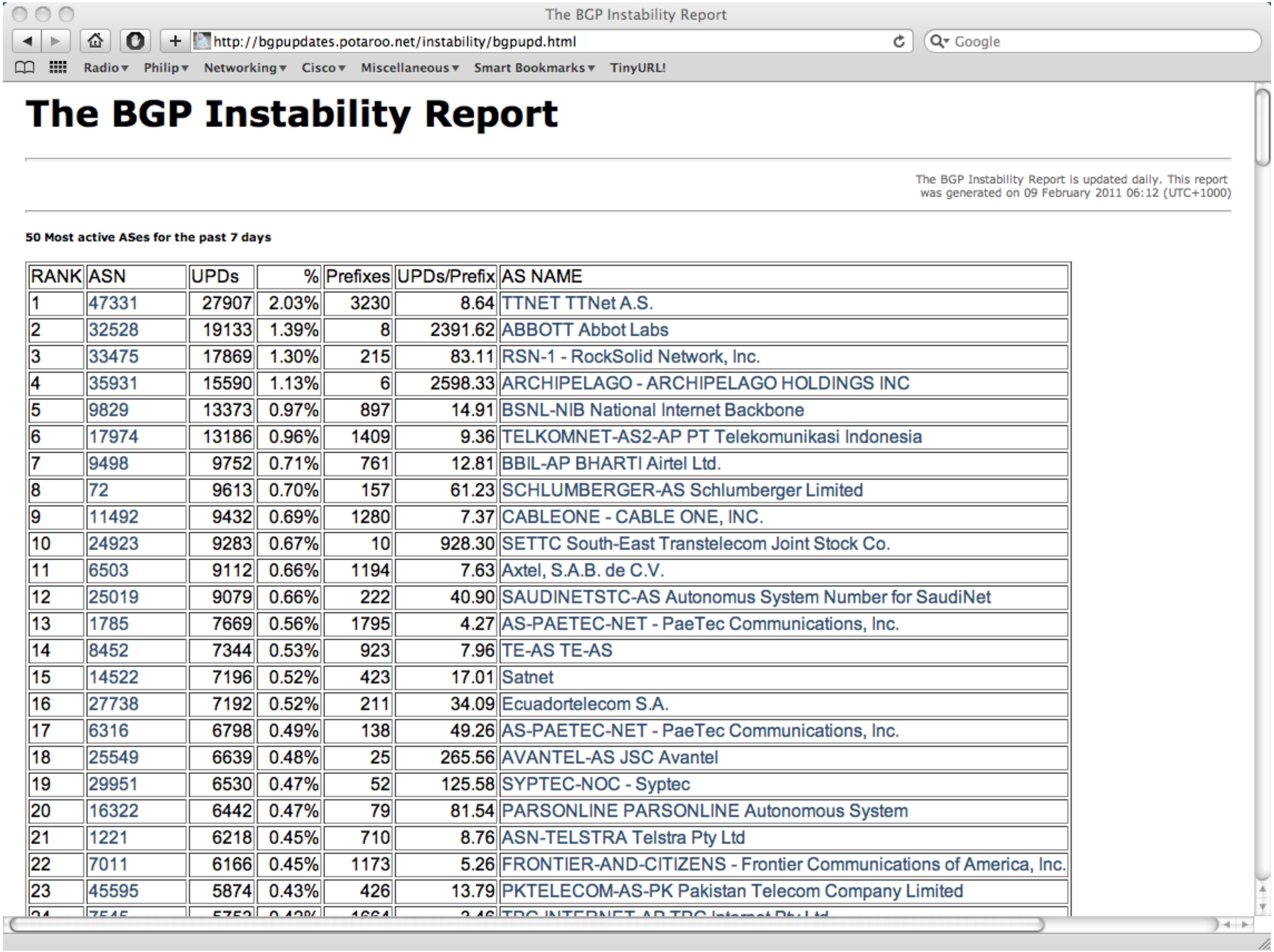
This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

```
 Rank AS                AS Name                                   Current  Wthdw  Aggte  Annce Redctn      %
   19 AS18566           COVAD - Covad Communications Co.             1197    888    213    522    675  56.39%


  Prefix                AS Path                             Aggregation Suggestion
  64.81.22.0/24         4777 2516 4565 18566
  64.81.96.0/21         4777 2516 4565 18566
  64.81.96.0/24         4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.96.0/21 4777 2516 4565 18566
  64.81.97.0/24         4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.96.0/21 4777 2516 4565 18566
  64.81.98.0/24         4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.96.0/21 4777 2516 4565 18566
  64.81.99.0/24         4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.96.0/21 4777 2516 4565 18566
  64.81.100.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.96.0/21 4777 2516 4565 18566
  64.81.101.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.96.0/21 4777 2516 4565 18566
  64.81.102.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.96.0/21 4777 2516 4565 18566
  64.81.103.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.96.0/21 4777 2516 4565 18566
  64.81.104.0/22        4777 2516 4565 18566
  64.81.108.0/23        4777 2516 4565 18566 + Announce - aggregate of 64.81.108.0/24 (4777 2516 4565 18566) and 64.81.109.0/24 (4777 25
  64.81.108.0/24        4777 2516 4565 18566 - Withdrawn - aggregated with 64.81.109.0/24 (4777 2516 4565 18566)
  64.81.109.0/24        4777 2516 4565 18566 - Withdrawn - aggregated with 64.81.108.0/24 (4777 2516 4565 18566)
  64.81.110.0/24        4777 2516 4565 18566
  64.105.0.0/16         4777 2516 3356 18566
  64.105.0.0/23         4777 2516 3356 18566 - Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
  64.105.4.0/22         4777 2516 4565 18566 + Announce - aggregate of 64.105.4.0/23 (4777 2516 4565 18566) and 64.105.6.0/23 (4777 2516
  64.105.4.0/23         4777 2516 4565 18566 - Withdrawn - aggregated with 64.105.6.0/23 (4777 2516 4565 18566)
  64.105.6.0/23         4777 2516 4565 18566 - Withdrawn - aggregated with 64.105.4.0/23 (4777 2516 4565 18566)
  64.105.8.0/23         4777 2516 3356 18566 - Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
  64.105.10.0/23        4777 2516 4565 18566
  64.105.14.0/23        4777 2516 4565 18566
  64.105.16.0/24        4777 2516 3356 18566 - Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
  64.105.17.0/24        4777 2516 4565 18566
  64.105.18.0/23        4777 2516 3356 18566 - Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
  64.105.20.0/22        4777 2516 4565 18566 + Announce - aggregate of 64.105.20.0/23 (4777 2516 4565 18566) and 64.105.22.0/23 (4777 25
  64.105.20.0/23        4777 2516 4565 18566 - Withdrawn - aggregated with 64.105.22.0/23 (4777 2516 4565 18566)
  64.105.22.0/23        4777 2516 4565 18566 - Withdrawn - aggregated with 64.105.20.0/23 (4777 2516 4565 18566)
  64.105.24.0/21        4777 2516 4565 18566
  64.105.32.0/20        4777 2516 4565 18566 + Announce - aggregate of 64.105.32.0/21 (4777 2516 4565 18566) and 64.105.40.0/21 (4777 25
  64.105.32.0/21        4777 2516 4565 18566 - Withdrawn - aggregated with 64.105.40.0/21 (4777 2516 4565 18566)
```

