BGP Techniques for Internet Service Providers

Philip Smith

SANOG 23

15 January 2014

Thimphu

Last updated 12 January 2014

Presentation Slides

- Will be available on
  - http://thyme.apnic.net/ftp/seminars/SANOG23-BGP-Techniques.pdf
  - And on the SANOG 23 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

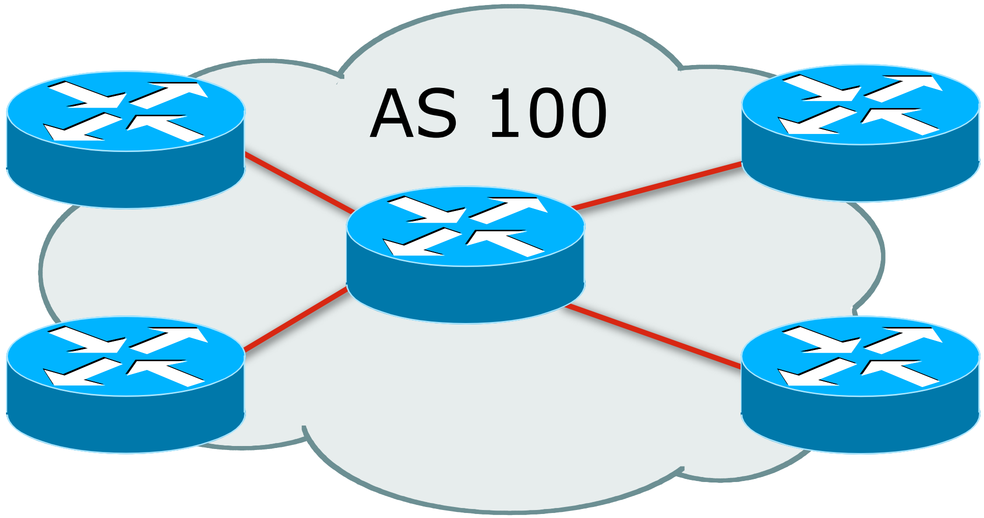
BGP Basics

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 – RFC6793)

- Usage:

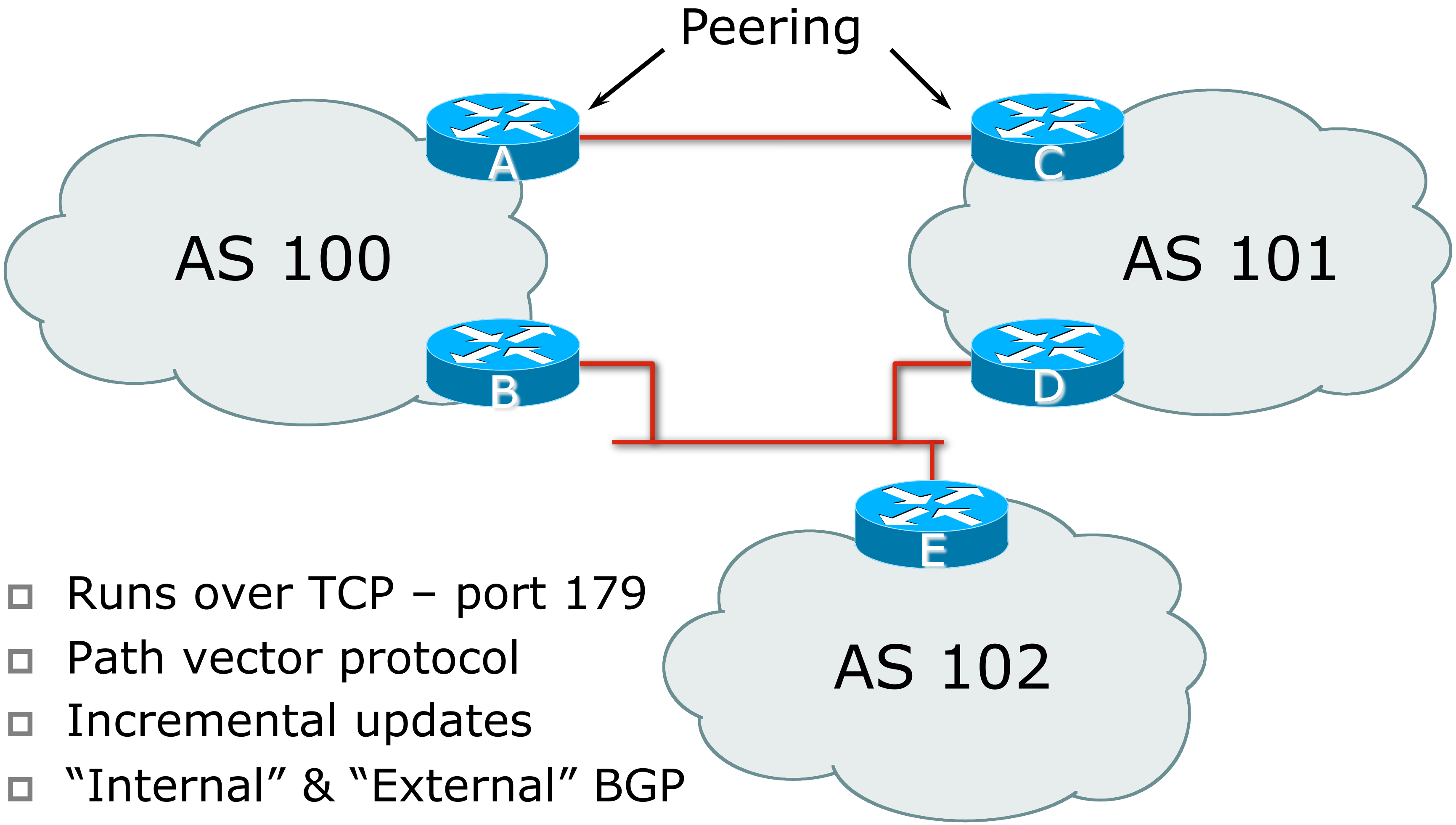
0 and 65535	(reserved)
1-64495	(public Internet)
64496-64511	(documentation – RFC5398)
64512-65534	(private use only)
23456	(represent 32-bit range in 16-bit world)
65536-65551	(documentation – RFC5398)
65552-4199999999	(public Internet)
4200000000-4294967295	(private use only – RFC6996)

- 32-bit range representation specified in RFC5396
  - Defines "asplain" (traditional format) as standard notation

Autonomous System Number (ASN)

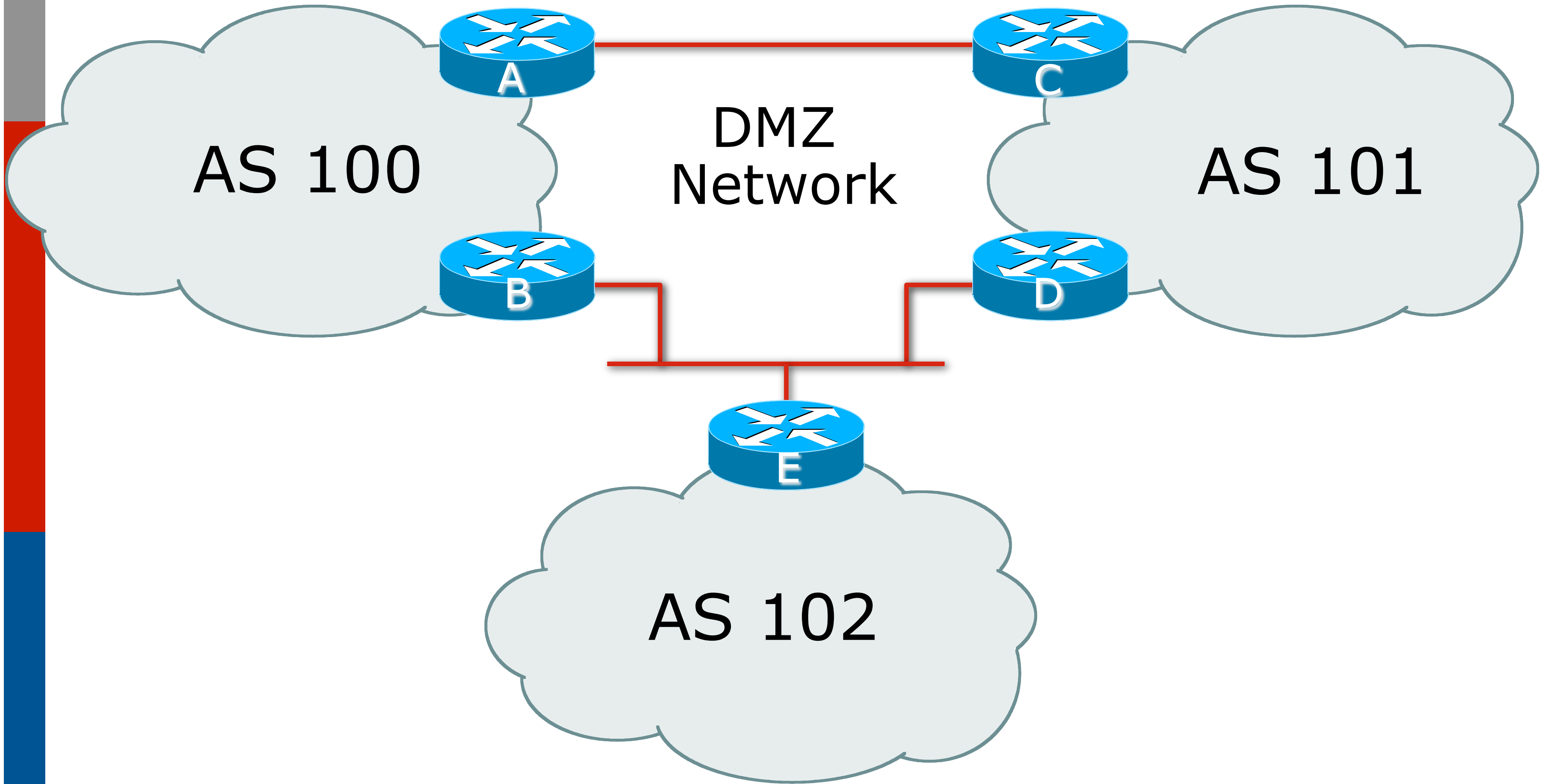
- ASNs are distributed by the Regional Internet Registries
  - They are also available from upstream ISPs who are members of one of the RIRs
- Current 16-bit ASN assignments up to 63999 have been made to the RIRs
  - Around 46000 are visible on the Internet
  - Around 500 left unassigned
- Each RIR has also received a block of 32-bit ASNs
  - Out of 5600 assignments, around 4400 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)



- DMZ is the link or network shared between ASes

BGP General Operation

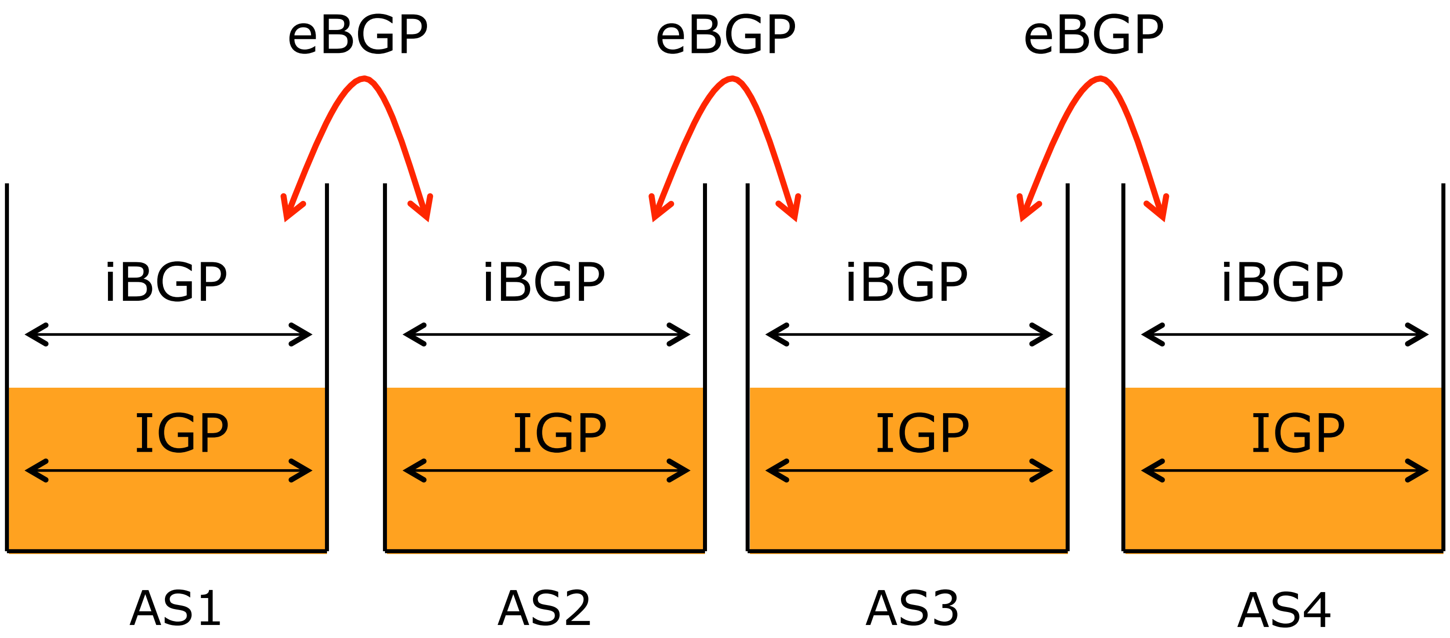
- Learns multiple paths via internal and external BGP speakers
- Picks the best path and installs in the forwarding table
- Best path is sent to external BGP neighbours
- Policies are applied by influencing the best path selection

eBGP & iBGP

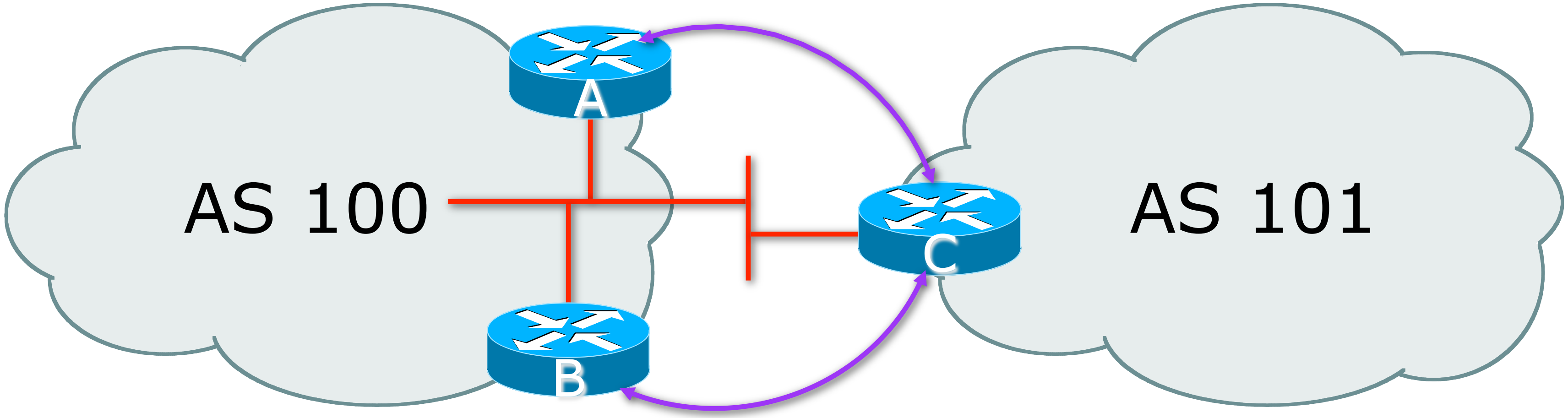
- BGP used internally (iBGP) and externally (eBGP)
- iBGP used to carry
  - Some/all Internet prefixes across ISP backbone
  - ISP's customer prefixes
- eBGP used to
  - Exchange prefixes with other ASes
  - Implement routing policy

BGP/IGP model used in ISP networks

- Model representation



External BGP Peering (eBGP)

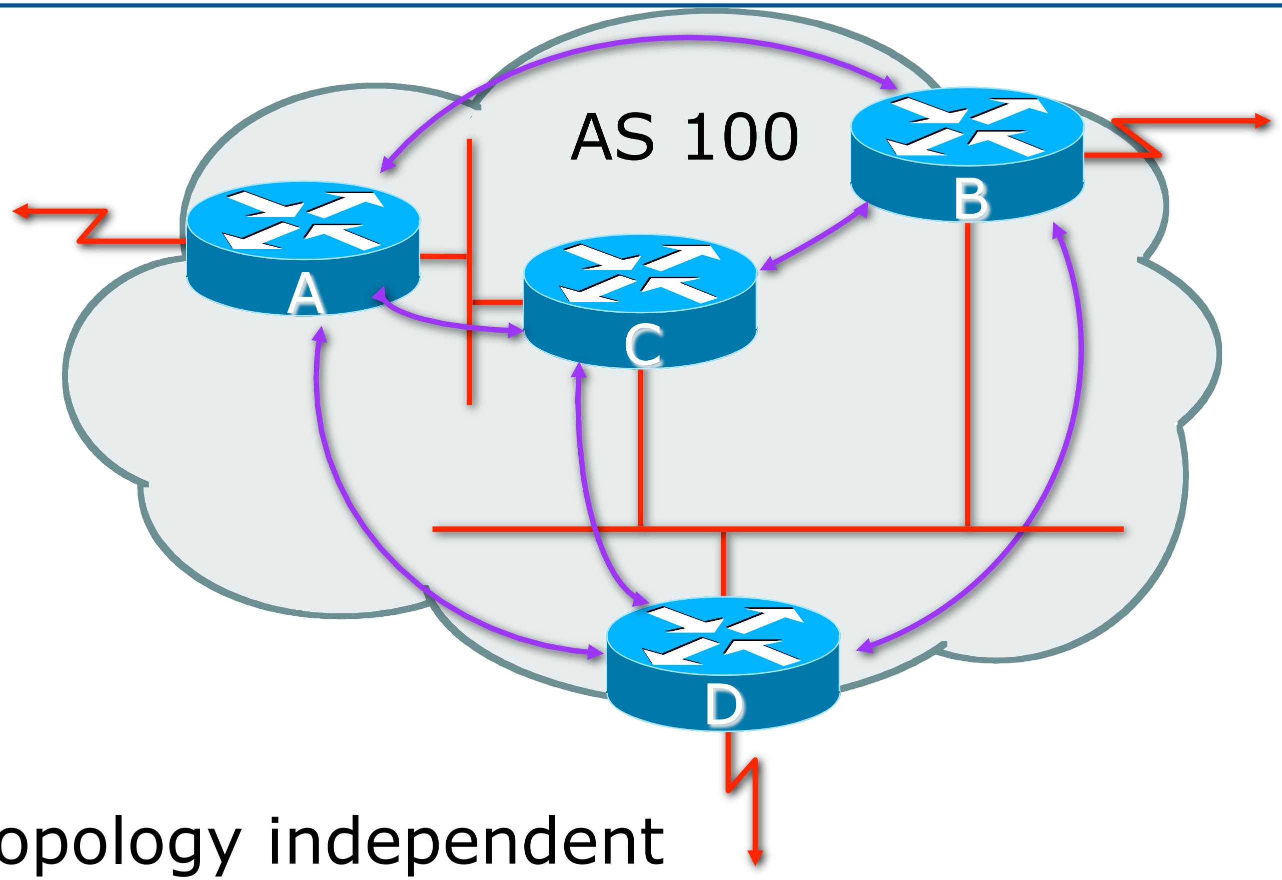


- Between BGP speakers in different AS
- Should be directly connected
- **Never** run an IGP between eBGP peers

