BGP Techniques for Internet Service Providers

Philip Smith <pfs@cisco.com>

NZNOG'08

23rd-25th January 2008

Dunedin

Presentation Slides

- Will be available on

  ftp://ftp-eng.cisco.com

  /pfs/seminars/NZNOG08-BGP-Techniques.pdf

  And on the NZNOG'08 website

- Feel free to ask questions any time

BGP Techniques for Internet Service Providers

- BGP Basics
- Scaling BGP
- 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

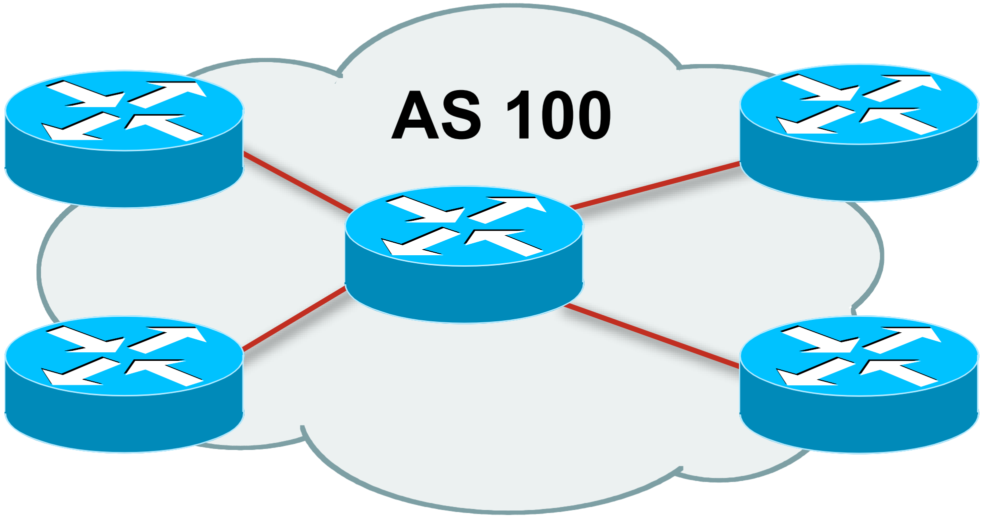
    RFC4276 gives an implementation report on BGP

    RFC4277 describes operational experiences using BGP

- The Autonomous System is BGP's fundamental operating unit

    It is used to uniquely identify networks with a common routing policy

Autonomous System (AS)



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

Autonomous System Number (ASN)

- An ASN is a 16 bit integer

  1-64511 are for use on the public Internet

  64512-65534 are for private use only

  0 and 65535 are reserved

- ASNs are now extended to 32 bit!

  RFC4893 is standards document describing 32-bit ASNs

  Representation still under discussion:

    32-bit notation or “16.16” notation

    Latter documented in Internet Draft:

      draft-michaelson-4byte-as-representation-05.txt

  AS 23456 is used to represent 32-bit ASNs in 16-bit ASN world

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 45055 have been made to the RIRs