Importance of Aggregation

- Size of routing table
  - Router Memory is not so much of a problem as it was in the 1990s
  - 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 09 February 2011 06:12 (UTC+1000)

50 Most active ASes for the past 7 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	47331	27907	2.03%	3230	8.64	TTNET TTNet A.S.
2	32528	19133	1.39%	8	2391.62	ABBOTT Abbot Labs
3	33475	17869	1.30%	215	83.11	RSN-1 - RockSolid Network, Inc.
4	35931	15590	1.13%	6	2598.33	ARCHIPELAGO - ARCHIPELAGO HOLDINGS INC
5	9829	13373	0.97%	897	14.91	BSNL-NIB National Internet Backbone
6	17974	13186	0.96%	1409	9.36	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
7	9498	9752	0.71%	761	12.81	BBIL-AP BHARTI Airtel Ltd.
8	72	9613	0.70%	157	61.23	SCHLUMBERGER-AS Schlumberger Limited
9	11492	9432	0.69%	1280	7.37	CABLEONE - CABLE ONE, INC.
10	24923	9283	0.67%	10	928.30	SETTC South-East Transtelecom Joint Stock Co.
11	6503	9112	0.66%	1194	7.63	Axtel, S.A.B. de C.V.
12	25019	9079	0.66%	222	40.90	SAUDINETSTC-AS Autonomus System Number for SaudiNet
13	1785	7669	0.56%	1795	4.27	AS-PAETEC-NET - PaeTec Communications, Inc.
14	8452	7344	0.53%	923	7.96	TE-AS TE-AS
15	14522	7196	0.52%	423	17.01	Satnet
16	27738	7192	0.52%	211	34.09	Ecuadortelecom S.A.
17	6316	6798	0.49%	138	49.26	AS-PAETEC-NET - PaeTec Communications, Inc.
18	25549	6639	0.48%	25	265.56	AVANTEL-AS JSC Avantel
19	29951	6530	0.47%	52	125.58	SYPTEC-NOC - Syptec
20	16322	6442	0.47%	79	81.54	PARSONLINE PARSONLINE Autonomous System
21	1221	6218	0.45%	710	8.76	ASN-TELSTRA Telstra Pty Ltd
22	7011	6166	0.45%	1173	5.26	FRONTIER-AND-CITIZENS - Frontier Communications of America, Inc.
23	45595	5874	0.43%	426	13.79	PKTELECOM-AS-PK Pakistan Telecom Company Limited