Internal BGP (iBGP)

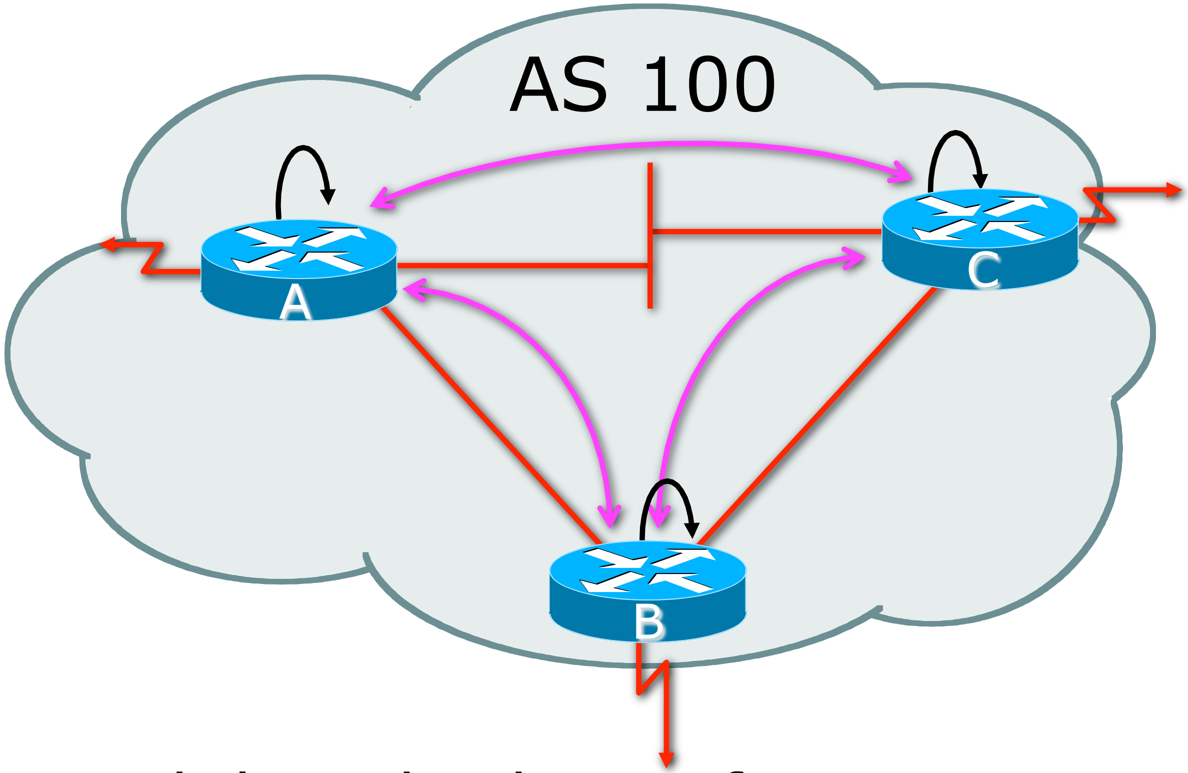
- BGP peer within the same AS
- Not required to be directly connected
  - IGP takes care of inter-BGP speaker connectivity
- iBGP speakers must to be fully meshed:
  - They originate connected networks
  - They pass on prefixes learned from outside the ASN
  - They do not pass on prefixes learned from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

Internal BGP Peering (iBGP)



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

Peering between 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

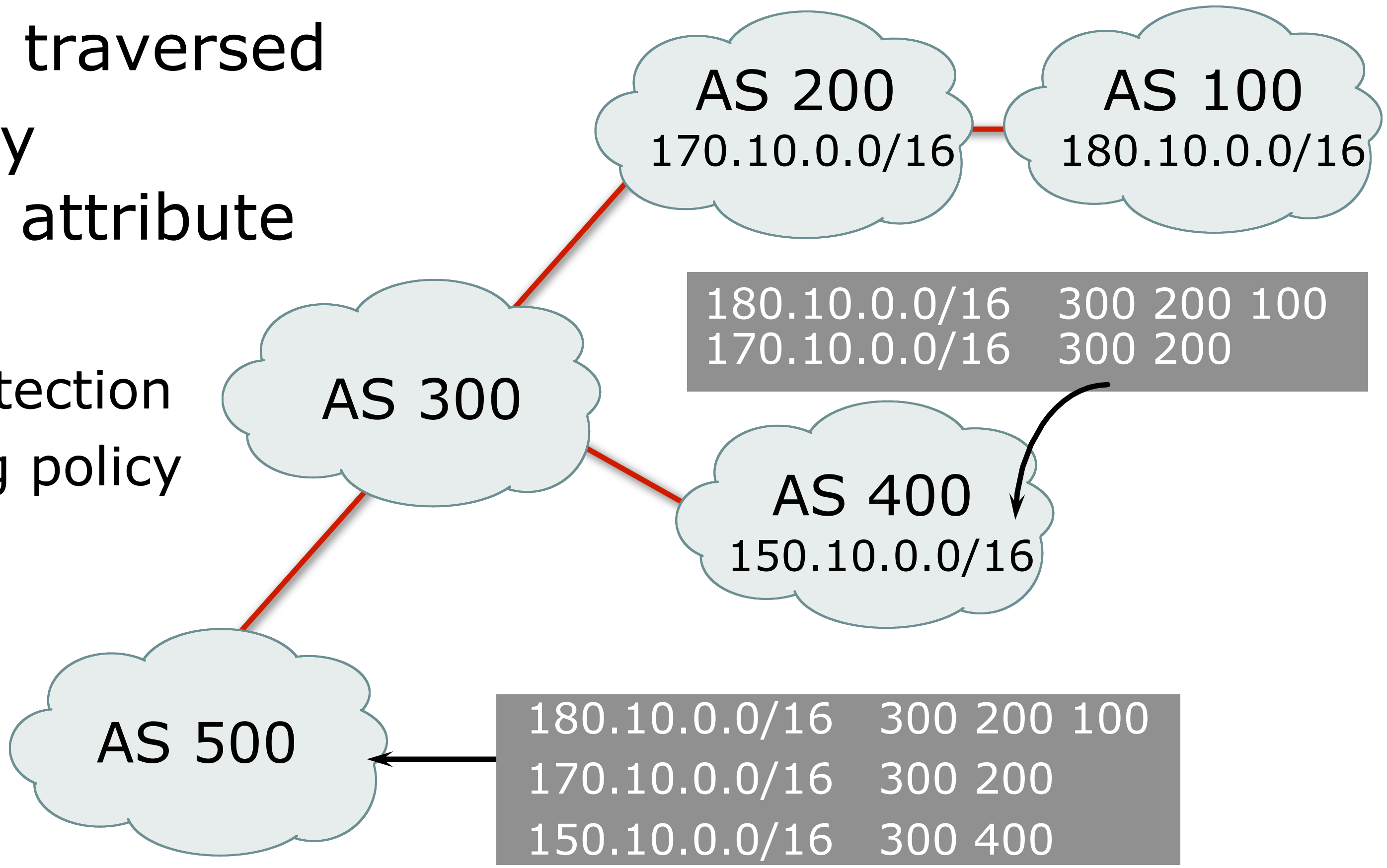
Information about BGP

BGP Attributes

- Carry various information about or characteristics of the prefix being propagated
  - AS-PATH
  - NEXT-HOP
  - ORIGIN
  - AGGREGATOR
  - LOCAL_PREFERENCE
  - Multi-Exit Discriminator
  - (Weight)
  - COMMUNITY

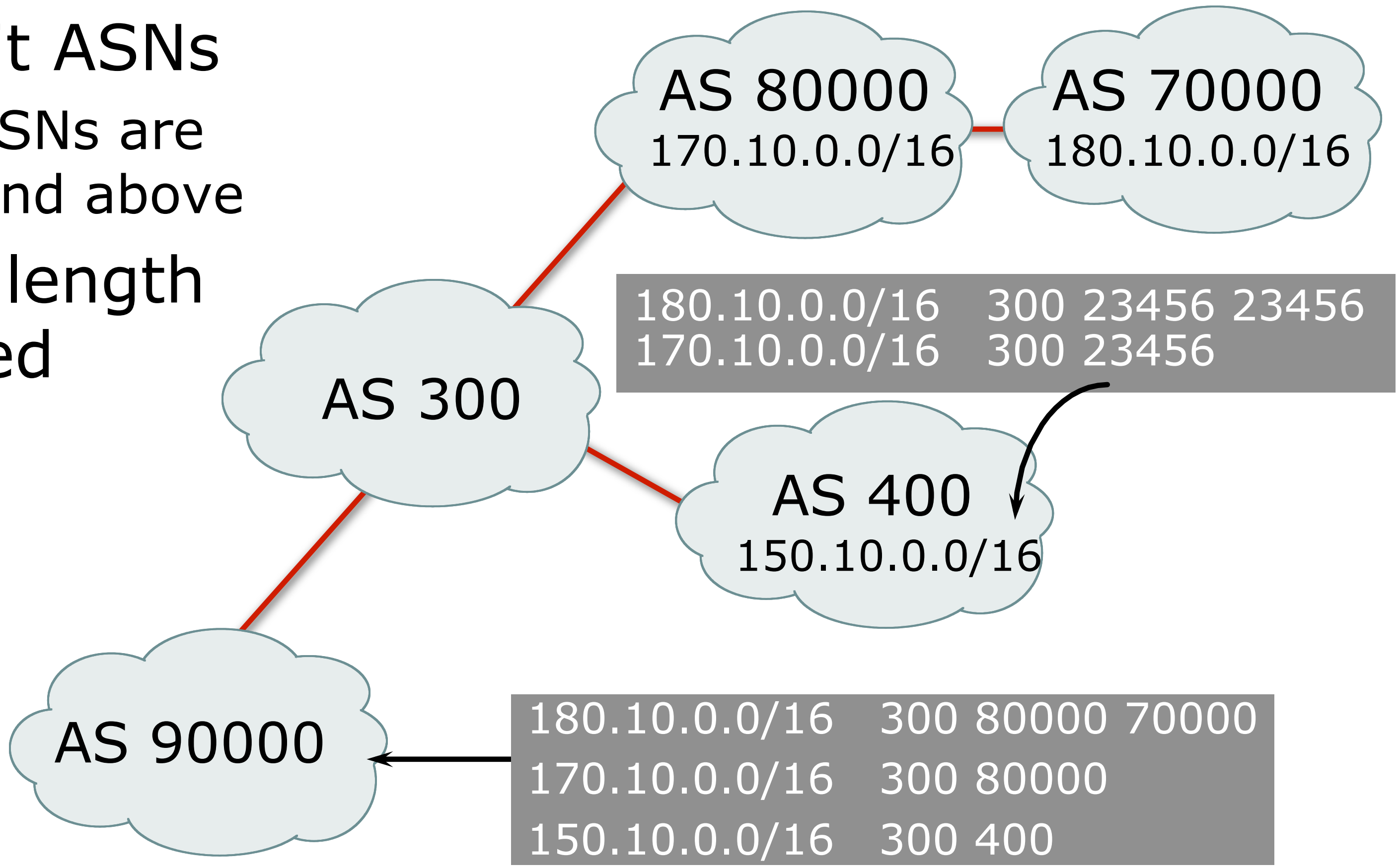
AS-Path

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

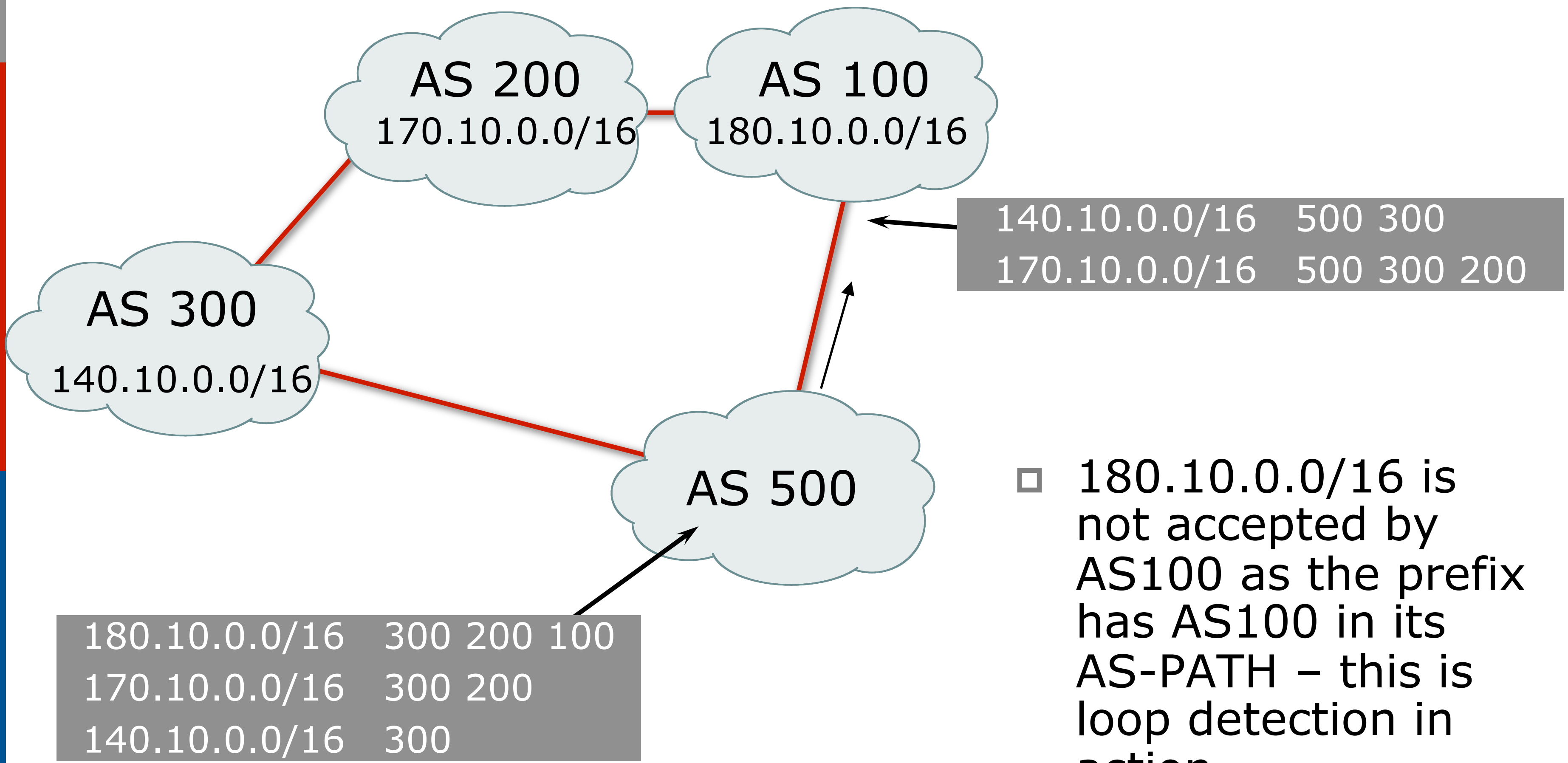


AS-Path (with 16 and 32-bit ASNs)