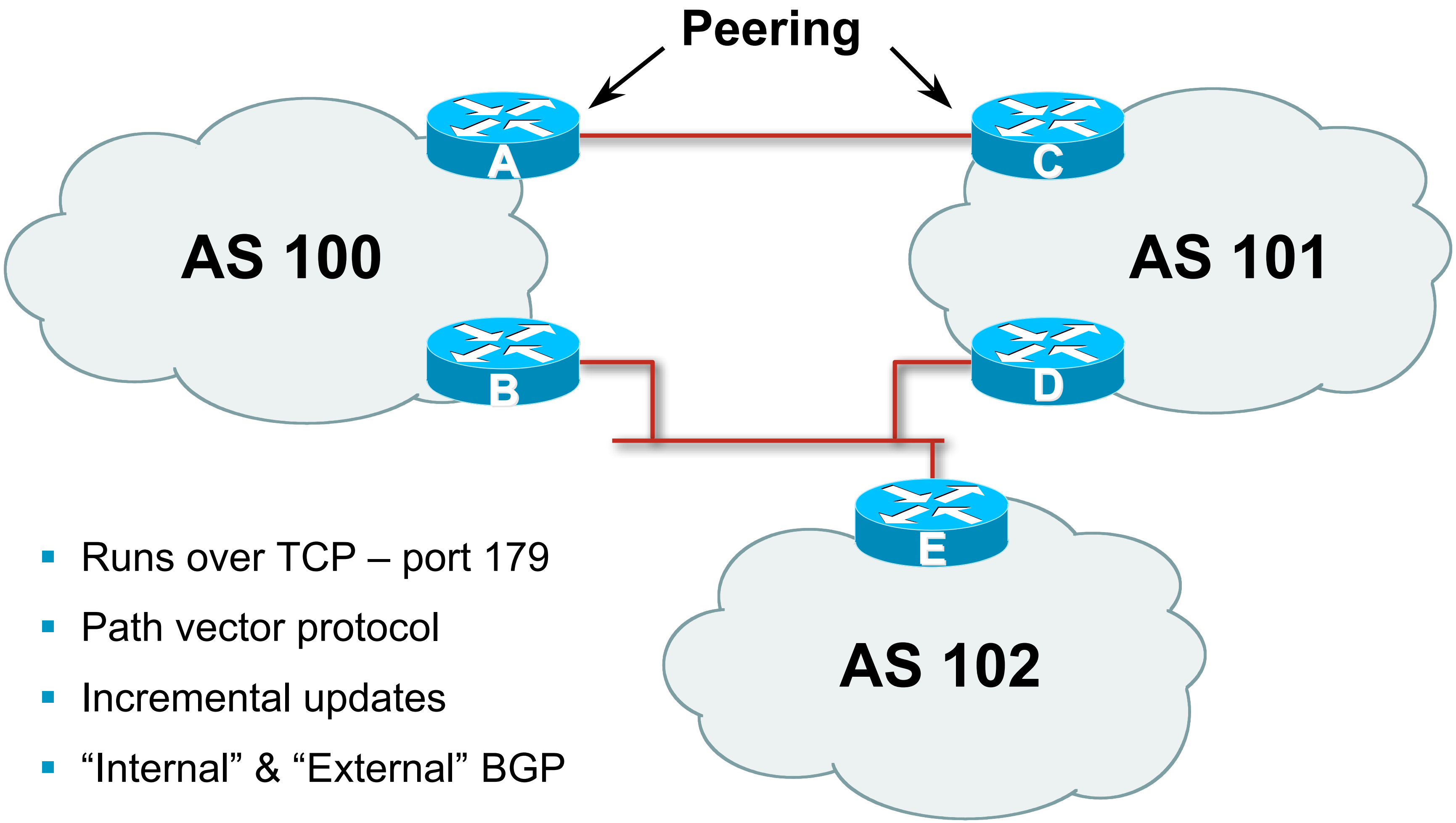
  Around 27200 are visible on the Internet

- The RIRs also have received 1024 32-bit ASNs each

  10 are visible on the Internet (early adopters)

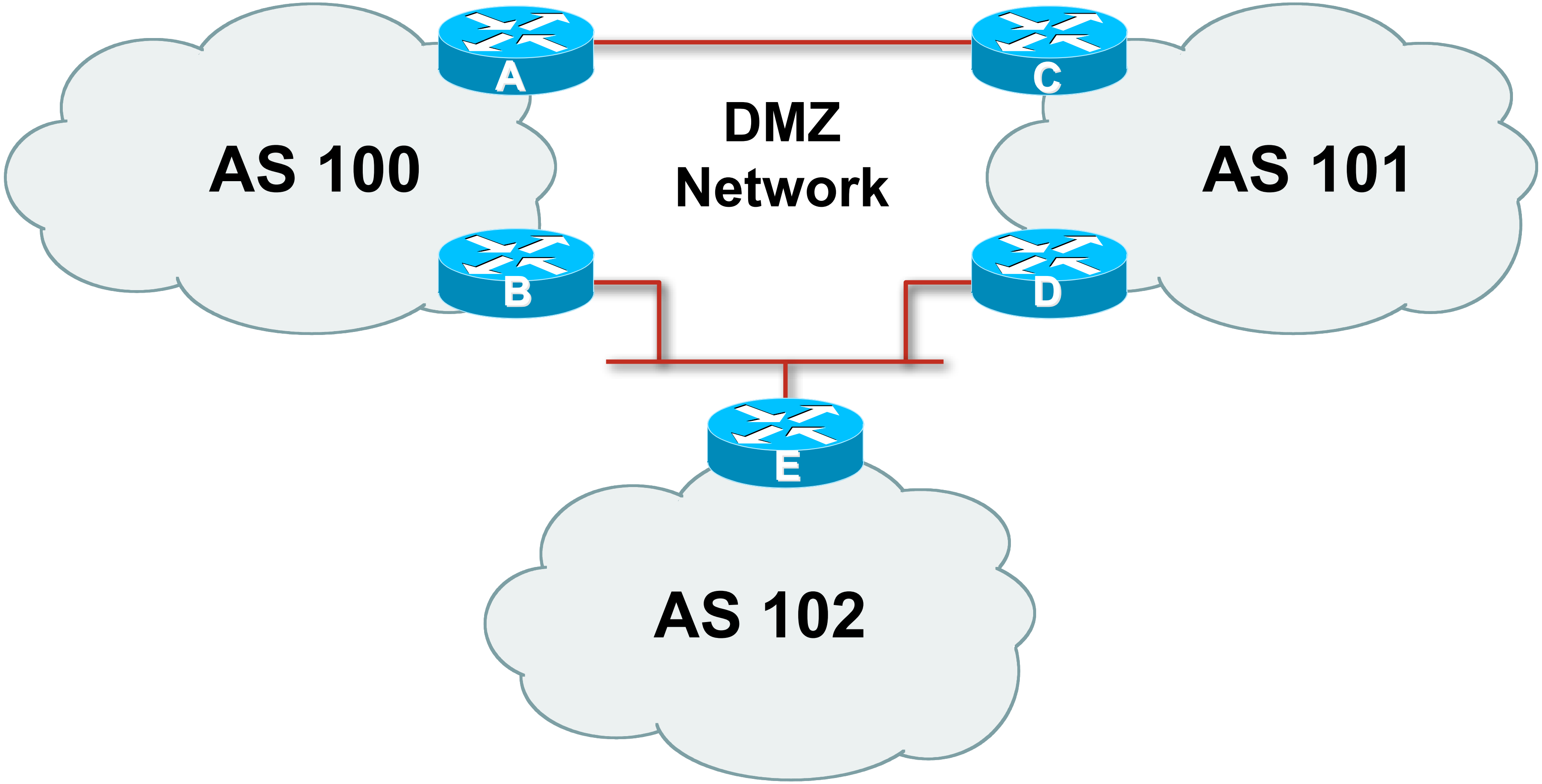
- See **www.iana.org/assignments/as-numbers**