The BGP Instability Report

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

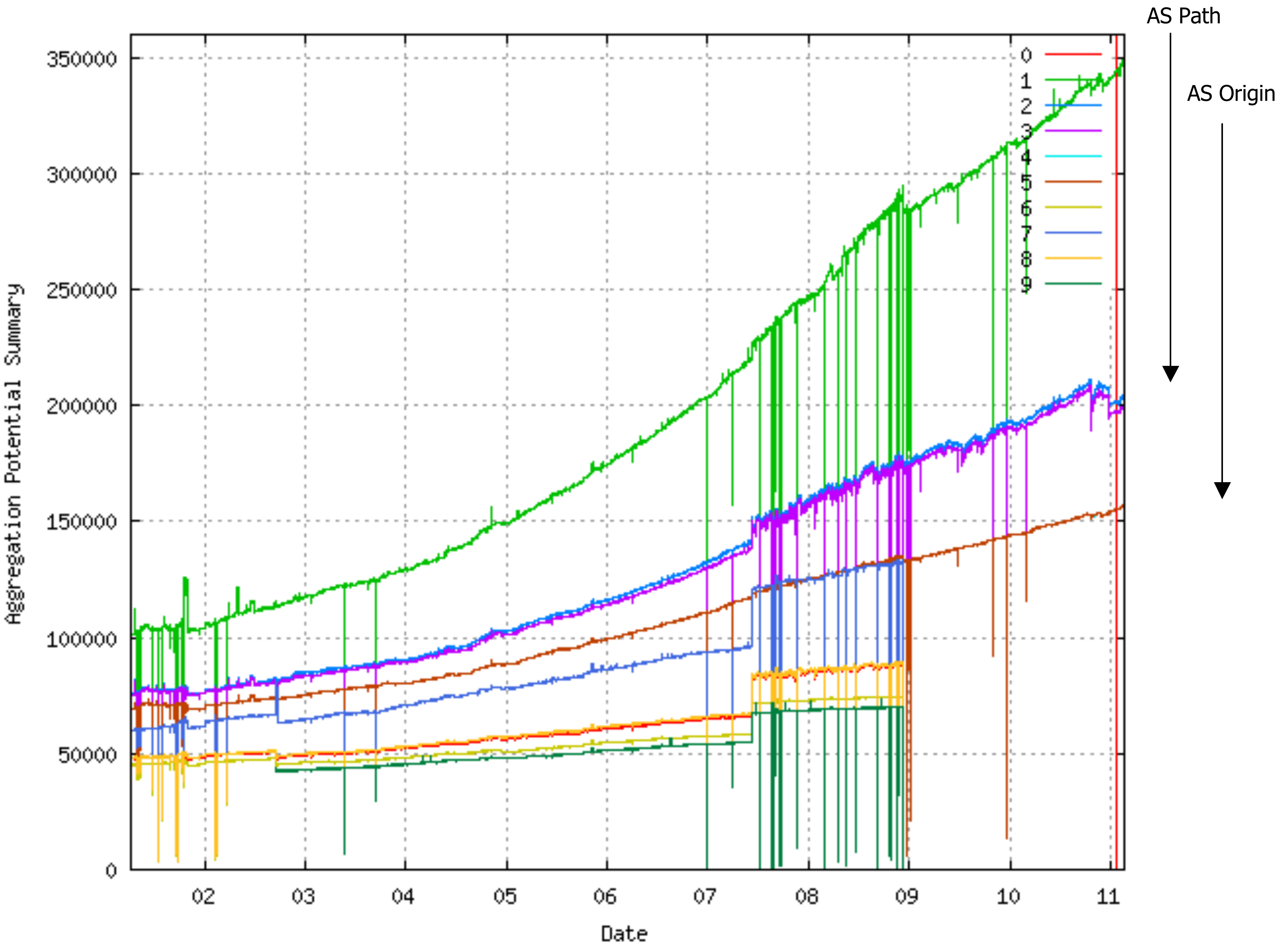
Google

Radio ▾ Philip ▾ Networking ▾ Cisco ▾ Miscellaneous ▾ Smart Bookmarks ▾ TinyURL!