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

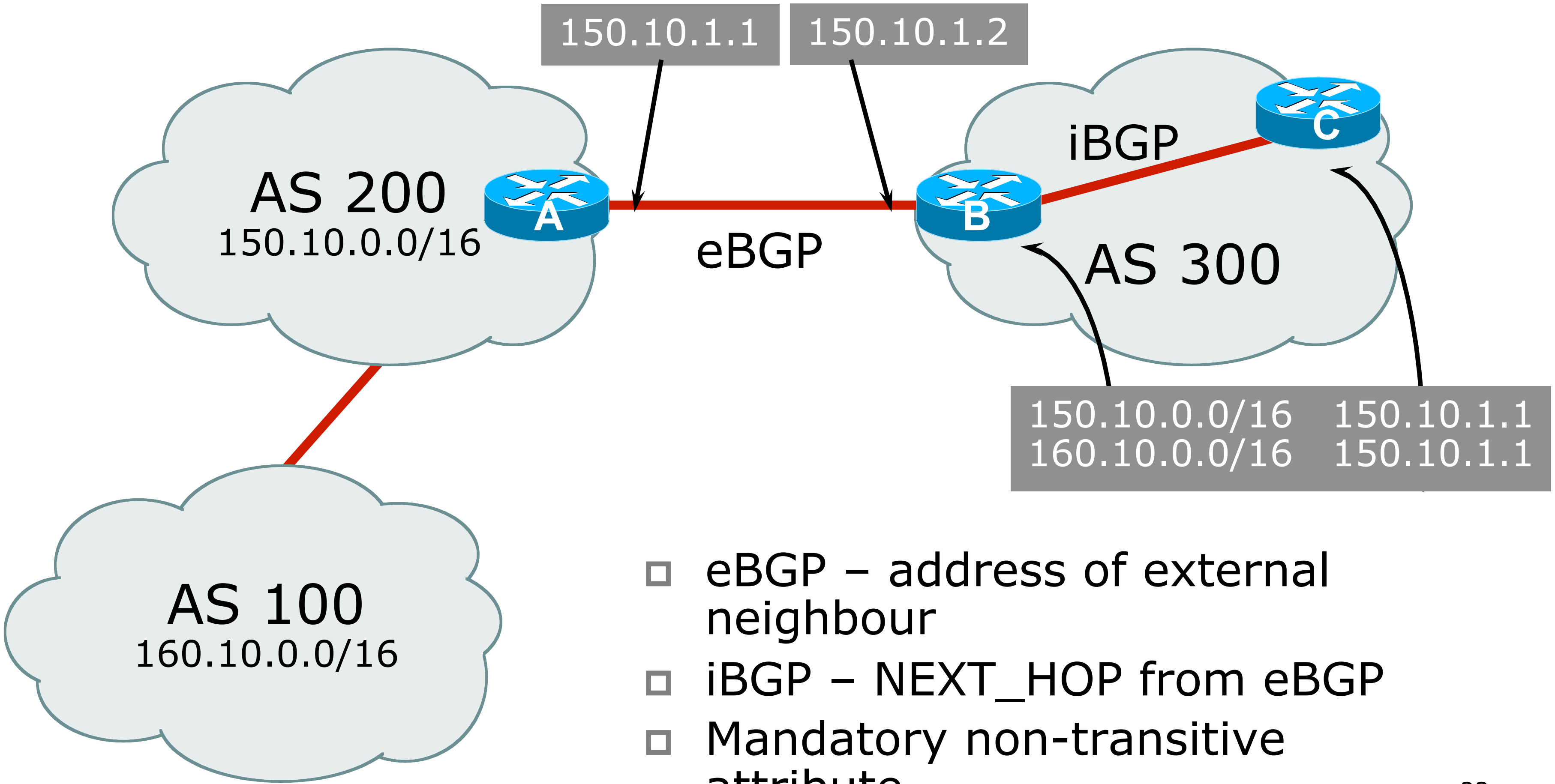


AS-Path loop detection



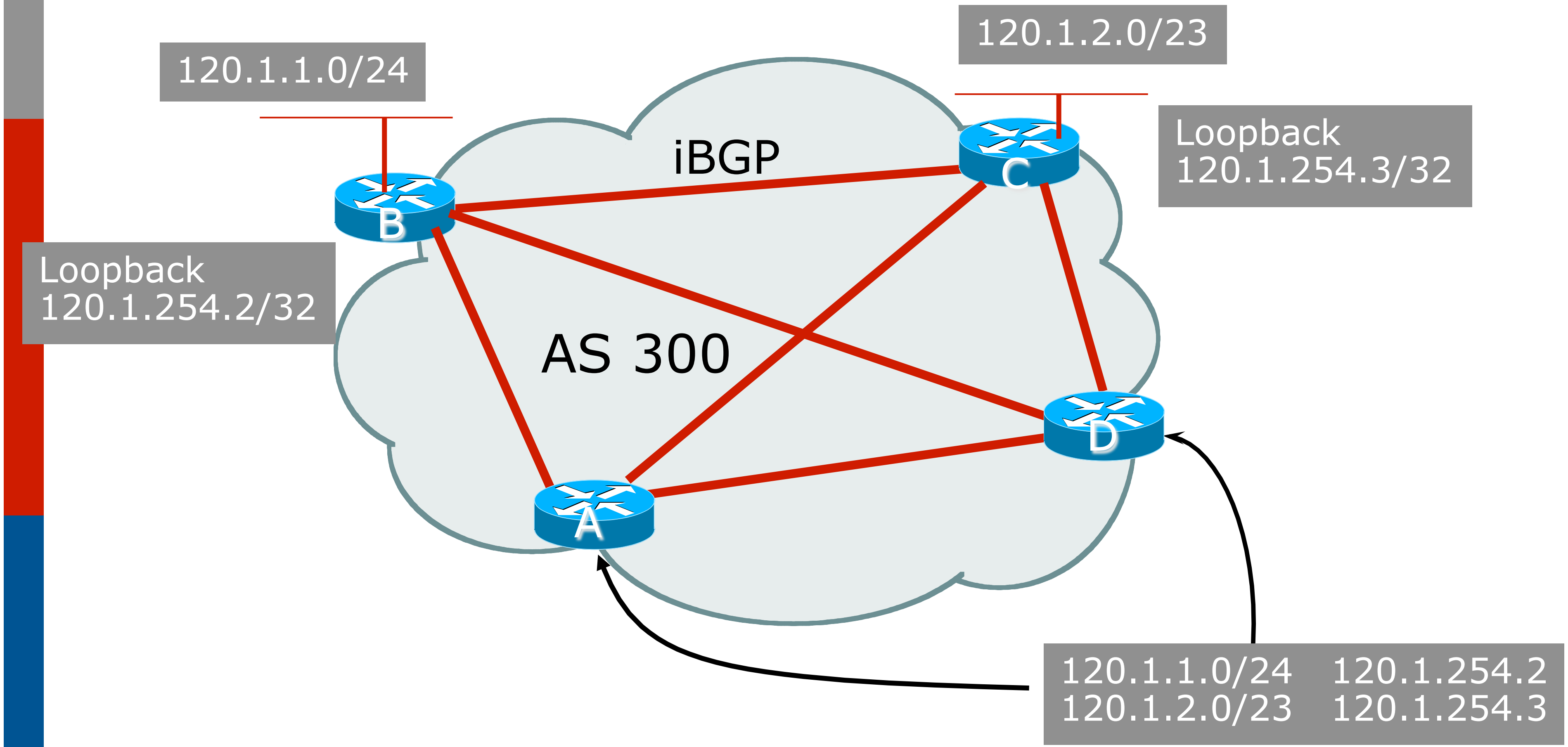
- 180.10.0.0/16 is not accepted by AS100 as the prefix has AS100 in its AS-PATH – this is loop detection in action

Next Hop



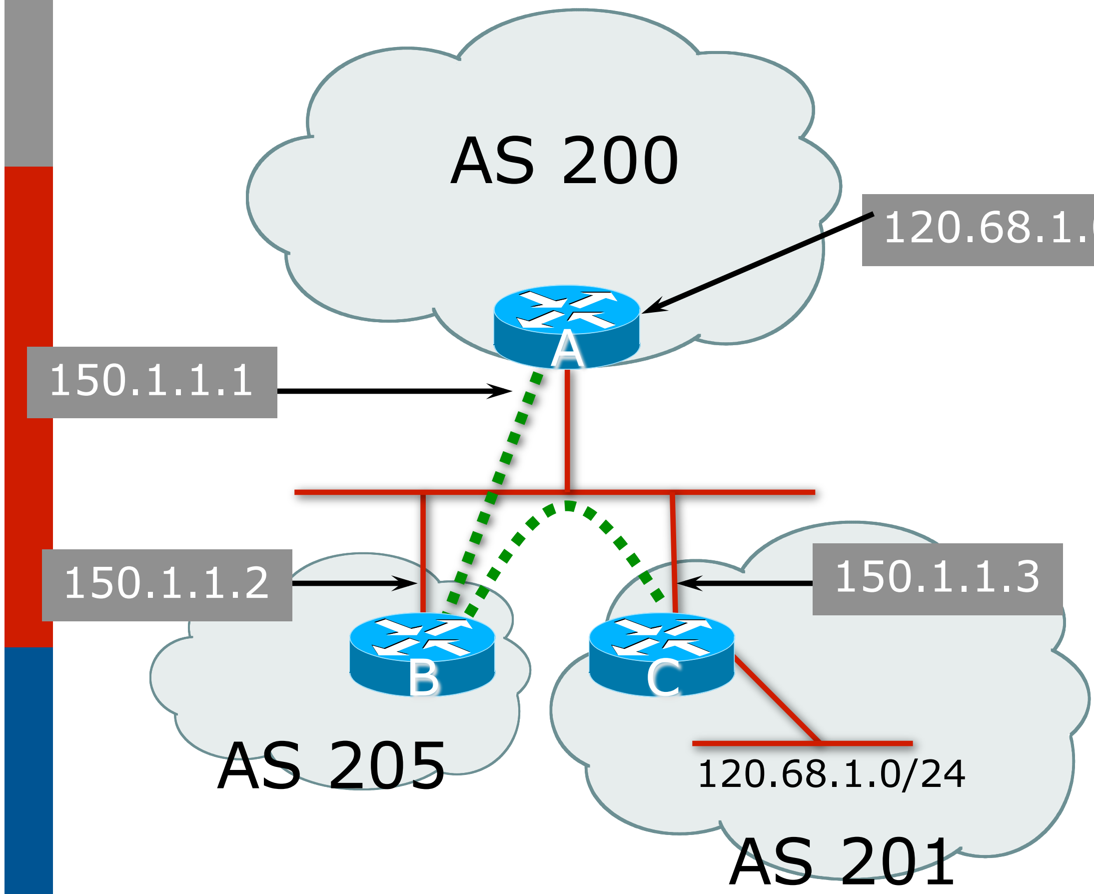
- eBGP – address of external neighbour
- iBGP – NEXT_HOP from eBGP
- Mandatory non-transitive attribute

iBGP Next Hop



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

Third Party Next Hop



- eBGP between Router A and Router B
- eBGP between Router B and Router C
- 120.68.1/24 prefix has next hop address of 150.1.1.3 – this is used by Router A instead of 150.1.1.2 as it is on same subnet as Router B
- 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

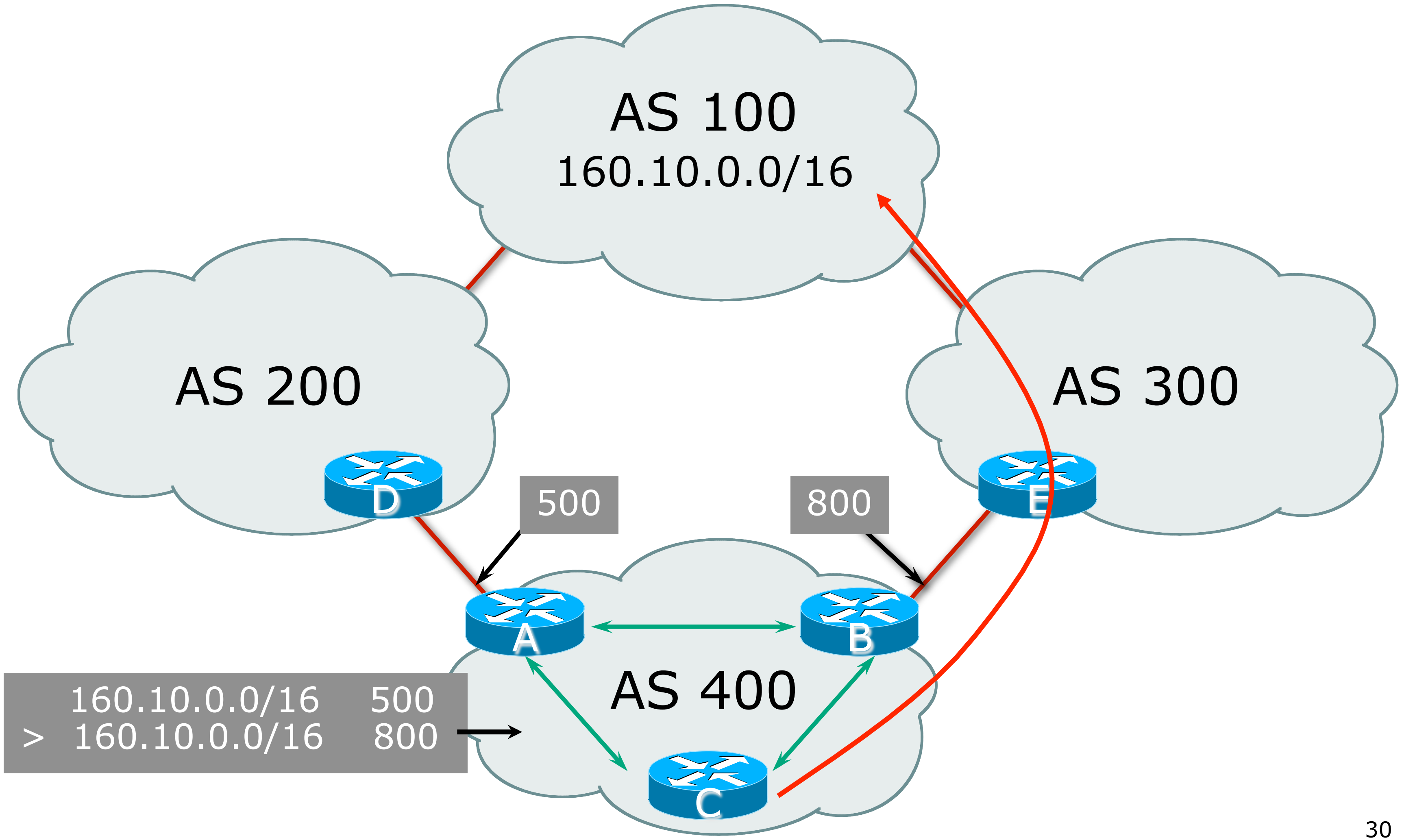
Origin

- 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

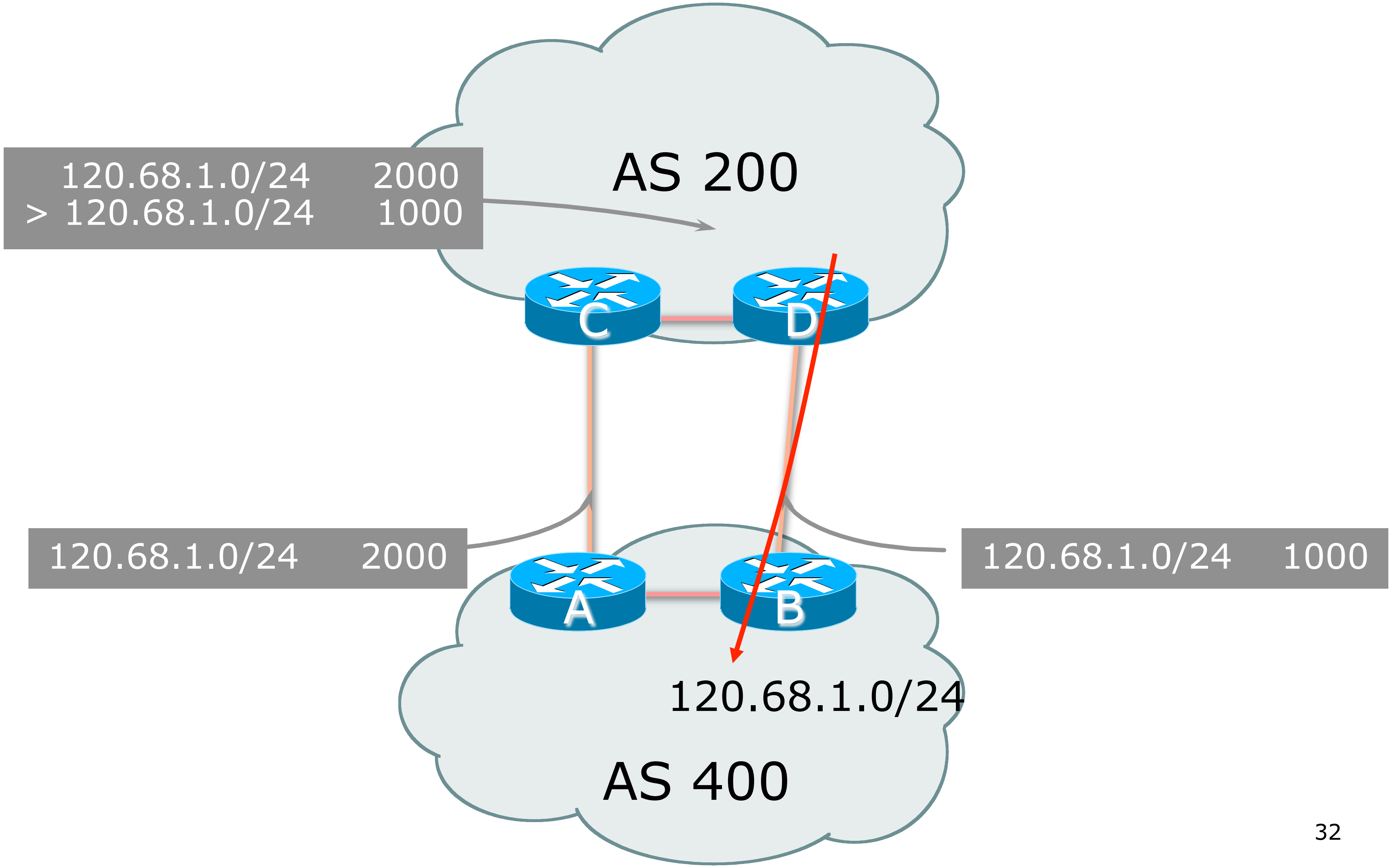
Local Preference



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)

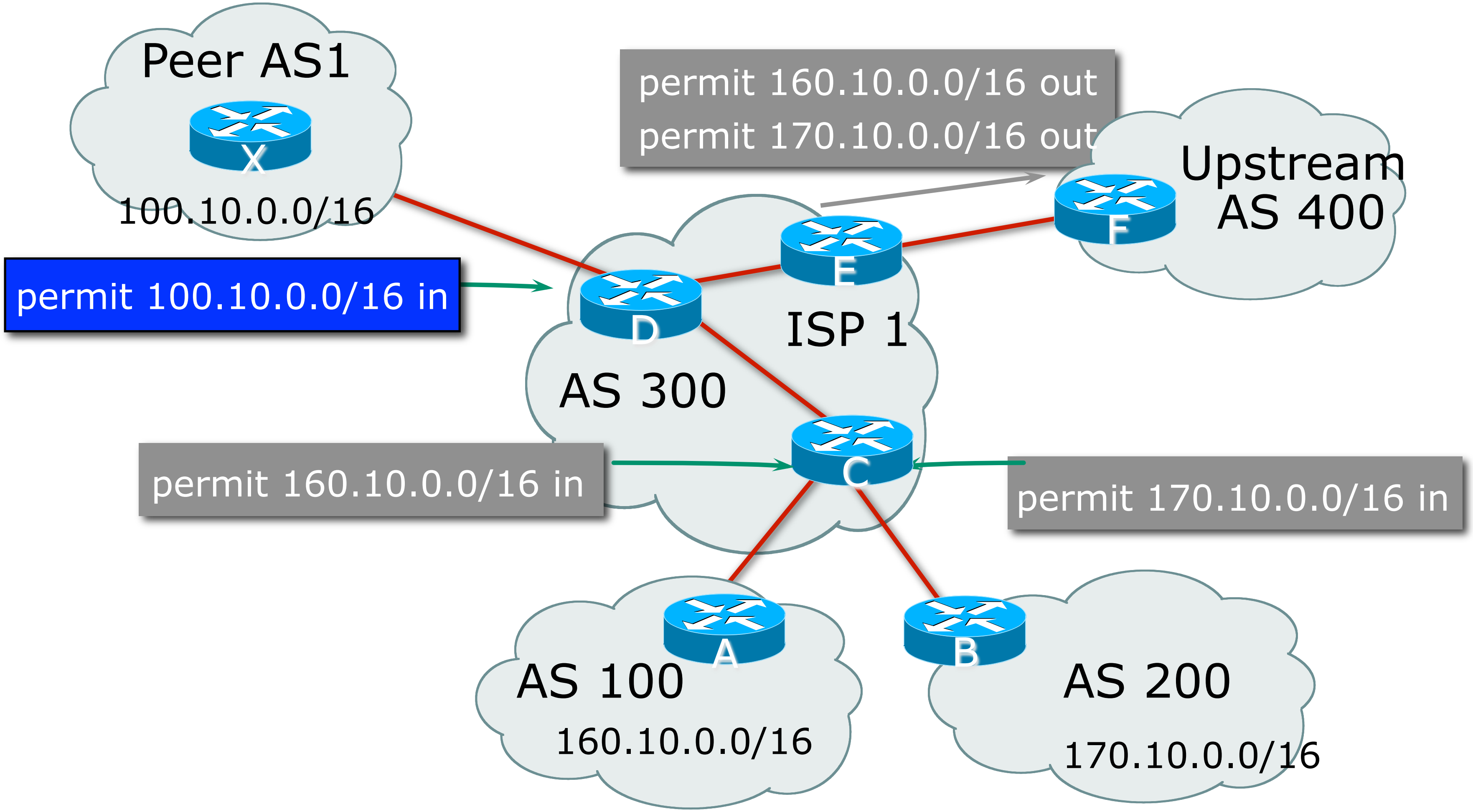
Multi-Exit Discriminator "metric confusion"

- 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"