BGP Basics



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

Demarcation Zone (DMZ)



- Shared network between ASes

BGP General Operation

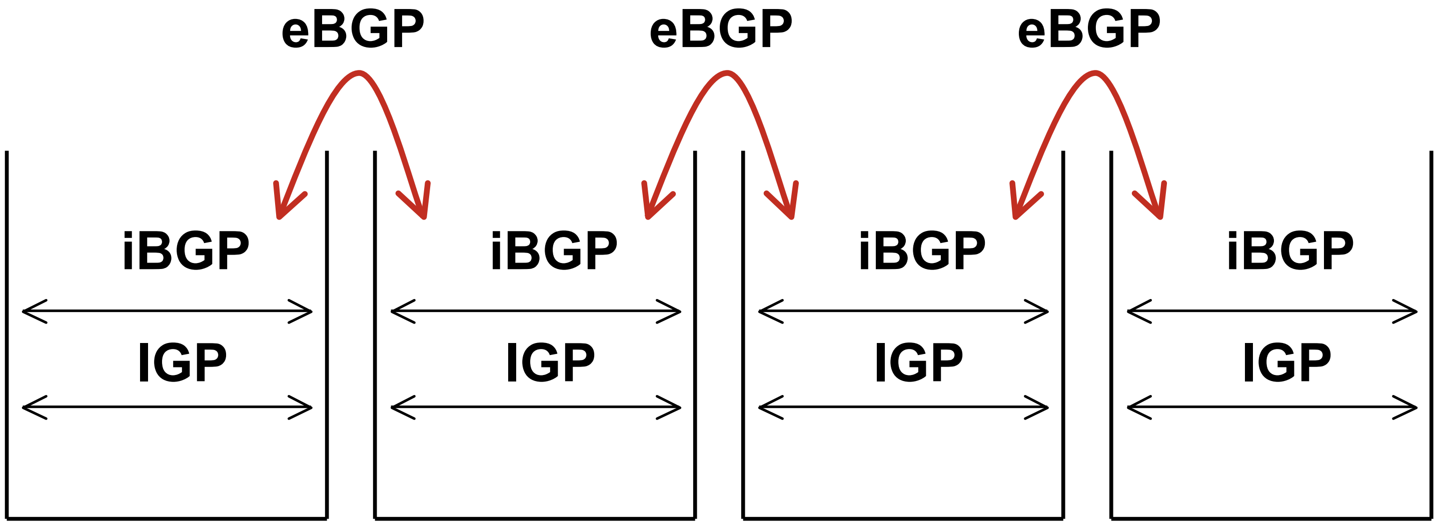
- Learns multiple paths via internal and external BGP speakers
- Picks the best path and installs in the forwarding table
- Best path is sent to external BGP neighbours
- Policies are applied by influencing the best path selection