50 Most active Prefixes for the past 7 days

RANK	PREFIX	UPDs	%	Origin AS -- AS NAME
1	63.211.68.0/22	10173	0.69%	35931 -- ARCHIPELAGO - ARCHIPELAGO HOLDINGS INC
2	130.36.34.0/24	9565	0.65%	32528 -- ABBOTT Abbot Labs
3	130.36.35.0/24	9565	0.65%	32528 -- ABBOTT Abbot Labs
4	213.129.96.0/19	9255	0.63%	24923 -- SETTC South-East Transtelecom Joint Stock Co.
5	202.92.235.0/24	6772	0.46%	9498 -- BBIL-AP BHARTI Airtel Ltd.
6	216.126.136.0/22	6422	0.44%	6316 -- AS-PAETEC-NET - PaeTec Communications, Inc.
7	198.140.43.0/24	5346	0.37%	35931 -- ARCHIPELAGO - ARCHIPELAGO HOLDINGS INC
8	68.65.152.0/22	3728	0.25%	11915 -- TELWEST-NETWORK-SVCS-STATIC - TEL WEST COMMUNICATIONS LLC
9	80.245.240.0/20	3612	0.25%	35738 -- KVANT-AS Kvant ltd.
10	95.170.128.0/19	3545	0.24%	25549 -- AVANTEL-AS JSC Avantel
11	202.153.174.0/24	3301	0.23%	17408 -- ABOVE-AS-AP AboveNet Communications Taiwan
12	183.88.0.0/16	3299	0.23%	45629 -- JASTEL-NETWORK-TH-AP Jasmine International Tower 45758 -- TRIPLETNET-AS-AP TripleT Internet Internet service provider Bangkok
13	223.206.0.0/16	3291	0.22%	45629 -- JASTEL-NETWORK-TH-AP Jasmine International Tower 45758 -- TRIPLETNET-AS-AP TripleT Internet Internet service provider Bangkok
14	206.184.16.0/24	3195	0.22%	174 -- COGENT Cogent/PSI
15	93.91.160.0/20	2963	0.20%	25549 -- AVANTEL-AS JSC Avantel
16	114.128.0.0/16	2931	0.20%	45629 -- JASTEL-NETWORK-TH-AP Jasmine International Tower 45758 -- TRIPLETNET-AS-AP TripleT Internet Internet service provider Bangkok
17	192.190.209.0/24	2806	0.19%	1221 -- ASN-TELSTRA Telstra Pty Ltd
18	192.190.214.0/24	2806	0.19%	1221 -- ASN-TELSTRA Telstra Pty Ltd
19	62.36.229.0/24	2462	0.17%	12479 -- UNI2-AS France Telecom Espana SA
20	213.108.216.0/21	2243	0.15%	49776 -- GORSET-AS Gorodskaya Set Ltd.
21	213.170.59.0/24	1706	0.12%	49600 -- LASEDA La Seda de Barcelona, S.A
22	208.54.82.0/24	1615	0.11%	701 -- UUNET - MCI Communications Services, Inc. d/b/a Verizon Business
23	159.18.255.0/24	1347	0.09%	6401 -- ALLST-6401 - Allstream Corp.
24	210.82.213.0/24	1311	0.09%	9929 -- CNCNET-CN China Netcom Corp.
25	210.82.212.0/24	1310	0.09%	9929 -- CNCNET-CN China Netcom Corp.
26	210.82.252.0/24	1310	0.09%	9929 -- CNCNET-CN China Netcom Corp.
27	210.82.242.0/23	1307	0.09%	9929 -- CNCNET-CN China Netcom Corp.