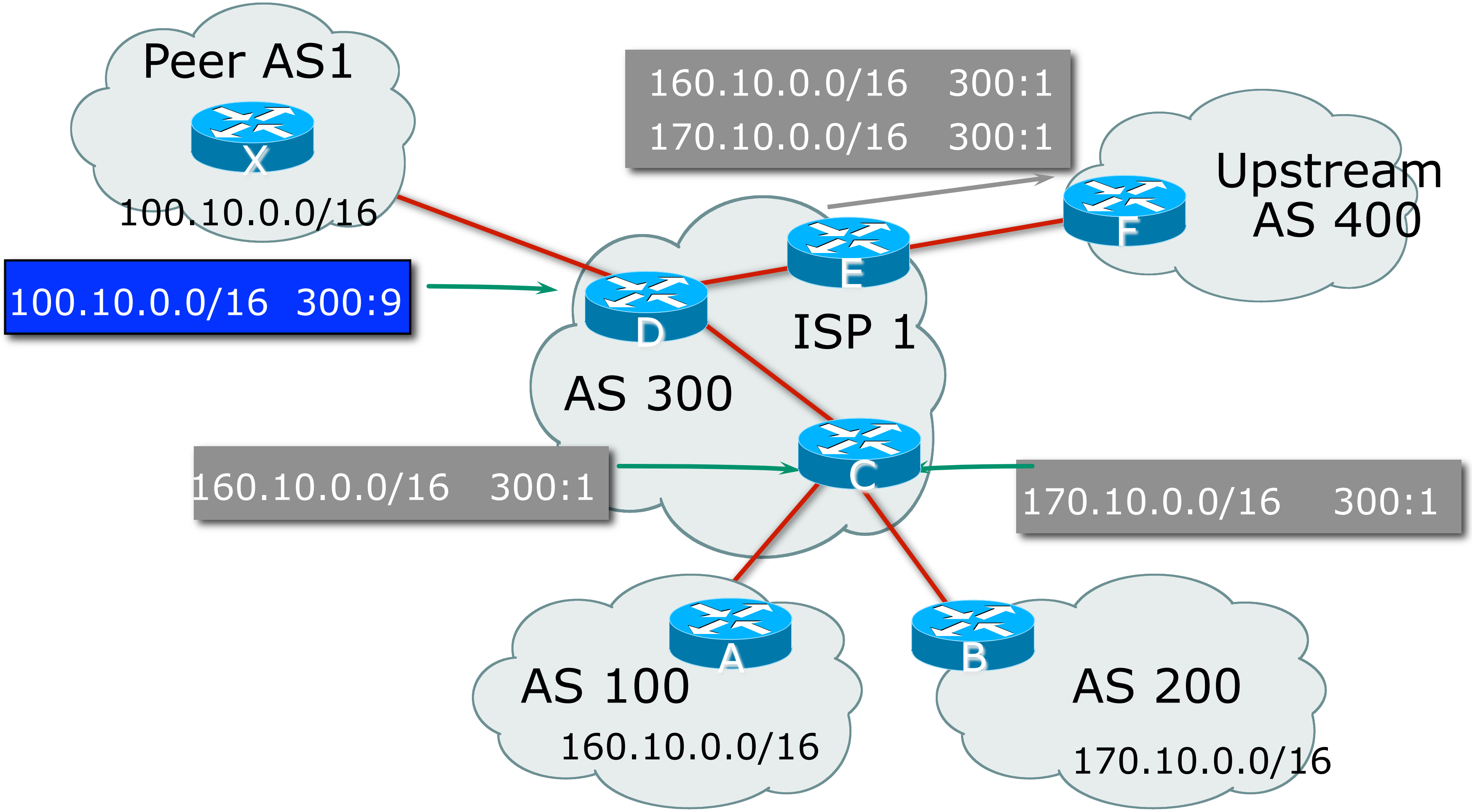
Community

- 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)



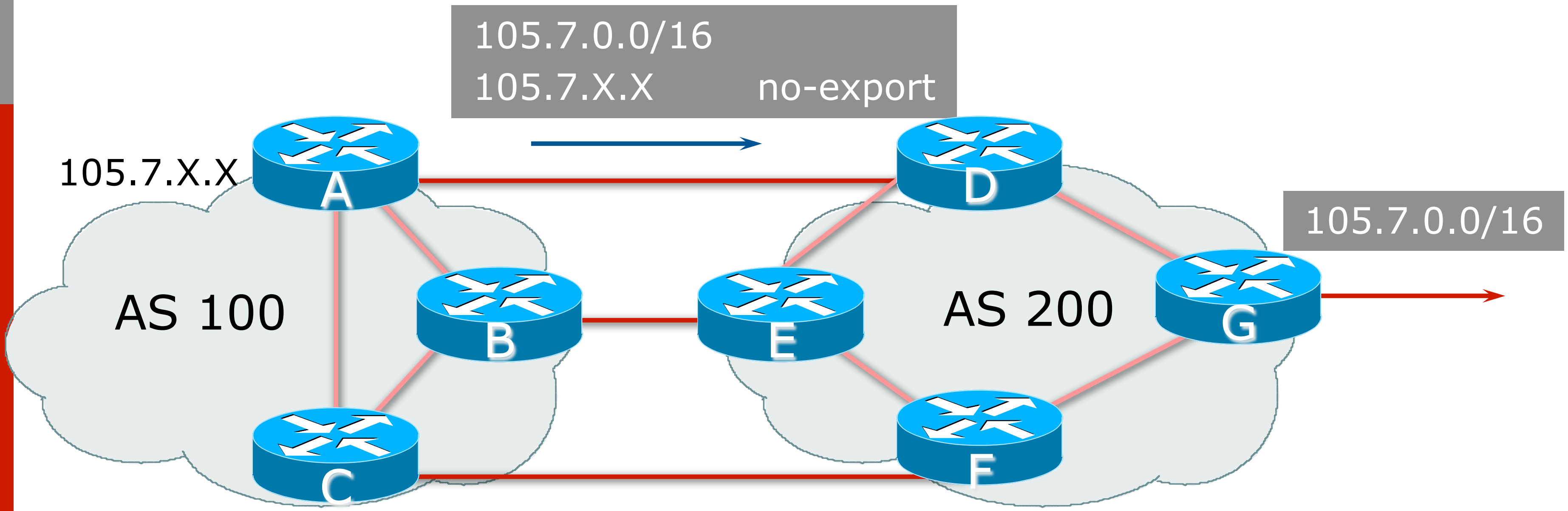
Community Example (after)



Well-Known Communities

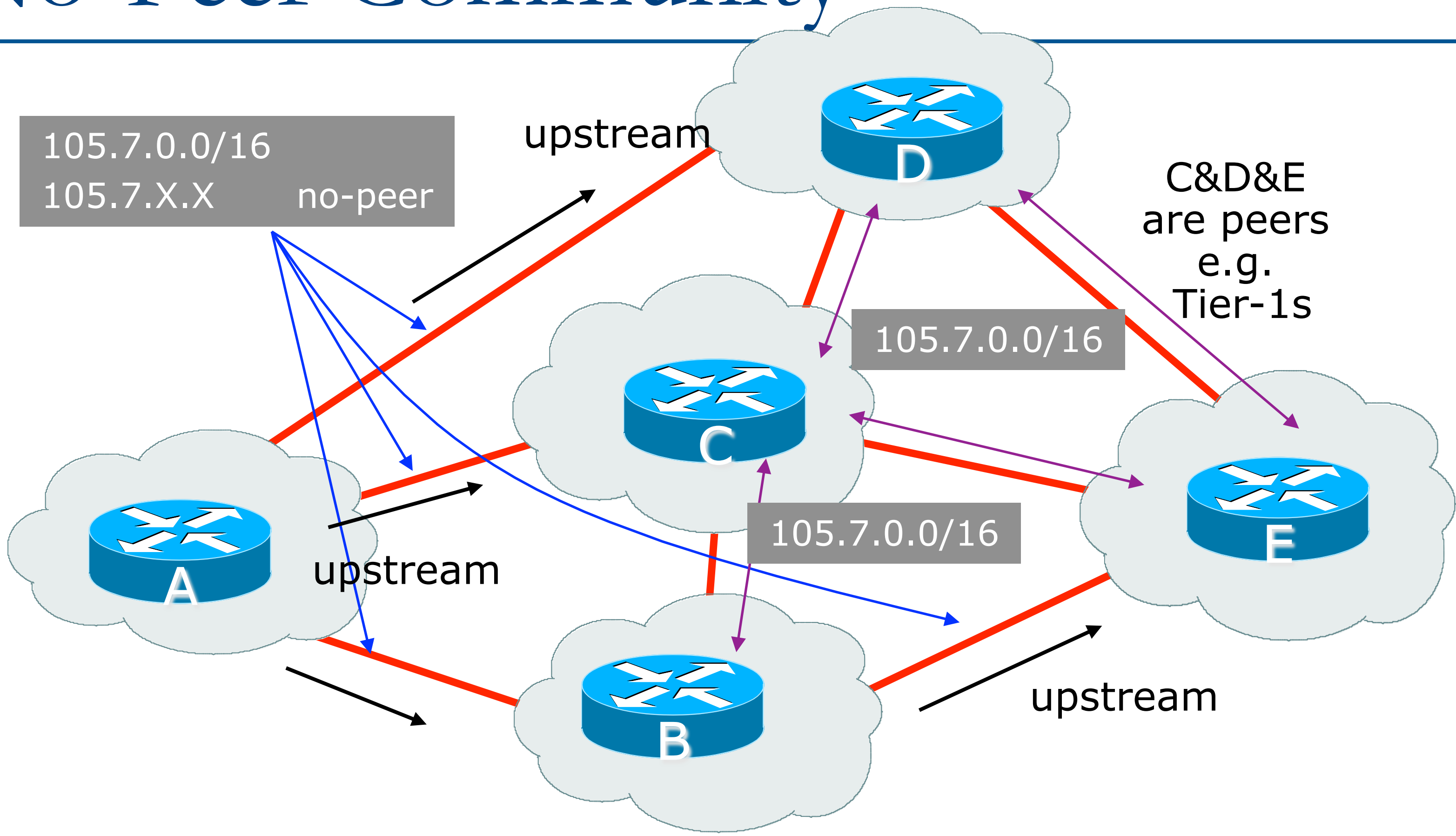
- Several well known communities
  - www.iana.org/assignments/bgp-well-known-communities
- no-export 65535:65281
  - do not advertise to any eBGP peers
- no-advertise 65535:65282
  - do not advertise to any BGP peer
- no-export-subconfed 65535:65283
  - do not advertise outside local AS (only used with 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 http://datatracker.ietf.org/doc/draft-ietf-idr-as4octet-extcomm-generic-subtype/ 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 Cisco IOS: Part One

1. Do not consider path if no route to next hop
2. Do not consider iBGP path if not synchronised (Cisco IOS)
3. Highest weight (local to router)
4. Highest local preference (global within AS)
5. Prefer locally originated route
6. Shortest AS path

BGP Path Selection Algorithm for Cisco IOS: Part Two

7. Lowest origin code
   - IGP < EGP < incomplete
8. Lowest Multi-Exit Discriminator (MED)
   - If bgp deterministic-med, order the paths by AS number before comparing
   - 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 Cisco IOS: Part Three

9. Prefer eBGP path over iBGP path
10. Path with lowest IGP metric to next-hop
11. For eBGP paths:
    - If multipath is enabled, install N parallel paths in forwarding table
    - If router-id is the same, go to next step
    - If router-id is not the same, select the oldest path

BGP Path Selection Algorithm for Cisco IOS: Part Four

12. Lowest router-id (originator-id for reflected routes)
13. Shortest cluster-list
    - Client must be aware of Route Reflector attributes!
14. Lowest neighbour 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                   [RFC6793]
66  Deprecated
67  Support for Dynamic Capability             [ID]
68  Multisession BGP                           [ID]
69  Add Path Capability                        [ID]
70  Enhanced Route Refresh 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 arrived in 2006
- 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
  - Supported on virtually all routers
  - find out from "show ip bgp neighbor"
  - Non-disruptive, "Good For the Internet"

- Only hard-reset a BGP peering as a last resort

Consider the impact to be equivalent to a router reboot

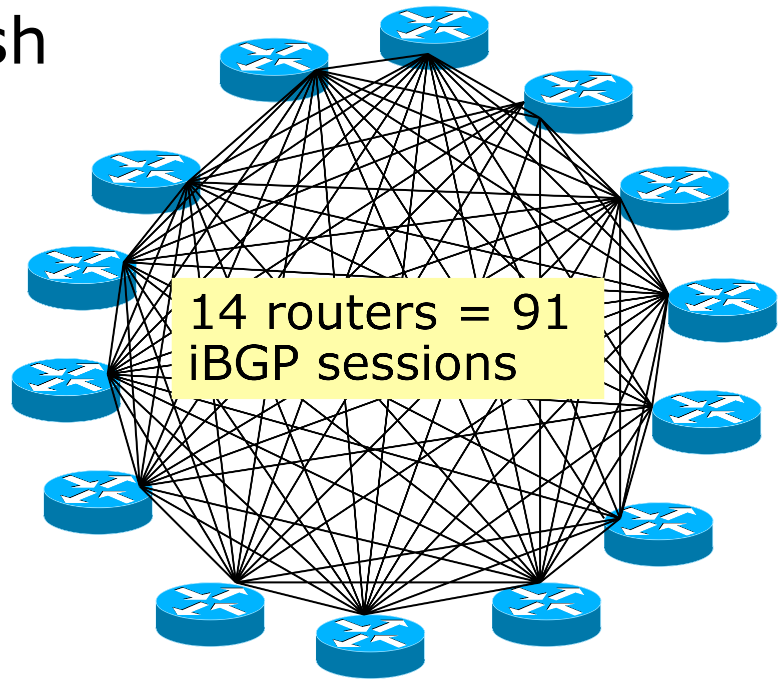
Route Reflectors

Scaling the iBGP mesh

Scaling iBGP mesh

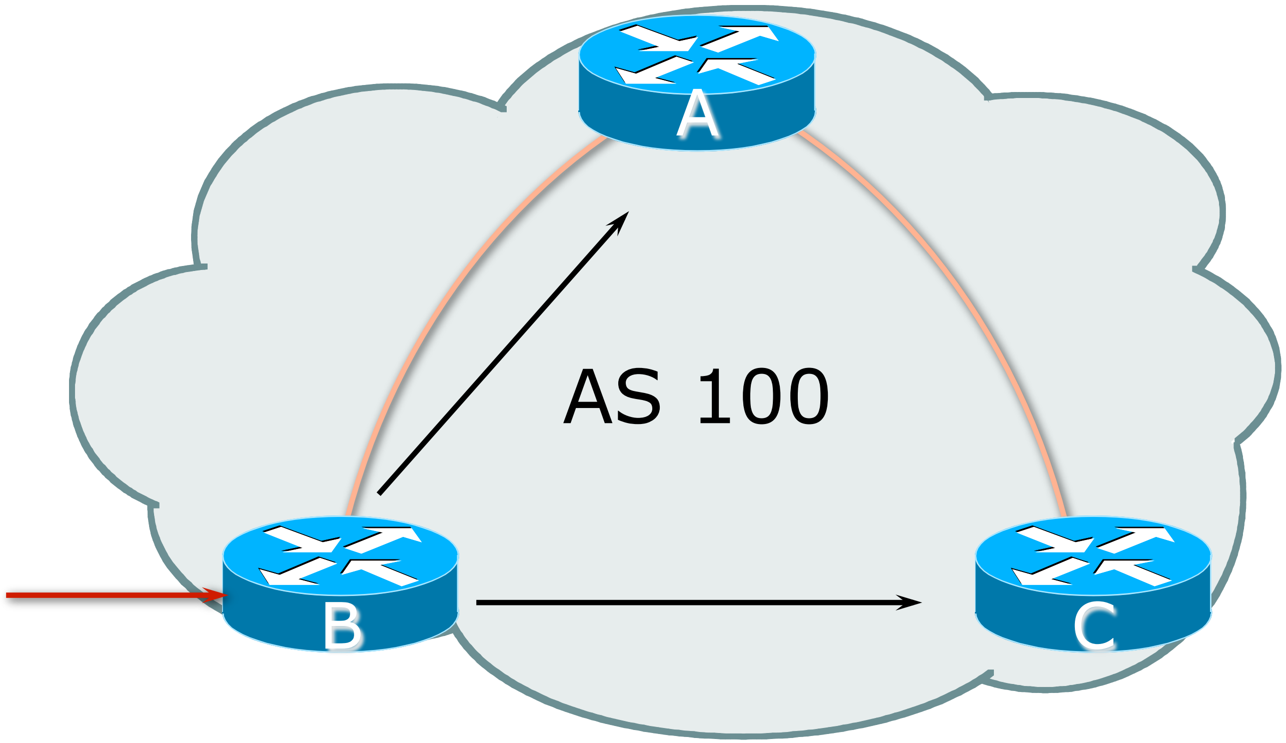
- Avoid ½n(n-1) iBGP mesh

n=1000 ⇒ nearly half a million ibgp sessions!

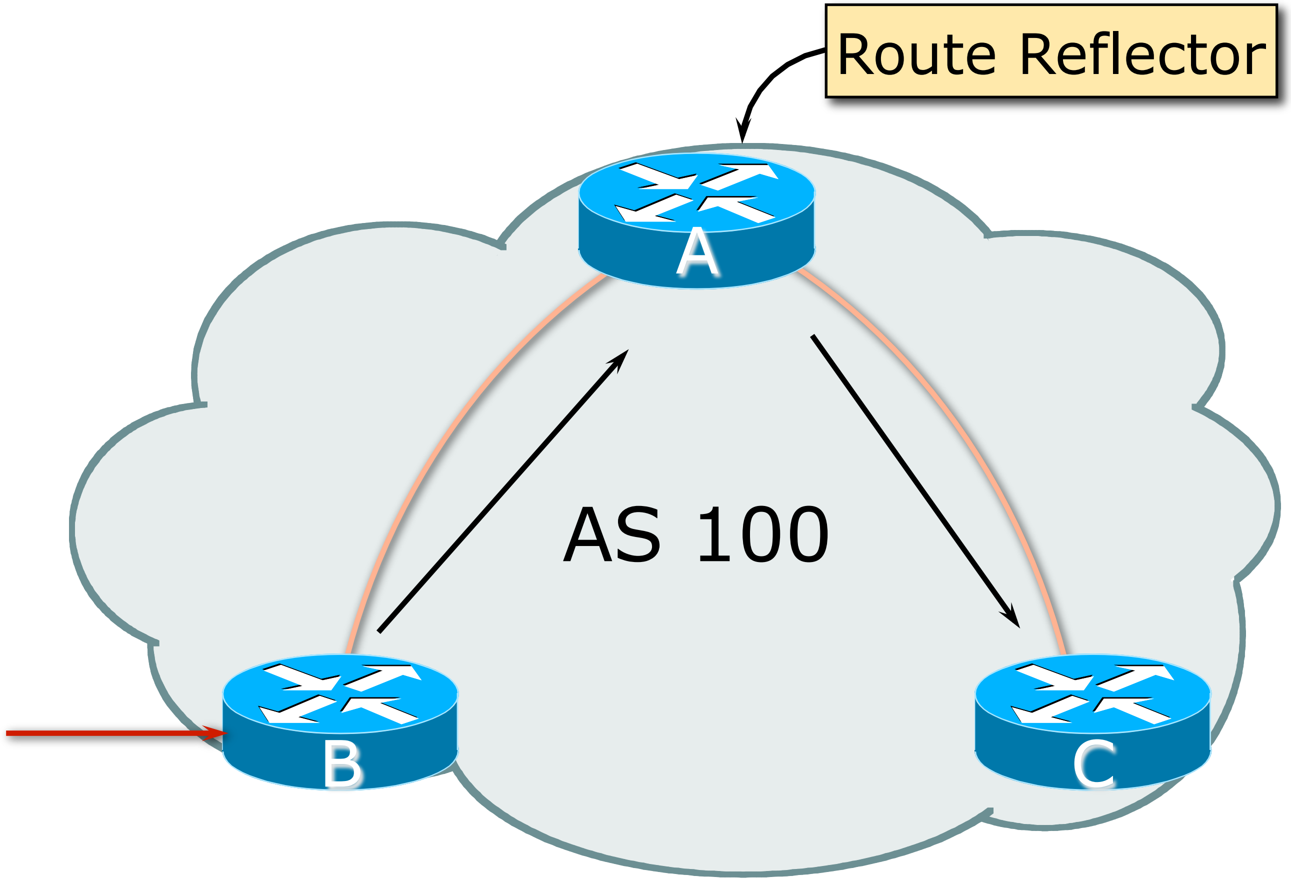


- Two solutions
  - Route reflector – simpler to deploy and run
  - Confederation – more complex, has corner case advantages