eBGP & iBGP

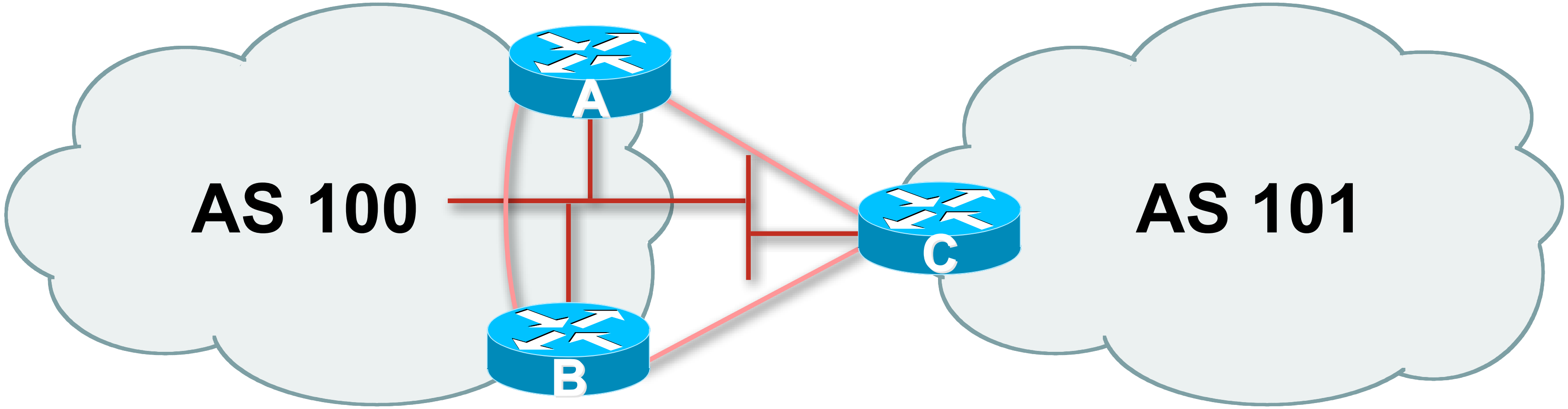
- BGP used internally (iBGP) and externally (eBGP)
- iBGP used to carry
  - some/all Internet prefixes across ISP backbone
  - ISP’s customer prefixes
- eBGP used to
  - exchange prefixes with other ASes
  - implement routing policy

BGP/IGP model used in ISP networks

- Model representation



External BGP Peering (eBGP)

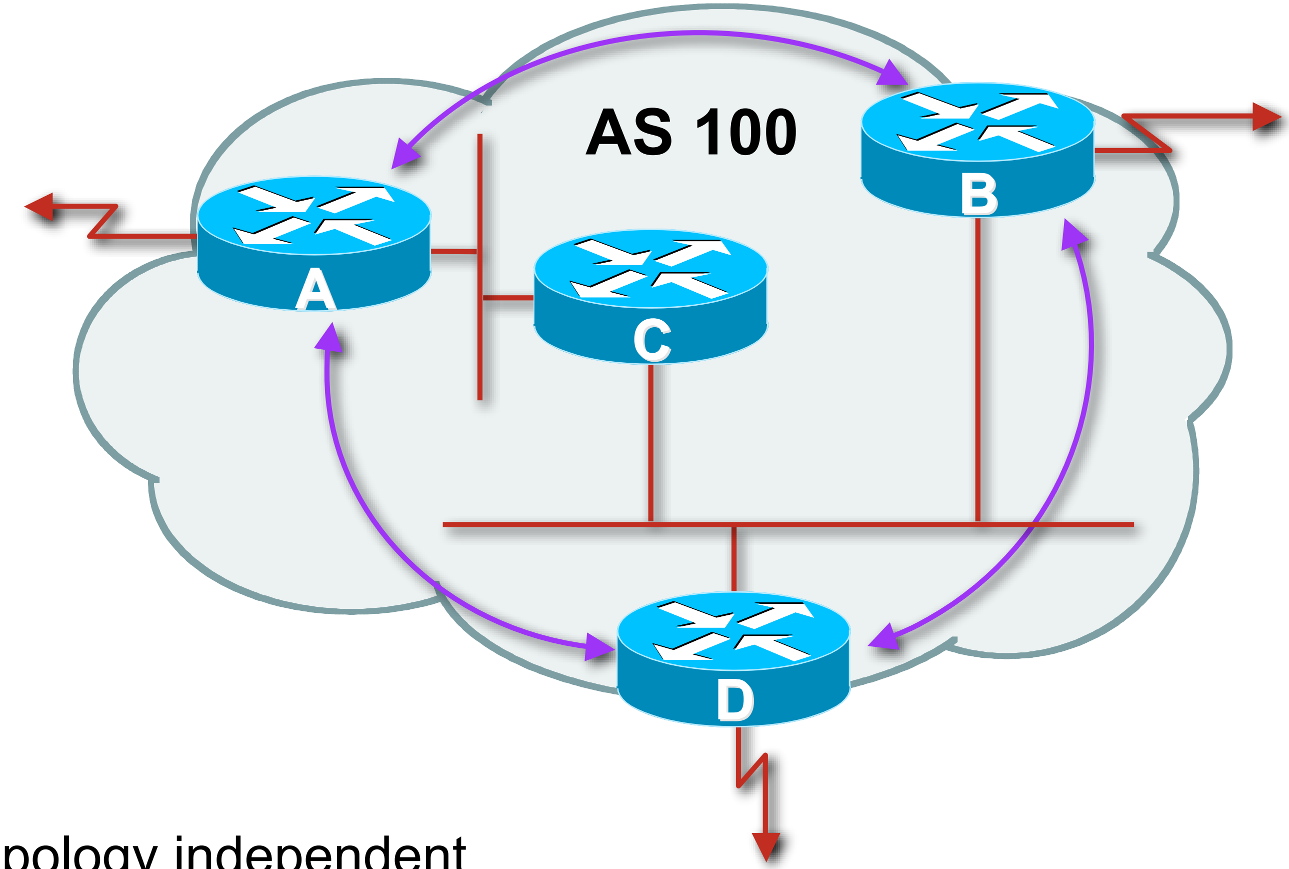


- Between BGP speakers in different AS
- Should be directly connected
- Never run an IGP between eBGP peers