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:

```
$ whois -h whois.apnic.net 202.12.29.0
inetnum:        202.12.28.0 - 202.12.29.255
netname:        APNIC-AP
descr:          Asia Pacific Network Information Center
descr:          Level 1 - 33 Park Road.
descr:          Milton QLD 4064
descr:          Australia
country:        AU
admin-c:        AIC1-AP
tech-c:         NO4-AP
mnt-by:         APNIC-HM
changed:        technical@apnic.net 19980918
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 193.128.0.0
inetnum:        193.128.0.0 - 193.133.255.255
netname:        UK-PIPEX-193-128-133
descr:          Verizon UK Limited
country:        GB
org:            ORG-UA24-RIPE
admin-c:        WERT1-RIPE
tech-c:         UPHM1-RIPE
status:         ALLOCATED UNSPECIFIED
remarks:        Please send abuse notification to abuse@uk.uu.net
mnt-by:         RIPE-NCC-HM-MNT
mnt-lower:      AS1849-MNT
mnt-routes:     AS1849-MNT
mnt-routes:     WCOM-EMEA-RICE-MNT
mnt-irt:        IRT-MCI-GB
source:         RIPE # Filtered
```

ALLOCATED – means that this is Provider Aggregatable address space and can only be announced by the ISP holding the allocation (in this case Verizon UK)

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
  - Traffic Engineering – see BGP Multihoming Tutorial
- Ask upstream/transit provider to either:
  - originate a default-route

    OR
  - announce one prefix you can use as default

Receiving Prefixes: From Upstream/Transit Provider

- If necessary to receive prefixes from any provider, care is required.
  - Don't accept default (unless you need it)
  - Don't accept your own prefixes
- For IPv4:
  - Don't accept private (RFC1918) and certain special use prefixes:
    - **http://www.rfc-editor.org/rfc/rfc5735.txt**
  - Don't accept prefixes longer than /24 (?)
- For IPv6:
  - Don't accept certain special use prefixes:
    - **http://www.rfc-editor.org/rfc/rfc5156.txt**
  - Don't accept prefixes longer than /48 (?)

Receiving Prefixes: From Upstream/Transit Provider

- Check Team Cymru's list of "bogons"

  www.team-cymru.org/Services/Bogons/http.html

- For IPv4 also consult:

  www.ietf.org/internet-drafts/draft-vegoda-no-more-unallocated-slash8s-00.txt

- For IPv6 also consult:

  www.space.net/~gert/RIPE/ipv6-filters.html

- Bogon Route Server:

  www.team-cymru.org/Services/Bogons/routeserver.html

  Supplies a BGP feed (IPv4 and/or IPv6) of address blocks which should not appear in the BGP table

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

```
*>  96.27.246.0/24       2497 1239 12026 12026 12026
12026 12026 12026 12026 12026 12026 12026 12026
12026 12026 12026 12026 12026 12026 12026 12026
12026 12026 12026 i
```

  This example shows 21 prepends (for no obvious reason)

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

BGP TTL "hack"

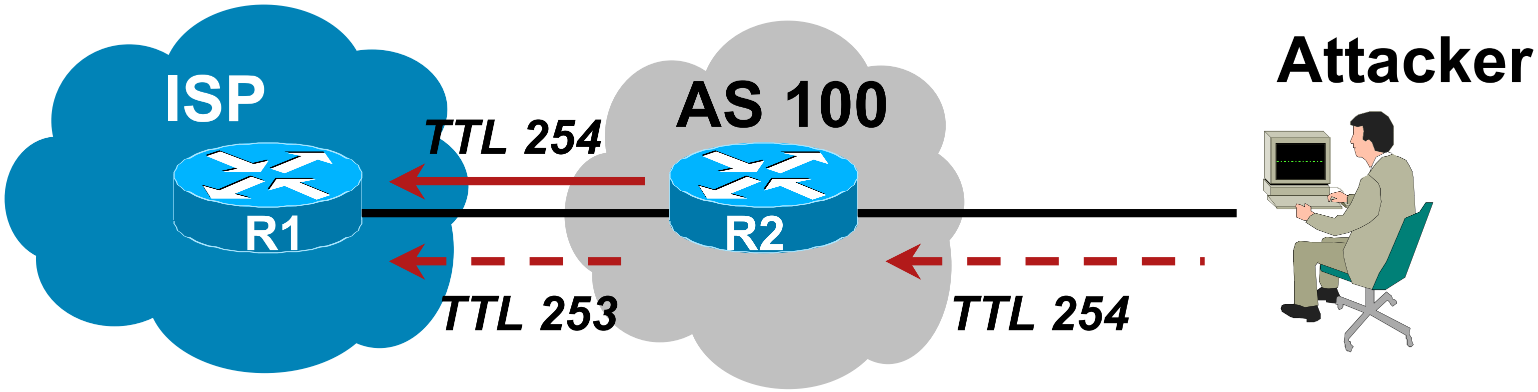
- Implement RFC5082 on BGP peerings

  (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 2011

Hong Kong, SAR, China

15 - 25 February 2011