Route Reflector: Principle

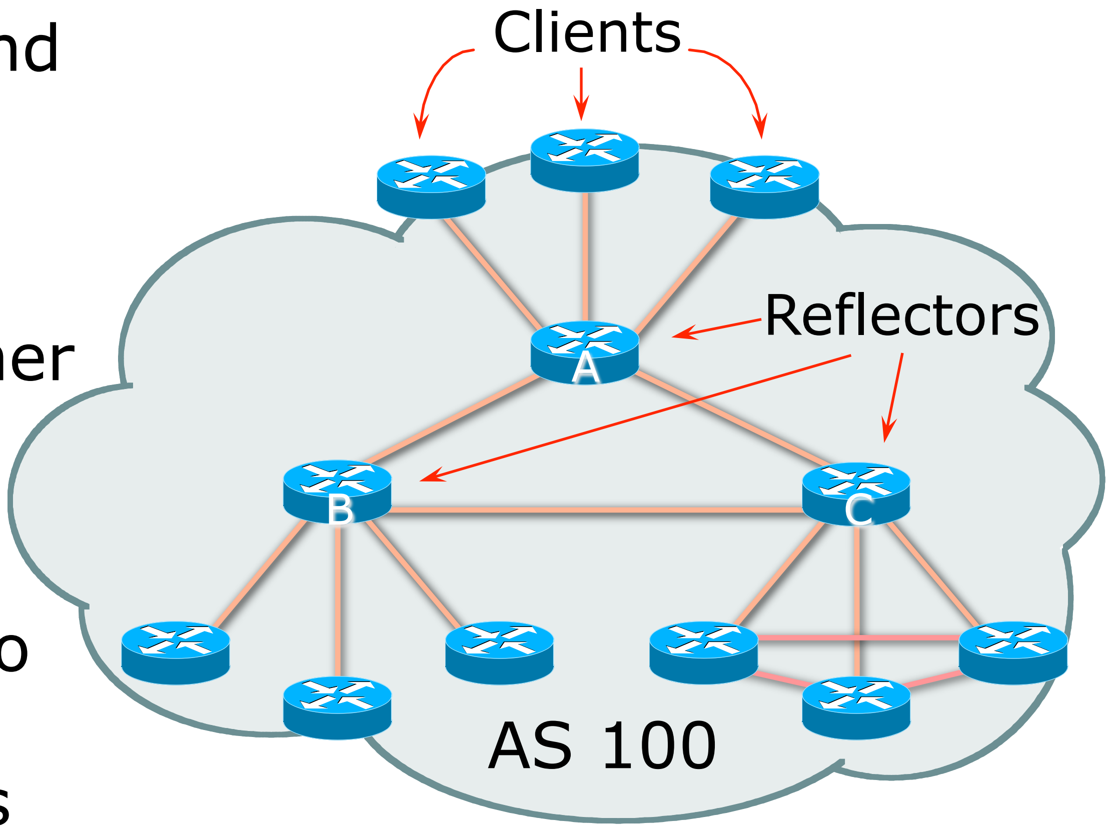


Route Reflector: Principle



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

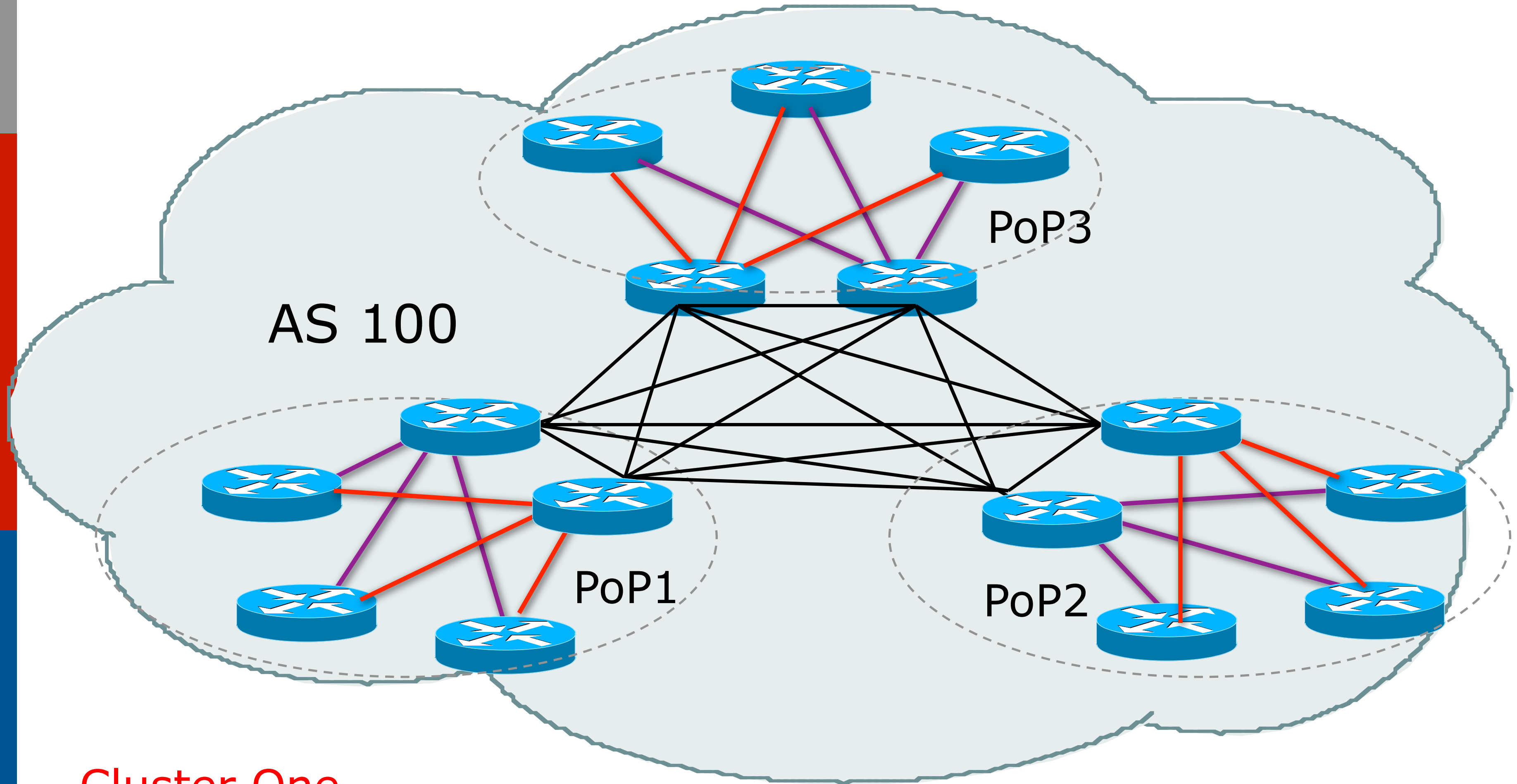
Route Reflector: Loop Avoidance

- 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

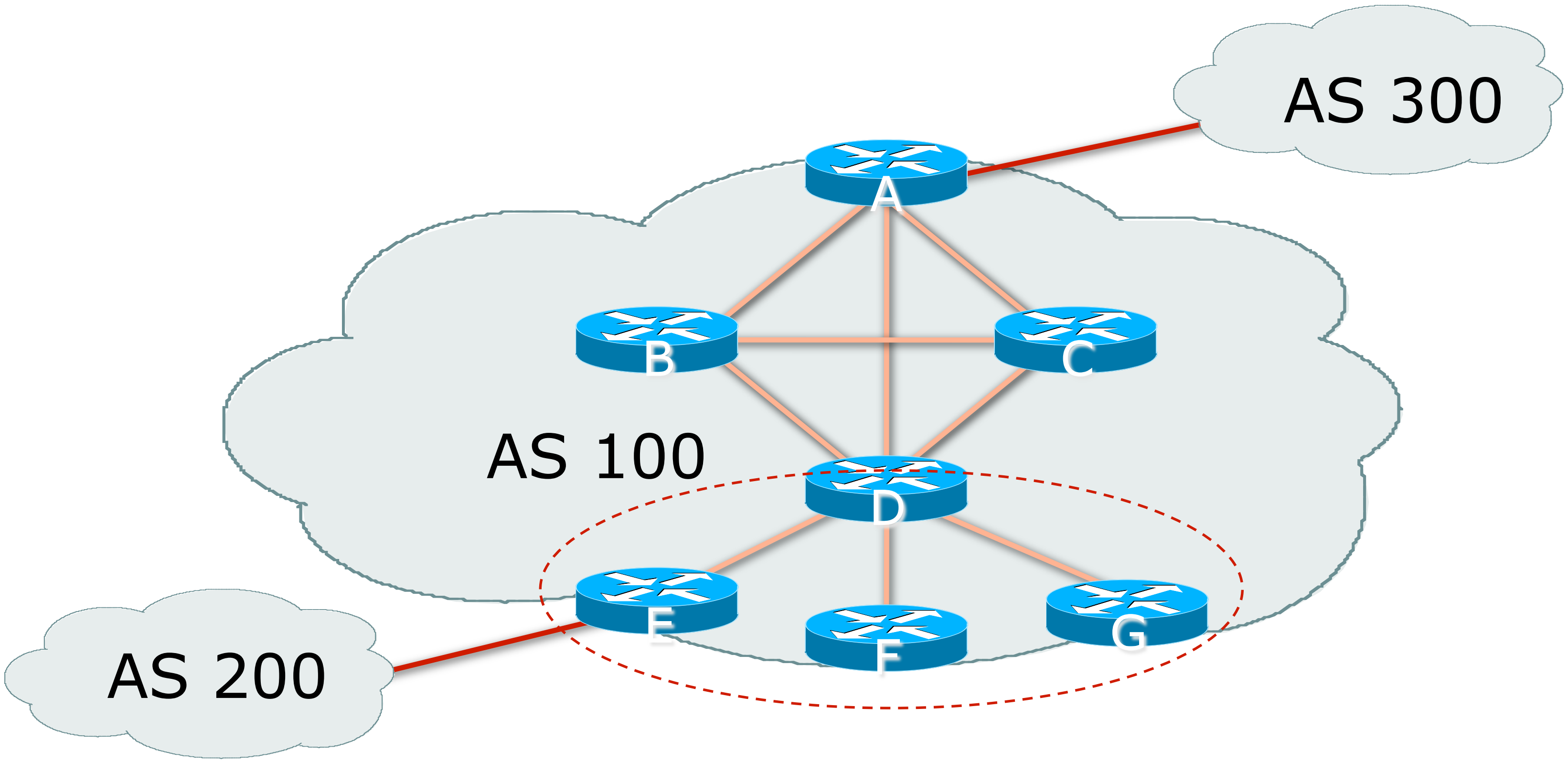
Route Reflector: Deployment

- 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 Reflectors: Migration



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

BGP Confederations

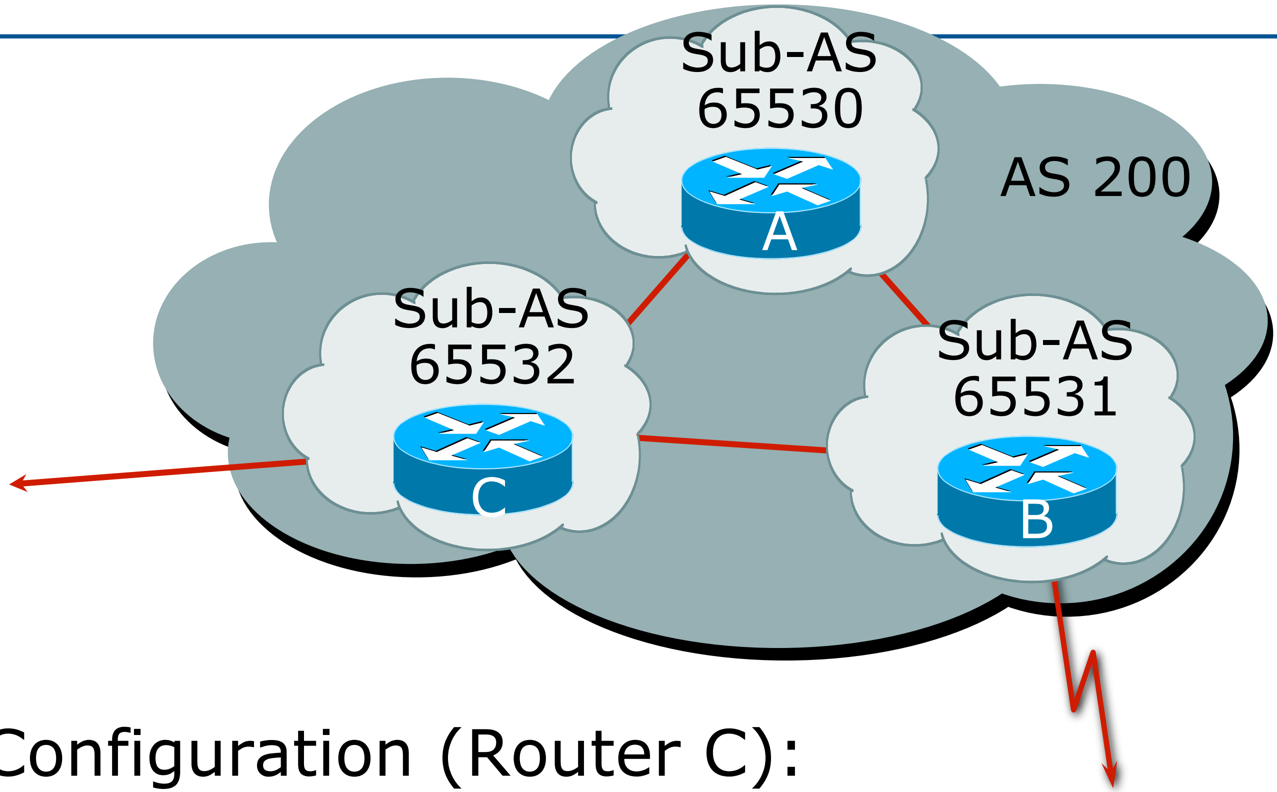
Confederations

- 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

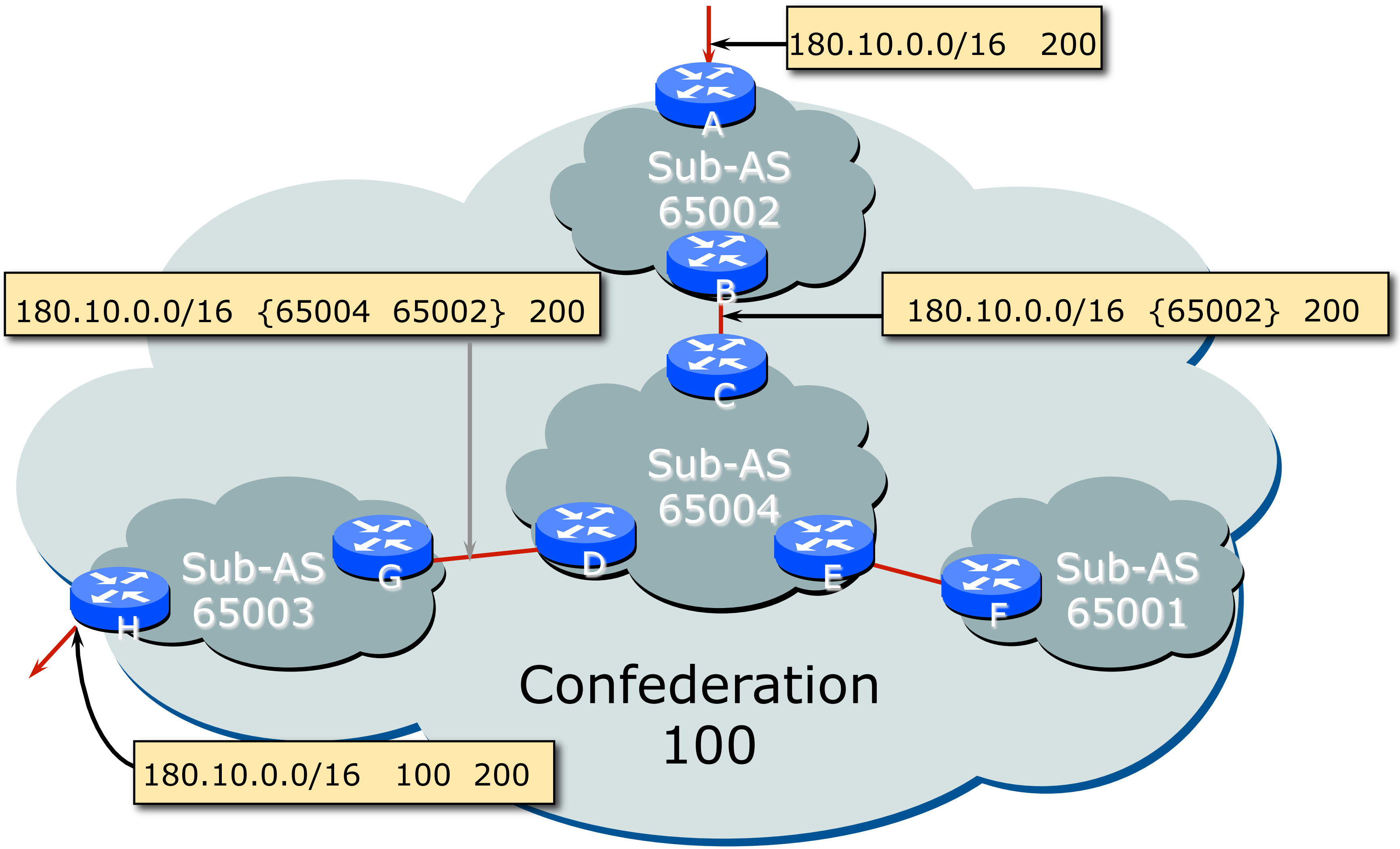
Confederations



- 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/rfc6793.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

- Nowadays:
  - Standard application process to the RIRs
  - Or via upstream provider
  - Sample RIR policy
    - www.apnic.net/docs/policy/asn-policy.html
- Bootstrap phase from 2007-2010
  - 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 (1)

- Initially three formats proposed for the 0-4294967295 ASN range :
  - asplain
  - asdot
  - asdot+
- In reality:
  - Most operators favour traditional plain format
  - 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+ !!!