Internal BGP (iBGP)

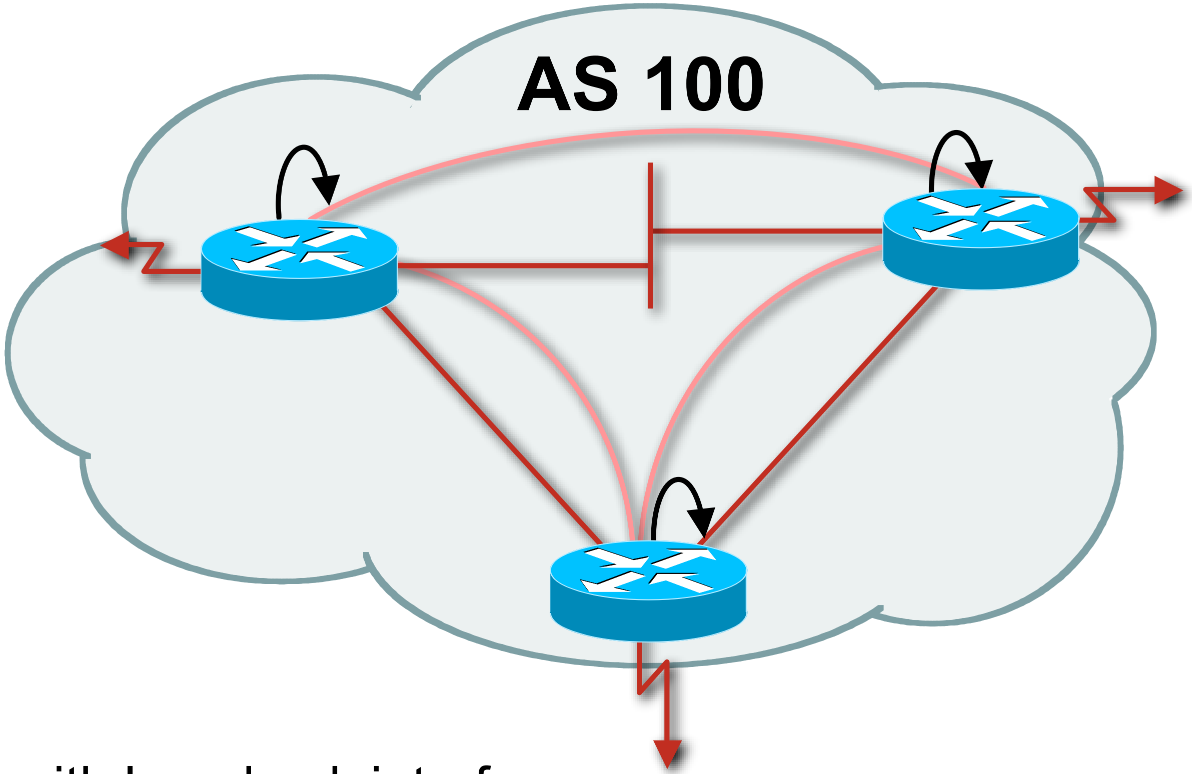
- BGP peer within the same AS
- Not required to be directly connected
  - IGP takes care of inter-BGP speaker connectivity
- iBGP speakers must to be fully meshed:
  - They originate connected networks
  - They pass on prefixes learned from outside the ASN
  - They do not pass on prefixes learned from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

Internal BGP Peering (iBGP)



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

Peering to Loopback Interfaces



- Peer with loop-back interface

  Loop-back interface does not go down – ever!