Representation (2)

- Rewriting regular expressions for asdot/asdot+ notation
- 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 (1)

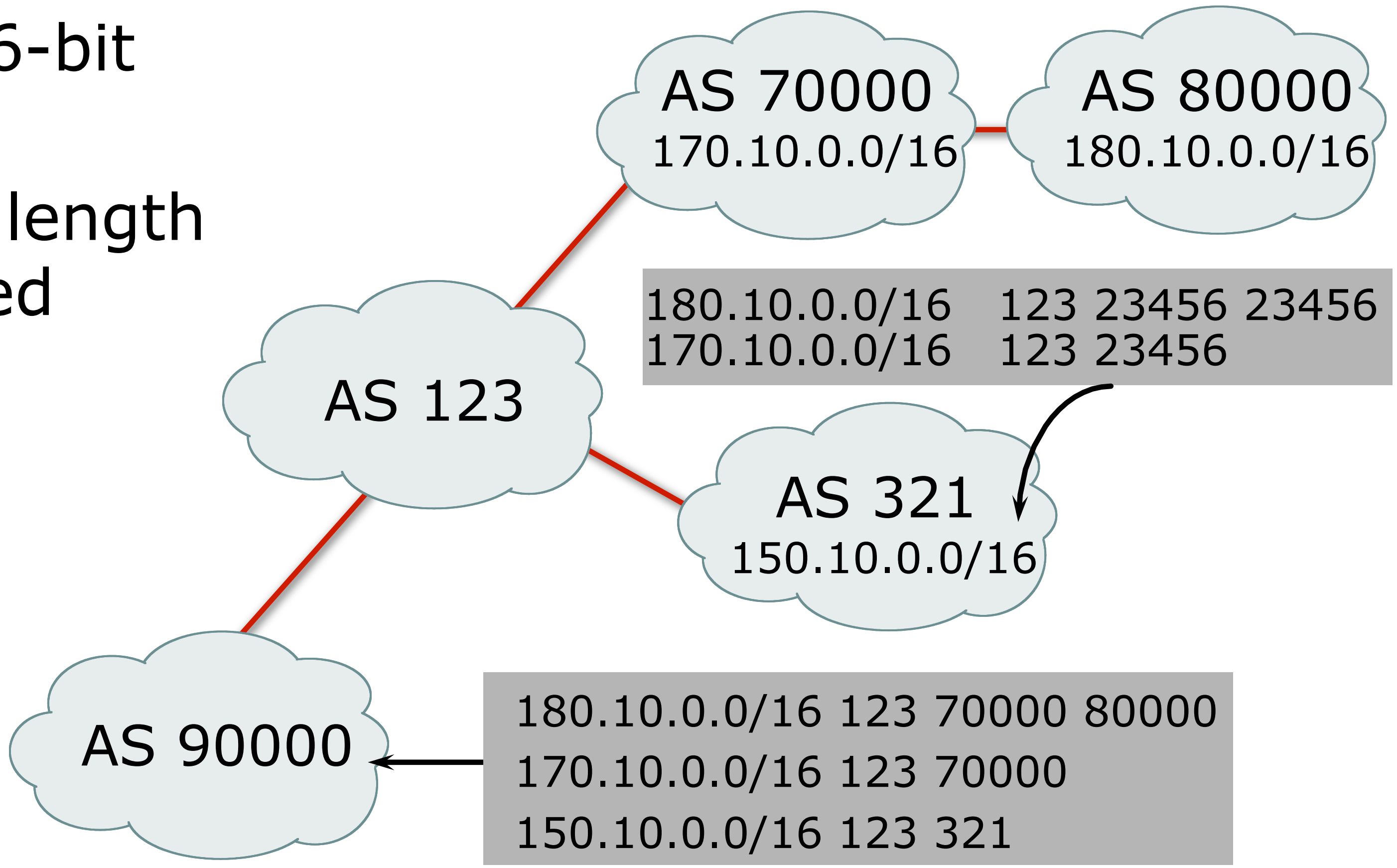
- Local router only supports 16-bit ASN and remote router uses 32-bit ASN
- BGP peering initiated:
  - Remote asks local if 32-bit supported (BGP capability negotiation)
  - When local says "no", remote then presents AS23456
  - Local needs to be configured to peer with remote using AS23456
- $\Rightarrow$ Operator of local router has to configure BGP peering with AS23456

Compatibility Mode (2)

- 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 each 32-bit ASN
  - 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

```
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
```

asplain format

- IPv4 prefix originated by AS3.5

```
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
```

asdot format

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
  - Workaround: use BGP communities instead
- 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!)**

If 32-bit ASN not supported:

- 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
- Unintended filtering by peers and upstreams:
  - Even if IRR supports 32-bit ASNs, not all tools in use can support
  - ISP may not support 32-bit ASNs which are in the IRR – and don't realise that AS23456 is the transition AS

Implementations (May 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

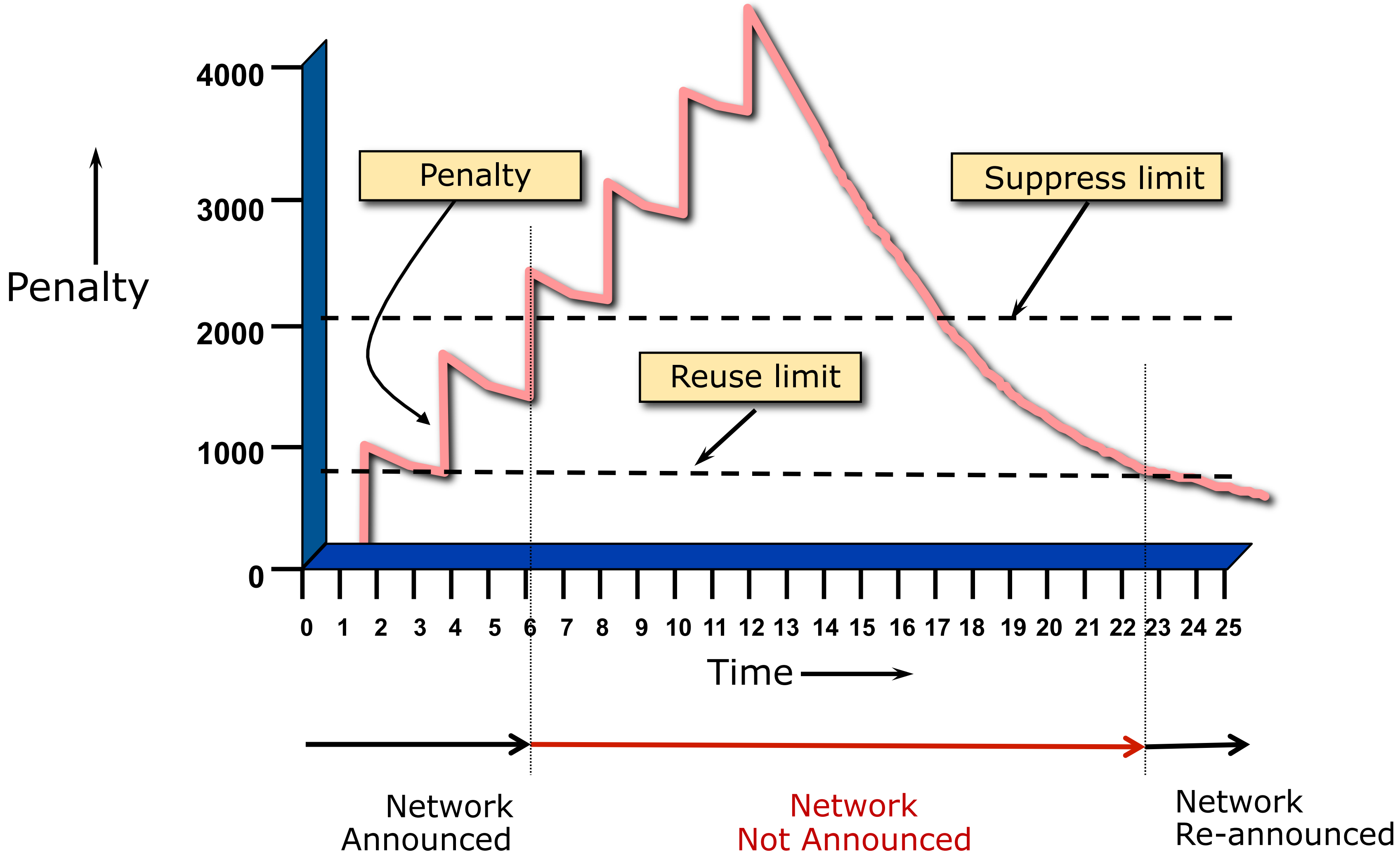
Route Flap Damping (continued)

- 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:

- Misconfigured Route Flap Damping will seriously impact access to:
  - Your network and
  - The Internet
- More background contained in RIPE Routing Working Group document:
  - www.ripe.net/ripe/docs/ripe-378
- Recommendations now in:
  - www.ripe.net/ripe/docs/ripe-580

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

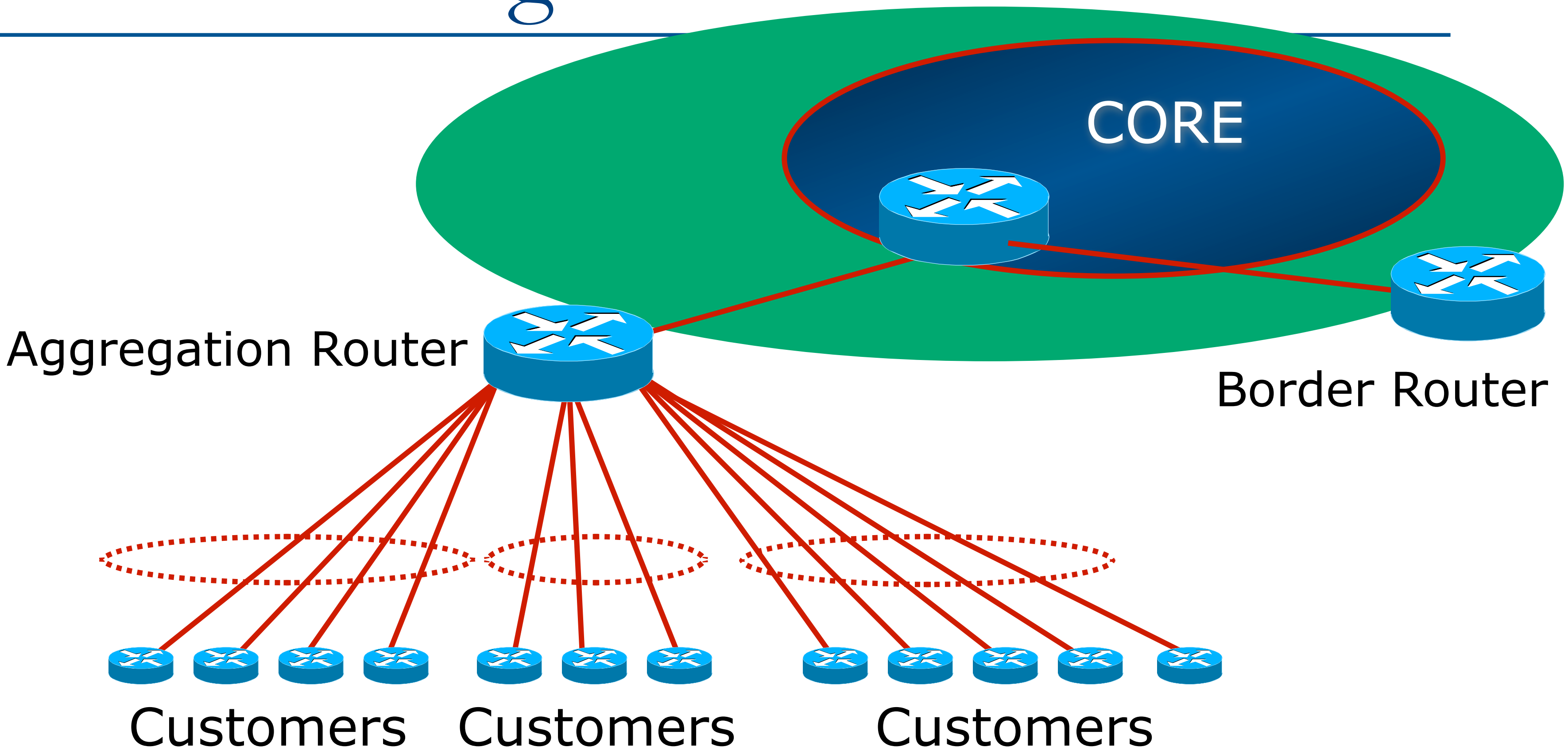
Community Example: Customer Edge

- 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
  - $\Rightarrow$ 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 the full BGP table!

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 notifying 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.

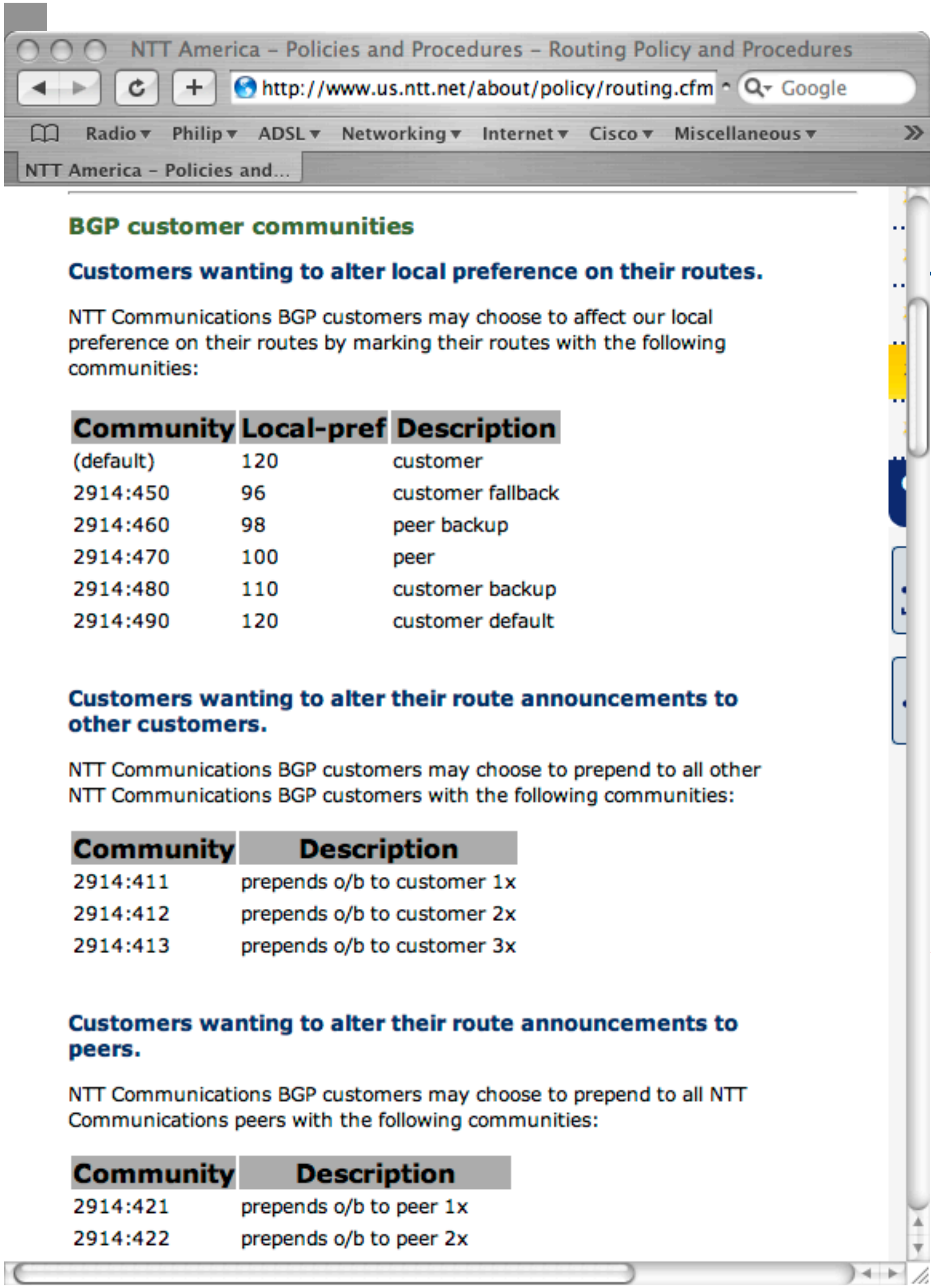
String	Resulting AS Path to ASXXX
65000:XXX	Do not advertise to ASXXX
65001:XXX	1239 (default) ...
65002:XXX	1239 1239 ...
65003:XXX	1239 1239 1239 ...
65004:XXX	1239 1239 1239 1239 ...

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

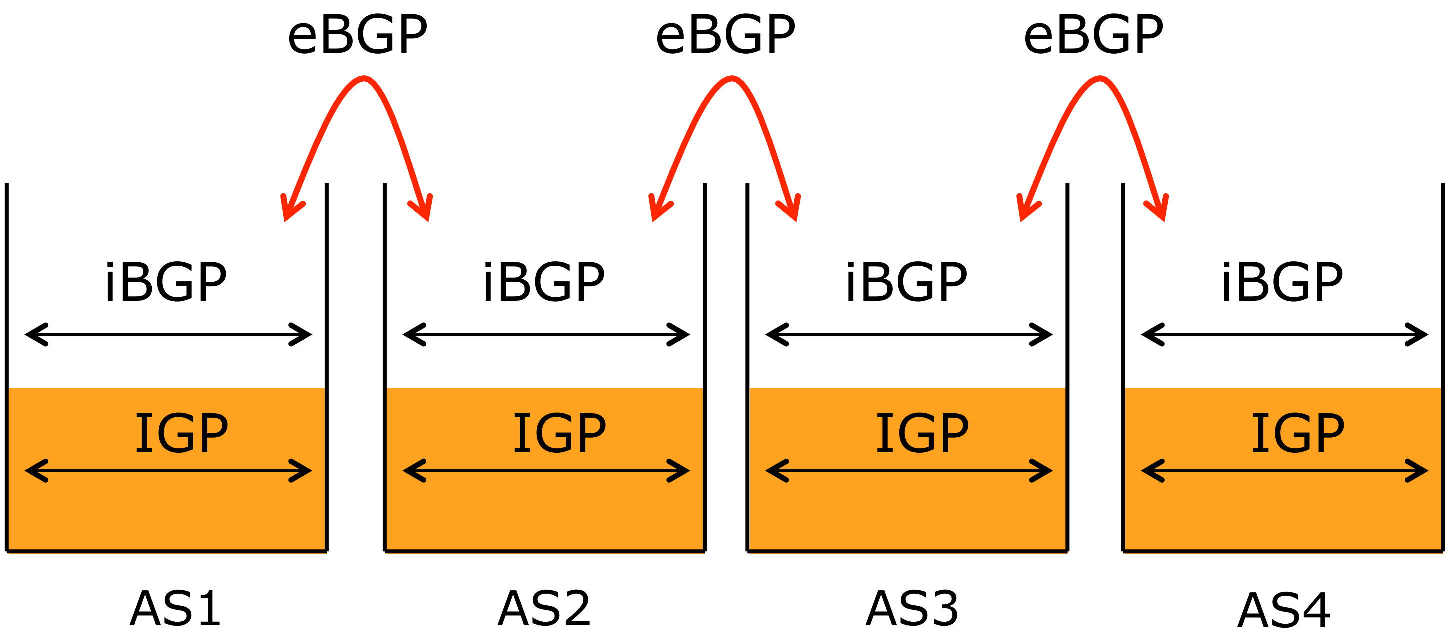
- Internal Routing Protocols (IGPs)
  - examples are ISIS and OSPF
  - used for carrying infrastructure addresses
  - NOT used for carrying Internet prefixes or customer prefixes
  - design goal is to minimise number of prefixes in IGP to aid scalability and rapid convergence

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

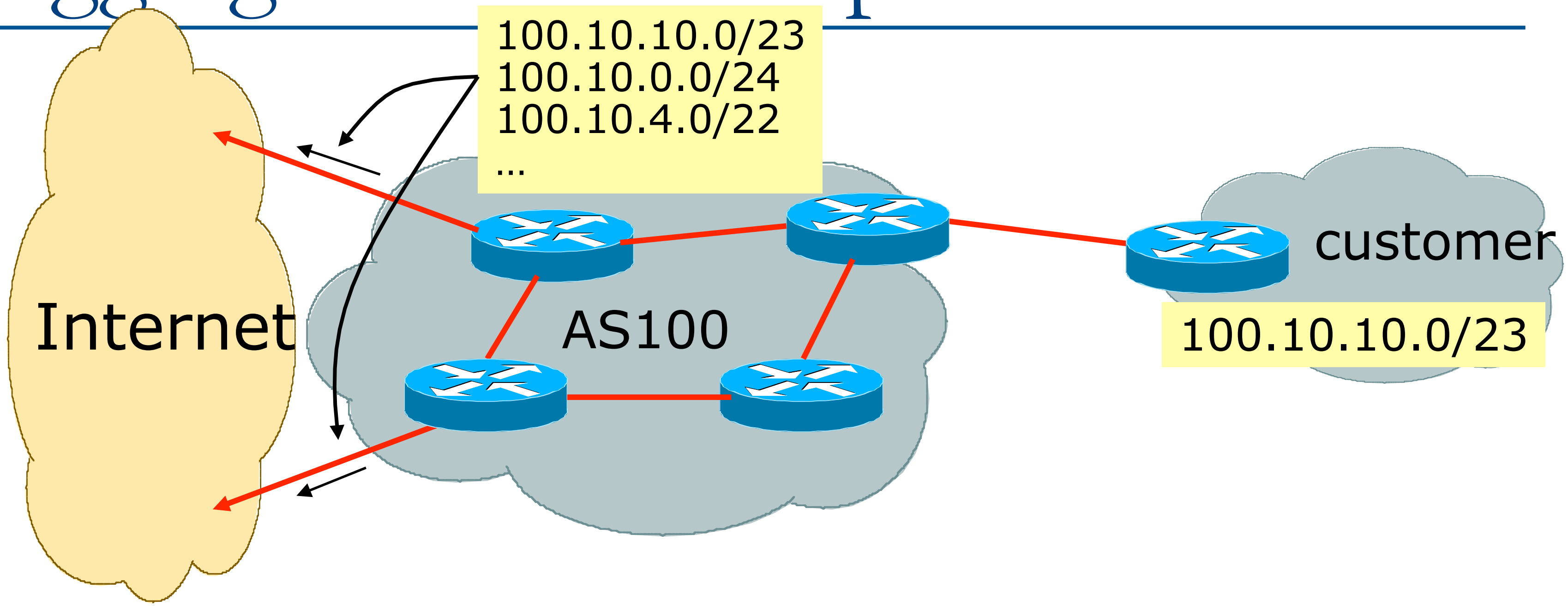
Aggregation

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
  - Jan 2014: 253000 /24s in IPv4 table of 477720 prefixes
- The same is happening for /48s with IPv6
  - Jan 2014: 6344 /48s in IPv6 table of 15577 prefixes

Aggregation – Example

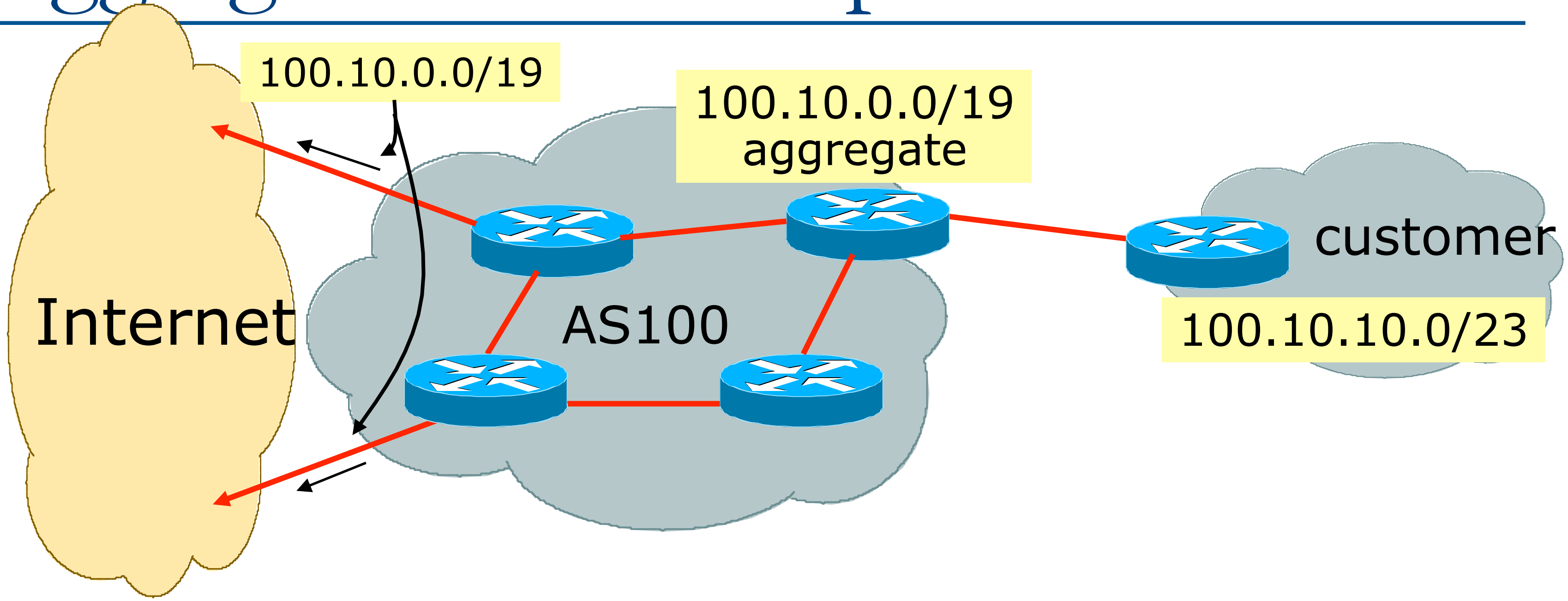


- Customer has /23 network assigned from AS100's /19 address block
- AS100 announces customers' individual networks to the Internet

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 (January 2014)

- Current Internet Routing Table Statistics
  - BGP Routing Table Entries 477720
  - Prefixes after maximum aggregation 190383
  - Unique prefixes in Internet 236793
  - Prefixes smaller than registry alloc 166265
  - /24s announced 253304
  - ASes in use 45866

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
    - (covers both IPv4 and IPv6 BGP tables)
  - 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 recommendations
  - IPv4: RIPE-399 — www.ripe.net/ripe/docs/ripe-399.html
  - IPv6: RIPE-532 — www.ripe.net/ripe/docs/ripe-532.html

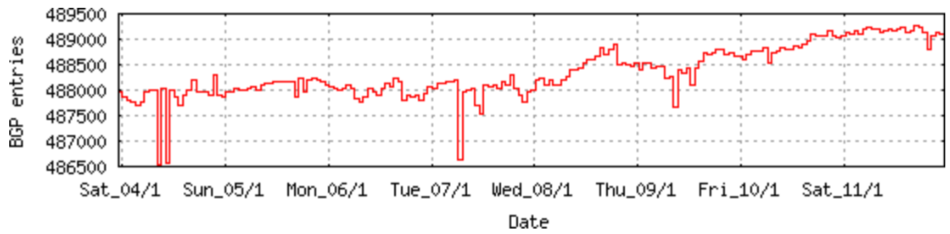
Efforts to Improve Aggregation The CIDR Report

- Also computes the size of the routing table assuming ISPs performed optimal aggregation
- Website allows searches and computations of aggregation to be made on a per AS basis
  - Flexible and powerful tool to aid ISPs
  - Intended to show how greater efficiency in terms of BGP table size can be obtained without loss of routing and policy information
  - Shows what forms of origin AS aggregation could be performed and the potential benefit of such actions to the total table size
  - Very effectively challenges the traffic engineering excuse

Status Summary

Table History

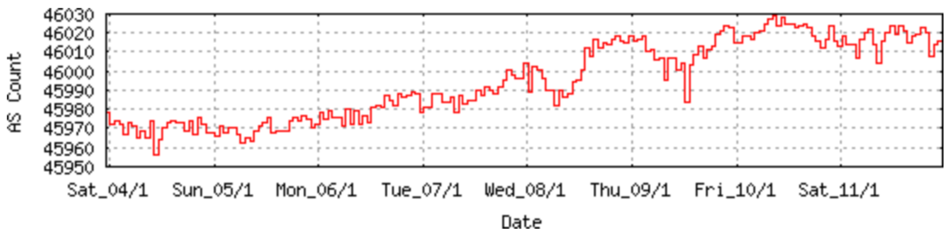
Date	Prefixes	CIDR Aggregated
04-01-14	487980	276353
05-01-14	487866	276635
06-01-14	488108	276237
07-01-14	488074	276227
08-01-14	488015	276657
09-01-14	488520	276469
10-01-14	488669	276220
11-01-14	489052	275931



Plot: BGP Table Size

AS Summary

46007	Number of ASes in routing system
18872	Number of ASes announcing only one prefix
4426	Largest number of prefixes announced by an AS AS7029: WINDSTREAM - Windstream Communications Inc
118834176	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 transit policies. Aggregation is also proposed across non-advertised address space ('holes').

--- 11Jan14 ---

ASnum	NetsNow	NetsAggr	NetGain	% Gain	Description
Table	489033	275870	213163	43.6%	All ASes
AS28573	3441	95	3346	97.2%	NET Serviços de Comunicação S.A.
AS6389	3029	56	2973	98.2%	BELLSOUTH-NET-BLK - BellSouth.net Inc.
AS7029	4426	1657	2769	62.6%	WINDSTREAM - Windstream Communications Inc
AS17974	2736	185	2551	93.2%	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
AS4766	2944	962	1982	67.3%	KIXS-AS-KR Korea Telecom
AS18881	1795	31	1764	98.3%	Global Village Telecom
AS36998	1805	47	1758	97.4%	SDN-MOBITEL
AS1785	2148	397	1751	81.5%	AS-PAETEC-NET - PaeTec Communications, Inc.
AS10620	2696	1078	1618	60.0%	Telmex Colombia S.A.
AS22773	2326	827	1499	64.4%	ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
AS18566	2048	565	1483	72.4%	MEGAPATH5-US - MegaPath Corporation
AS4323	2932	1506	1426	48.6%	TWTC - tw telecom holdings, inc.
AS7303	1743	457	1286	73.8%	Telecom Argentina S.A.
AS4755	1810	593	1217	67.2%	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
AS7552	1259	157	1102	87.5%	VIETEL-AS-AP Viettel Corporation
AS22561	1258	226	1032	82.0%	AS22561 - CenturyTel Internet Holdings, Inc.
AS9829	1574	654	920	58.4%	BSNL-NIB National Internet Backbone
AS7545	2132	1309	823	38.6%	TPG-INTERNET-AP TPG Telecom Limited
AS18101	988	185	803	81.3%	RELIANCE-COMMUNICATIONS-IN Reliance Communications Ltd.DAKC MUMBAI
AS35908	901	103	798	88.6%	VPLSNET - Krypt Technologies
AS4808	1146	383	763	66.6%	CHINA169-BJ CNCGROUP IP network China169 Beijing Province Network
AS701	1507	776	731	48.5%	UUNET - MCI Communications Services, Inc. d/b/a Verizon Business
AS8151	1385	664	721	52.1%	Uninet S.A. de C.V.
AS24560	1089	368	721	66.2%	AIRTELBROADBAND-AS-AP Bharti Airtel Ltd., Telemedia Services
AS6983	1295	581	714	55.1%	ITCDELTA - ITC^Deltacom
AS13977	848	143	705	83.1%	CTELCO - FAIRPOINT COMMUNICATIONS, INC.

Top 20 Added Routes this week per Originating AS

Prefixes	ASnum	AS Description
232	AS7029	WINDSTREAM - Windstream Communications Inc
174	AS4	ISI-AS - University of Southern California
118	AS13188	BANKINFORM-AS TOV "Bank-Inform"
87	AS8402	CORBINA-AS OJSC "Vimpelcom"
86	AS18809	Cable Onda
68	AS2	UDEL-DCN - University of Delaware
66	AS9198	KAZTELECOM-AS JSC Kazakhtelecom
58	AS13370	LOCALTEL - LocalTel Communications
49	AS18001	DIALOG-AS Dialog Axiata PLC.
47	AS3	MIT-GATEWAYS - Massachusetts Institute of Technology
45	AS7602	SPT-AS-VN Saigon Postel Corporation
45	AS6	BULL-NETWORK for further information please visit http://www.bull.com
43	AS28554	Cablemas Telecomunicaciones SA de CV
36	AS9808	CMNET-GD Guangdong Mobile Communication Co.Ltd.
34	AS3801	MISNET - Mikrotec Internet Services
30	AS13783	RADFORD-UNIV-AS - Radford University
28	AS1452	DNIC-ASBLK-01451-01456 - Headquarters, USAISC
28	AS18403	FPT-AS-AP The Corporation for Financing & Promoting Technology
27	AS36954	MLTL-AS
25	AS28573	NET Serviços de Comunicação S.A.

Top 20 Withdrawn Routes this week per Originating AS

Prefixes	ASnum	AS Description
-140	AS4	ISI-AS - University of Southern California
-77	AS49687	REQ SC RoSite Equipment SRL
-76	AS7643	VNPT-AS-VN Vietnam Posts and Telecommunications (VNPT)
-74	AS9198	KAZTELECOM-AS JSC Kazakhtelecom
-64	AS8402	CORBINA-AS OJSC "Vimpelcom"
-56	AS2	UDEL-DCN - University of Delaware
-55	AS16852	LVLT-16852 - Level 3 Communications, Inc.
-53	AS36998	SDN-MOBITEL
-47	AS37986	TULIP Tulip Telecom Ltd.
-38	AS51645	IRKUTSK-AS CJSC "ER-Telecom Holding"
-36	AS3593	FRONTIER-EPIX - Frontier Communications of America, Inc.
-35	AS28509	Cablemas Telecomunicaciones SA de CV

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
4279	4426	AS7029	WINDSTREAM - Windstream Communications Inc
3722	4736	AS4	ISI-AS - University of Southern California
3707	6157	AS3	MIT-GATEWAYS - Massachusetts Institute of Technology
3434	3441	AS28573	NET Serviços de Comunicação S.A.
2992	3029	AS6389	BELLSOUTH-NET-BLK - BellSouth.net Inc.
2862	2944	AS4766	KIXS-AS-KR Korea Telecom
2723	2736	AS17974	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
2722	2932	AS4323	TWTC - tw telecom holdings, inc.
2695	2696	AS10620	Telmex Colombia S.A.
2303	2872	AS2	UDEL-DCN - University of Delaware
2256	2326	AS22773	ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
2069	2148	AS1785	AS-PAETEC-NET - PaeTec Communications, Inc.
2054	2132	AS7545	TPG-INTERNET-AP TPG Telecom Limited
2029	2048	AS18566	MEGAPATH5-US - MegaPath Corporation
1811	1825	AS8402	CORBINA-AS OJSC "Vimpelcom"
1795	1810	AS4755	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
1771	1795	AS18881	Global Village Telecom
1736	1743	AS7303	Telecom Argentina S.A.
1626	1681	AS20115	CHARTER-NET-HKY-NC - Charter Communications
1574	1574	AS9829	BSNL-NIB National Internet Backbone

Report: ASes ordered by number of more specific prefixes
Report: More Specific prefix list (by AS)
Report: More Specific prefix list (ordered by prefix)

Announced Prefixes

Rank	AS	Type	Originate Addr Space	(pfx)	Transit Addr space	(pfx)	Description
14	AS6389	ORG+TRN	Originate: 30062848	/7.16	Transit: 504832	/13.05	BELLSOUTH-NET-BLK - BellSouth.net Inc.

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	%
3	AS6389	BELLSOUTH-NET-BLK - BellSouth.net Inc.	3029	2978	5	56	2973	98.15%

Prefix	AS Path	Aggregation Suggestion
12.81.90.0/23	4777 2516 3356 7018 6389	
12.81.120.0/24	4777 2516 3356 7018 6389	
12.83.3.0/24	4777 2516 3356 7018 6389	
12.83.5.0/24	4777 2516 3356 7018 6389	
12.83.7.0/24	4777 2516 3356 7018 6389	
65.0.0.0/12	4777 2516 3356 7018 6389	
65.0.0.0/18	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.0.0.0/19	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.0.40.0/22	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.0.50.0/23	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.0.64.0/18	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.0.128.0/18	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.0.192.0/19	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.0.224.0/19	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.1.0.0/19	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.1.32.0/19	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.1.64.0/19	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.1.128.0/18	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.1.224.0/20	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.1.240.0/20	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.2.0.0/16	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.2.0.0/17	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.2.128.0/17	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.3.224.0/19	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.4.64.0/18	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.4.192.0/18	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.5.1.0/24	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.5.12.0/22	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.5.16.0/22	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.5.20.0/23	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.5.21.0/24	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.5.22.0/23	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389
65.5.24.0/22	4777 2516 3356 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 4777 2516 3356 7018 6389

AS Report

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

Reader

Apple Smart Bookmarks Popular Philip TinyURL! Networking Miscellaneous Most Visited APNIC Radio Most Visited

AS Report

Announced Prefixes

Rank	AS	Type	Originate Addr Space	(pfx)	Transit Addr space	(pfx)	Description
187	AS18566	ORG+TRN	Originate: 2773760	/10.60	Transit: 389632	/13.43	MEGAPATH5-US - MegaPath Corporation

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	%
12	AS18566	MEGAPATH5-US - MegaPath Corporation	2048	1638	155	565	1483	72.41%

Prefix	AS Path	Aggregation Suggestion
64.81.16.0/22	4777 2516 3356 18566	
64.81.20.0/22	4777 2516 2828 18566	
64.81.22.0/24	4777 2516 4565 18566	
64.81.24.0/21	4777 2516 3356 18566	+ Announce - aggregate of 64.81.24.0/22 (4777 2516 3356 18566) and 64.81.28.0/22 (4777 2516 335
64.81.24.0/22	4777 2516 3356 18566	- Withdrawn - aggregated with 64.81.28.0/22 (4777 2516 3356 18566)
64.81.28.0/22	4777 2516 3356 18566	- Withdrawn - aggregated with 64.81.24.0/22 (4777 2516 3356 18566)
64.81.32.0/20	4777 2516 4565 18566	
64.81.32.0/24	4777 2516 4565 18566	- Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
64.81.33.0/24	4777 2516 4565 18566	- Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
64.81.34.0/24	4777 2516 4565 18566	- Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
64.81.35.0/24	4777 2516 4565 18566	- Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
64.81.36.0/24	4777 2516 4565 18566	- Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
64.81.37.0/24	4777 2516 4565 18566	- Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
64.81.38.0/24	4777 2516 4565 18566	- Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
64.81.39.0/24	4777 2516 4565 18566	- Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
64.81.40.0/24	4777 2516 4565 18566	- Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
64.81.44.0/24	4777 2516 4565 18566	- Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
64.81.48.0/20	4777 2516 3356 18566	
64.81.48.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.49.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.50.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.51.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.52.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.53.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.54.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.55.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.56.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.57.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.58.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.59.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.60.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.61.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 3356 18566
64.81.64.0/20	4777 2516 3356 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 11 January 2014 06:17 (UTC+1000)