- Do not want iBGP session to depend on state of a single interface or the physical topology

BGP Attributes

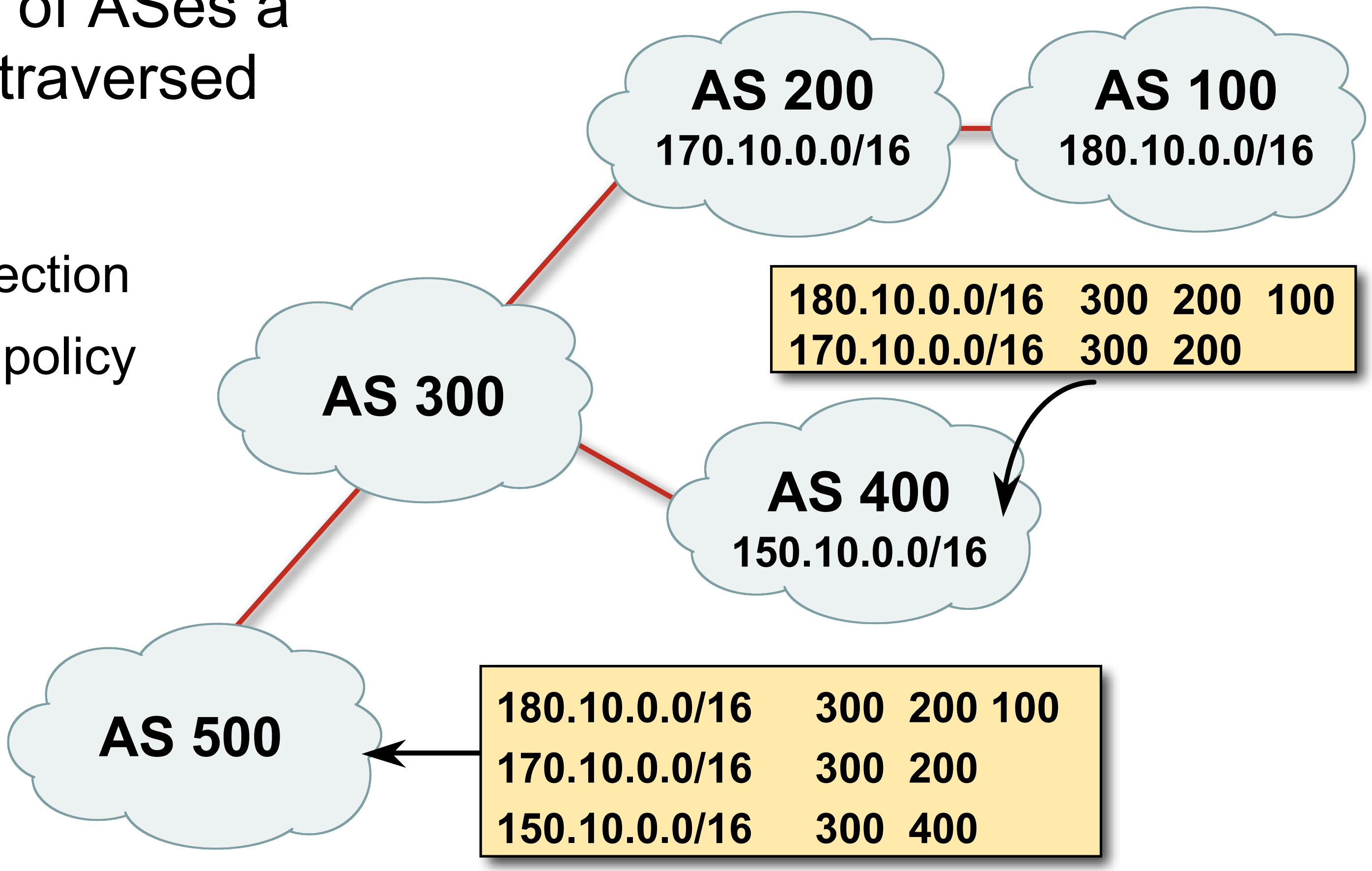
Information about BGP

AS-Path

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

  Loop detection

  Applying policy

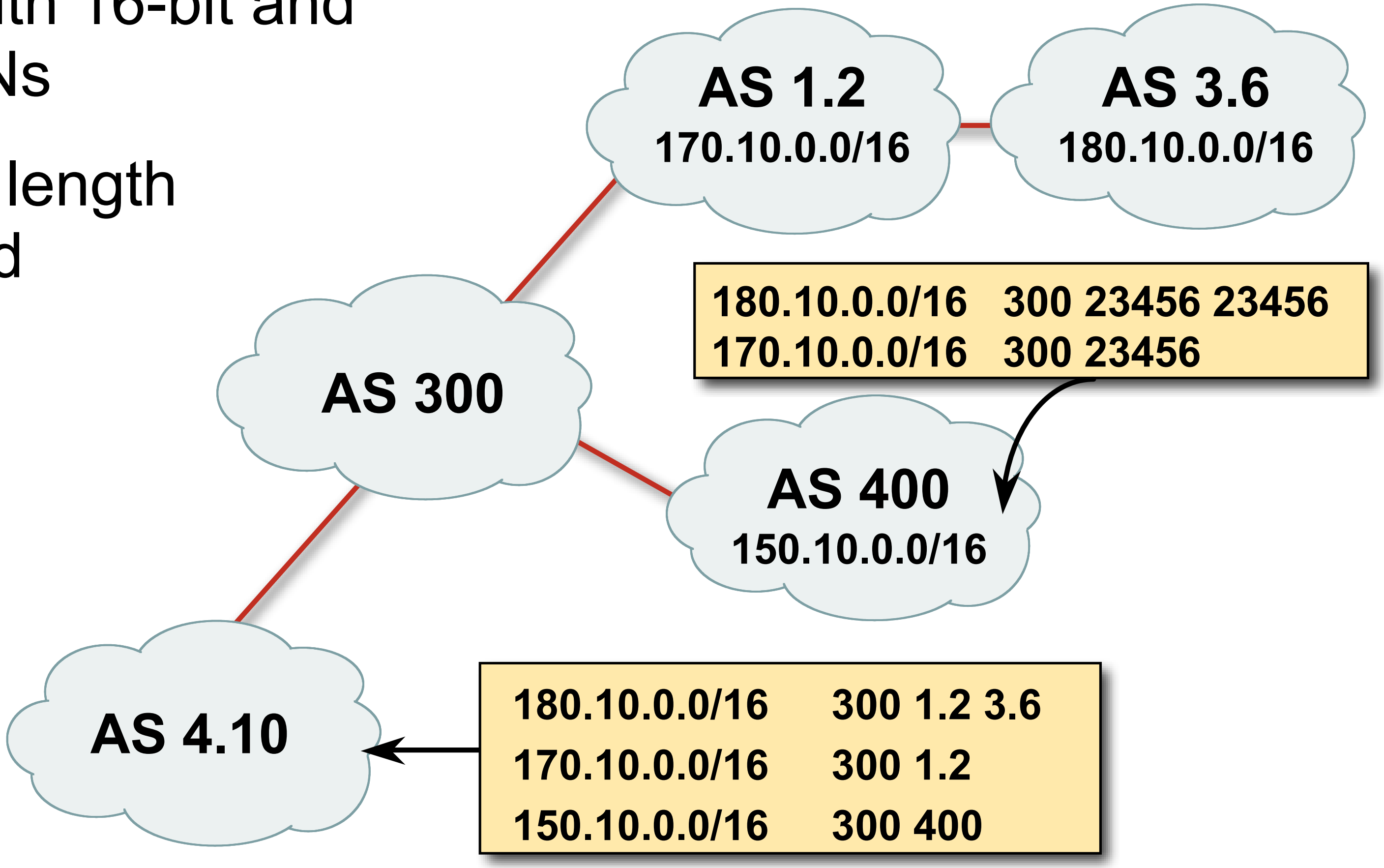