50 Most active ASes for the past 7 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	9829	52564	2.86%	1563	33.63	BSNL-NIB National Internet Backbone
2	4766	51980	2.83%	2946	17.64	KIXS-AS-KR Korea Telecom
3	3816	43341	2.36%	767	56.51	COLOMBIA TELECOMUNICACIONES S.A. ESP
4	8402	40335	2.20%	1995	20.22	CORBINA-AS OJSC "Vimpelcom"
5	14287	35253	1.92%	54	652.83	TRIAD-TELECOM - Triad Telecom, Inc.
6	4323	30619	1.67%	2949	10.38	TWTC - tw telecom holdings, inc.
7	701	29548	1.61%	8623	3.43	UUNET - MCI Communications Services, Inc. d/b/a Verizon Business
8	7029	22867	1.25%	4475	5.11	WINDSTREAM - Windstream Communications Inc
9	13118	20964	1.14%	44	476.45	ASN-YARTELECOM OJSC Rostelecom
10	4775	18599	1.01%	130	143.07	GLOBE-TELECOM-AS Globe Telecoms
11	7552	18089	0.99%	1541	11.74	VIETEL-AS-AP Viettel Corporation
12	41691	16864	0.92%	36	468.44	SUMTEL-AS-RIPE Summa Telecom LLC
13	4538	16169	0.88%	541	29.89	ERX-CERNET-BKB China Education and Research Network Center
14	45899	13866	0.76%	402	34.49	VNPT-AS-VN VNPT Corp
15	59217	13537	0.74%	1	13537.00	AZONNELIMITED-AS-AP Azonne Limited
16	17557	11875	0.65%	122	97.34	PKTELECOM-AS-PK Pakistan Telecommunication Company Limited
17	11976	10996	0.60%	204	53.90	FIDN - Fidelity Communication International Inc.
18	6629	10990	0.60%	68	161.62	NOAA-AS - NOAA
19	1221	10766	0.59%	671	16.04	ASN-TELSTRA Telstra Pty Ltd
20	64777	9985	0.54%	6928	1.44	-Private Use AS-
21	11054	9624	0.52%	23	418.43	LIVEPERSON LivePerson, Inc
22	9583	9496	0.52%	1339	7.09	SIFY-AS-IN Sify Limited
23	24560	9023	0.49%	1098	8.22	AIRTELBROADBAND-AS-AP Bharti Airtel Ltd., Telemedia Services
24	5668	8969	0.49%	159	56.41	AS-5668 - CenturyTel Internet Holdings, Inc

The BGP Instability Report

bgpupdates.potaroo.net/instability/bgpupd.html

Apple Smart Bookmarks Popular Philip TinyURL! Networking Miscellaneous Most Visited APNIC Radio Most Visited

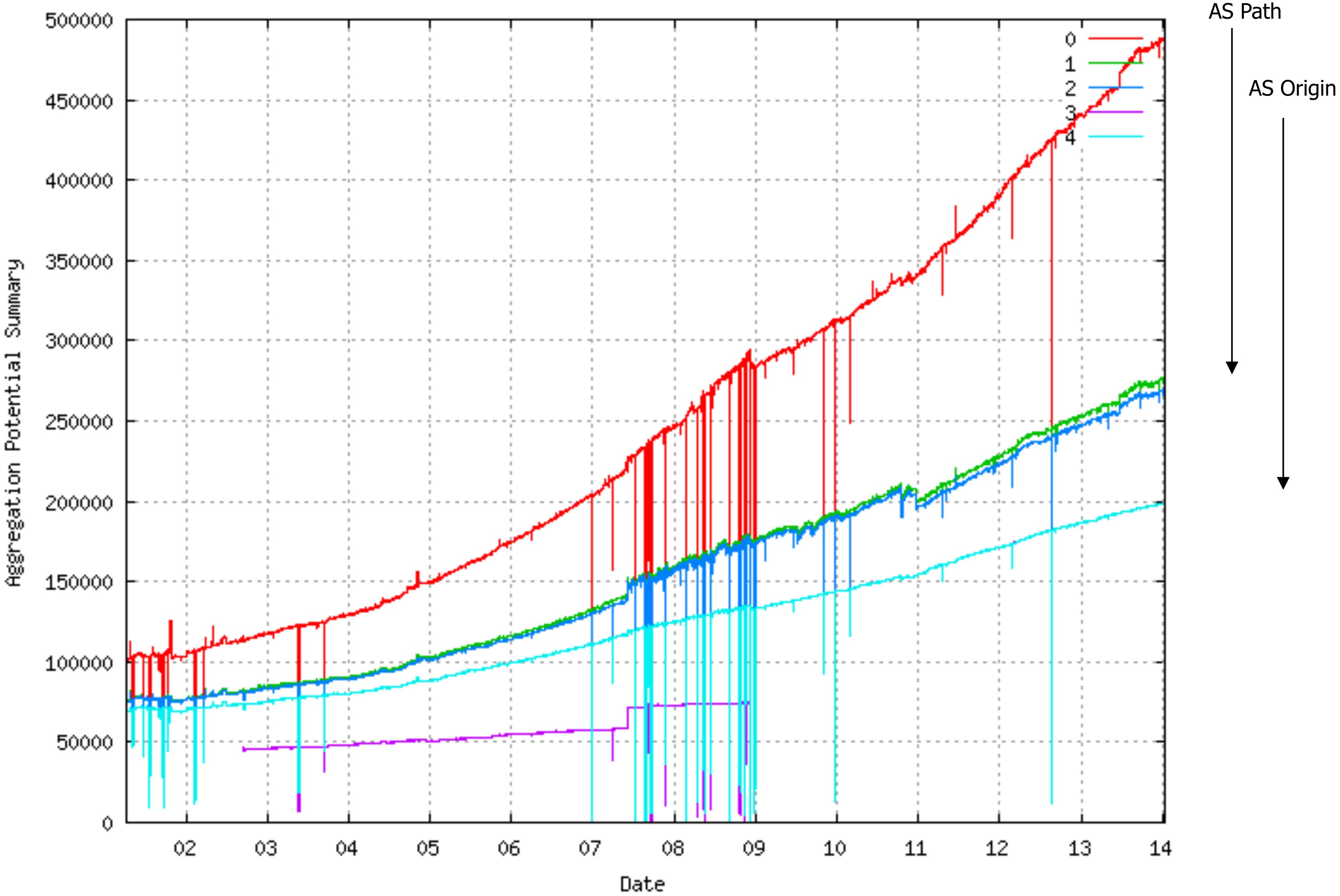
The BGP Instability Report

50 Most active Prefixes for the past 7 days

RANK	PREFIX	UPDs	%	Origin AS -- AS NAME
1	109.161.64.0/20	20959	1.05%	13118 -- ASN-YARTELECOM OJSC Rostelecom
2	103.243.164.0/22	13537	0.68%	59217 -- AZONNELIMITED-AS-AP Azonne Limited
3	89.221.206.0/24	10295	0.52%	41691 -- SUMTEL-AS-RIPE Summa Telecom LLC
4	192.58.232.0/24	10163	0.51%	6629 -- NOAA-AS - NOAA
5	120.28.62.0/24	9338	0.47%	4775 -- GLOBE-TELECOM-AS Globe Telecoms
6	222.127.0.0/24	9085	0.46%	4775 -- GLOBE-TELECOM-AS Globe Telecoms
7	208.70.20.0/22	7061	0.35%	14287 -- TRIAD-TELECOM - Triad Telecom, Inc.
8	208.73.244.0/22	7049	0.35%	14287 -- TRIAD-TELECOM - Triad Telecom, Inc.
9	208.88.232.0/22	7043	0.35%	14287 -- TRIAD-TELECOM - Triad Telecom, Inc.
10	208.78.116.0/22	7041	0.35%	14287 -- TRIAD-TELECOM - Triad Telecom, Inc.
11	216.162.0.0/20	7035	0.35%	14287 -- TRIAD-TELECOM - Triad Telecom, Inc.
12	206.152.15.0/24	6971	0.35%	54465 -- QPM-AS-1 - QuickPlay Media Inc.
13	148.218.0.0/16	6823	0.34%	11172 -- Alestra, S. de R.L. de C.V. 28477 -- UNIVERSIDAD AUTONOMA DEL ESTADO DE MORELOS
14	85.249.160.0/20	6549	0.33%	41691 -- SUMTEL-AS-RIPE Summa Telecom LLC
15	67.210.190.0/23	6226	0.31%	11976 -- FIDN - Fidelity Communication International Inc.
16	194.105.61.0/24	5827	0.29%	12922 -- MULTITRADE-AS Bank Outsourcer
17	199.187.118.0/24	4838	0.24%	11054 -- LIVEPERSON LivePerson, Inc
18	67.210.188.0/23	4765	0.24%	11976 -- FIDN - Fidelity Communication International Inc.
19	103.2.238.0/24	4584	0.23%	45194 -- SIPL-AS Syscon Infoway Pvt. Ltd., Internet Service Provider, India. 56097 -- JOISTER-AS Joister Group of Companies.
20	183.87.110.0/24	4584	0.23%	45194 -- SIPL-AS Syscon Infoway Pvt. Ltd., Internet Service Provider, India. 56097 -- JOISTER-AS Joister Group of Companies.
21	165.156.25.0/24	4501	0.23%	30437 -- GE-MS003 - General Electric Company
22	168.223.206.0/23	4213	0.21%	7202 -- FAMU - Florida A & M University
23	168.223.200.0/22	4185	0.21%	7202 -- FAMU - Florida A & M University
24	69.38.178.0/24	3923	0.20%	19406 -- TWRS-MA - Towerstream I, Inc.
26	170.128.223.0/24	3674	0.18%	11685 -- HNBCOL-AS - Huntington National Bank
27	205.166.53.0/24	3038	0.15%	22747 -- TCIS - TulsaConnect, LLC

Go to "http://bgpupdates.potaroo.net/cgi-bin/per-as?as=17557"

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 jwhois.apnic.net 202.12.29.0
inetnum:        202.12.28.0 - 202.12.29.255
netname:        APNIC-AP
descr:          Asia Pacific Network Information Centre
descr:          Regional Internet Registry for the Asia-Pacific
descr:          6 Cordelia Street
descr:          South Brisbane, QLD 4101
descr:          Australia
country:        AU
admin-c:        AIC1-AP
tech-c:         NO4-AP
mnt-by:         APNIC-HM
mnt-irt:        IRT-APNIC-AP
changed:        hm-changed@apnic.net
status:         ASSIGNED PORTABLE
changed:        hm-changed@apnic.net 20110309
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.rfc-editor.org/rfc/rfc6441.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

Preparing the network

Before we begin…

Preparing the Network

- We will deploy BGP across the network before we try and multihome
- BGP will be used therefore an ASN is required
- If multihoming to different ISPs, public ASN needed:
  - Either go to upstream ISP who is a registry member, or
  - Apply to the RIR yourself for a one off assignment, or
  - Ask an ISP who is a registry member, or
  - Join the RIR and get your own IP address allocation too
    - (this option strongly recommended)!

Preparing the Network
Initial Assumptions

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

Preparing the Network
First Step: IGP

- Decide on an IGP: OSPF or ISIS ☺
- Assign loopback interfaces and /32 address to each router which will run the IGP
  - Loopback is used for OSPF and BGP router id anchor
  - Used for iBGP and route origination
- Deploy IGP (e.g. OSPF)
  - IGP can be deployed with NO IMPACT on the existing static routing
  - e.g. OSPF distance might be 110, static distance is 1
  - Smallest distance wins

Preparing the Network IGP (cont)

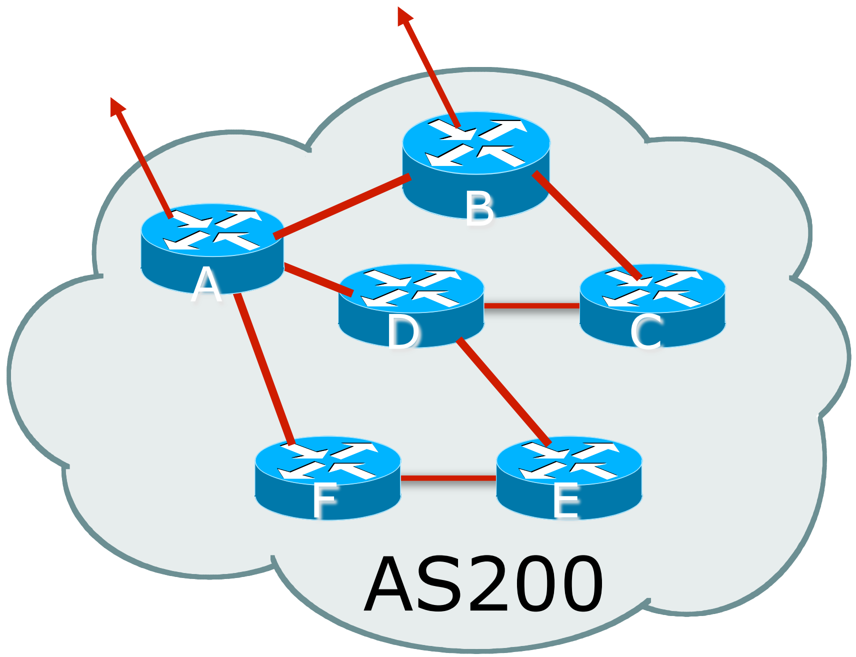
- Be prudent deploying IGP – keep the Link State Database Lean!
  - Router loopbacks go in IGP
  - WAN point to point links go in IGP
  - (In fact, any link where IGP dynamic routing will be run should go into IGP)
  - Summarise on area/level boundaries (if possible) – i.e. think about your IGP address plan

Preparing the Network IGP (cont)

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

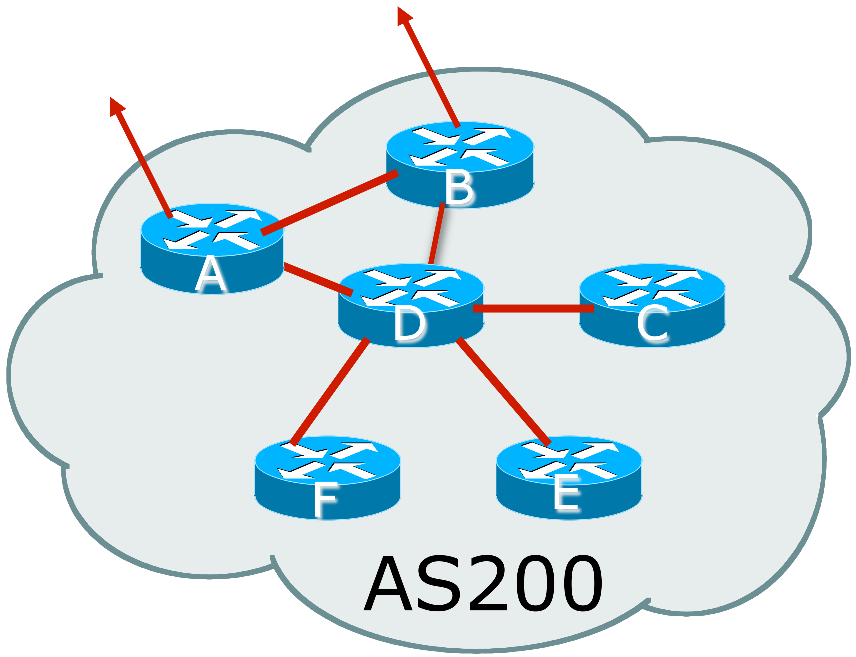
Preparing the Network
Second Step: iBGP

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



Preparing the Network
Second Step: iBGP (Transit Path)

- iBGP must run on all routers which are in the transit path between external connections
- Routers C, E and F are not in the transit path
  - Static routes or IGP will suffice
- Router D is in the transit path
  - Will need to be in iBGP mesh, otherwise routing loops will result



Preparing the Network Layers

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

Preparing the Network Aggregation Layer

- iBGP is optional
  - Many ISPs run iBGP here, either partial routing (more common) or full routing (less common)
  - Full routing is not needed unless customers want full table
  - Partial routing is cheaper/easier, might usually consist of internal prefixes and, optionally, external prefixes to aid external load balancing
    - Communities and peer-groups make this administratively easy
- Many aggregation devices can't run iBGP
  - Static routes from distribution devices for address pools
  - IGP for best exit

Preparing the Network Distribution Layer

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

Preparing the Network
Core Layer

- Core of network is usually the transit path
- iBGP necessary between core devices
  - Full routes or partial routes:
    - Transit ISPs carry full routes in core
    - Edge ISPs carry partial routes only
- Core layer includes AS border routers

Preparing the Network
iBGP Implementation

Decide on:

- Best iBGP policy
  - Will it be full routes everywhere, or partial, or some mix?
- iBGP scaling technique
  - Community policy?
  - Route-reflectors?
  - Techniques such as peer groups and peer templates?

Preparing the Network
iBGP Implementation

- Then deploy iBGP:
  - Step 1: Introduce iBGP mesh on chosen routers
    - make sure that iBGP distance is greater than IGP distance (it usually is)
  - Step 2: Install "customer" prefixes into iBGP

    Check! Does the network still work?
  - Step 3: Carefully remove the static routing for the prefixes now in IGP and iBGP

    Check! Does the network still work?
  - Step 4: Deployment of eBGP follows

Preparing the Network
iBGP Implementation

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

Preparing the Network
iBGP Implementation

- *Carefully remove static routes?*
- Work on one router at a time:
  - Check that static route for a particular destination is also learned by the iBGP
  - If so, remove it
  - If not, establish why and fix the problem
  - (Remember to look in the RIB, not the FIB!)
- Then the next router, until the whole PoP is done
- Then the next PoP, and so on until the network is now dependent on the IGP and iBGP you have deployed

Preparing the Network Completion

- Previous steps are NOT flag day steps
  - Each can be carried out during different maintenance periods, for example:
  - Step One on Week One
  - Step Two on Week Two
  - Step Three on Week Three
  - And so on
  - And with proper planning will have NO customer visible impact at all

Preparing the Network
Configuration Summary

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

Configuration Tips

Of passwords, tricks and templates

iBGP and IGPs Reminder!

- Make sure loopback is configured on router
  - iBGP between loopbacks, NOT real interfaces
- Make sure IGP carries loopback IPv4 /32 and IPv6 /128 address
- Consider the DMZ nets:
  - Use unnumbered interfaces?
  - Use next-hop-self on iBGP neighbours
  - Or carry the DMZ IPv4 /30s and IPv6 /127s 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 point-to-point 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

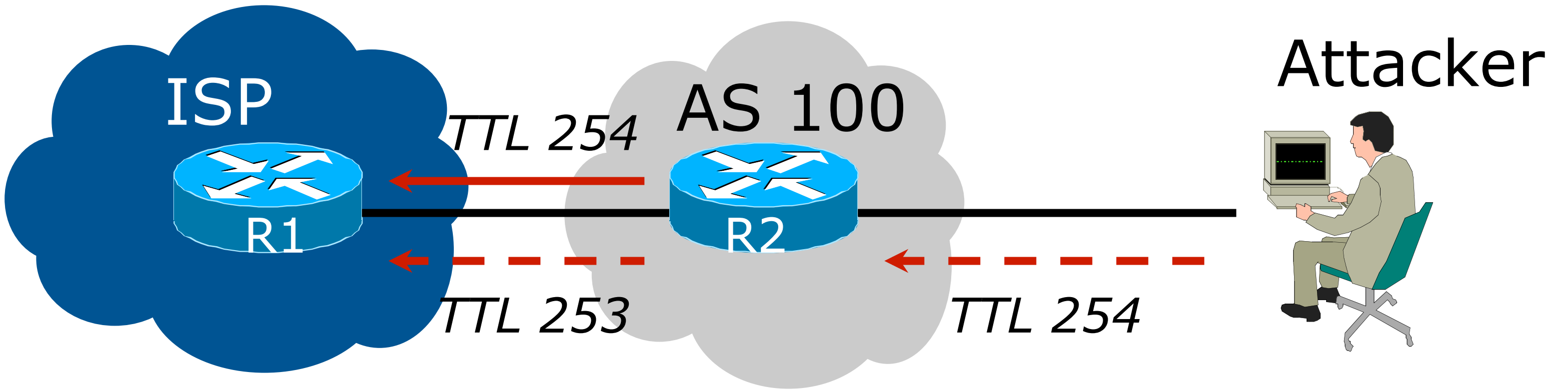
```
*>i193.105.15.0       2516 3257 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 i
```

  This example shows 100 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

SANOG 23

15 January 2014

Thimphu