180.10.0.0/16	300 200 100
170.10.0.0/16	300 200

180.10.0.0/16	300 200 100
170.10.0.0/16	300 200
150.10.0.0/16	300 400

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

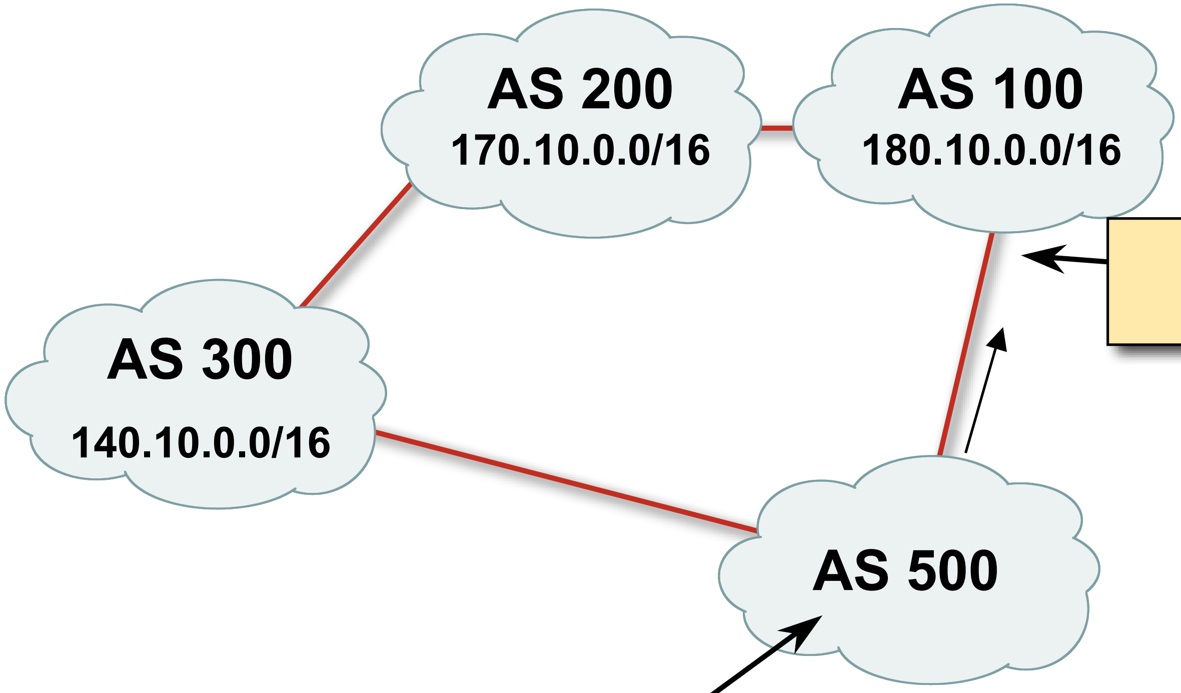
- Internet with 16-bit and 32-bit ASNs
- AS-PATH length maintained



180.10.0.0/16	300 23456 23456
170.10.0.0/16	300 23456

180.10.0.0/16	300 1.2 3.6
170.10.0.0/16	300 1.2
150.10.0.0/16	300 400

AS-Path loop detection

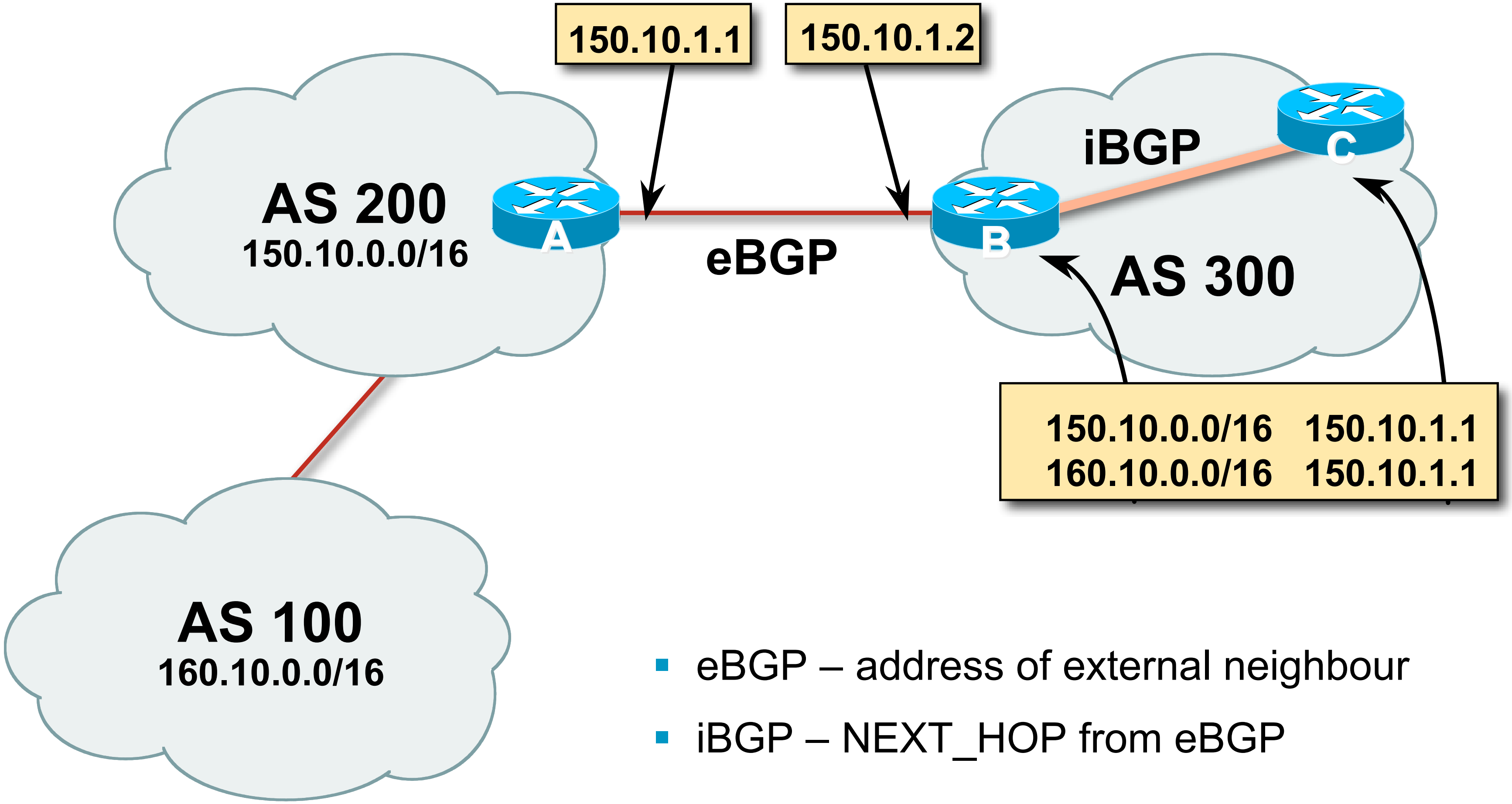


140.10.0.0/16	500 300
170.10.0.0/16	500 300 200

180.10.0.0/16	300 200 100
170.10.0.0/16	300 200
140.10.0.0/16	300

- 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

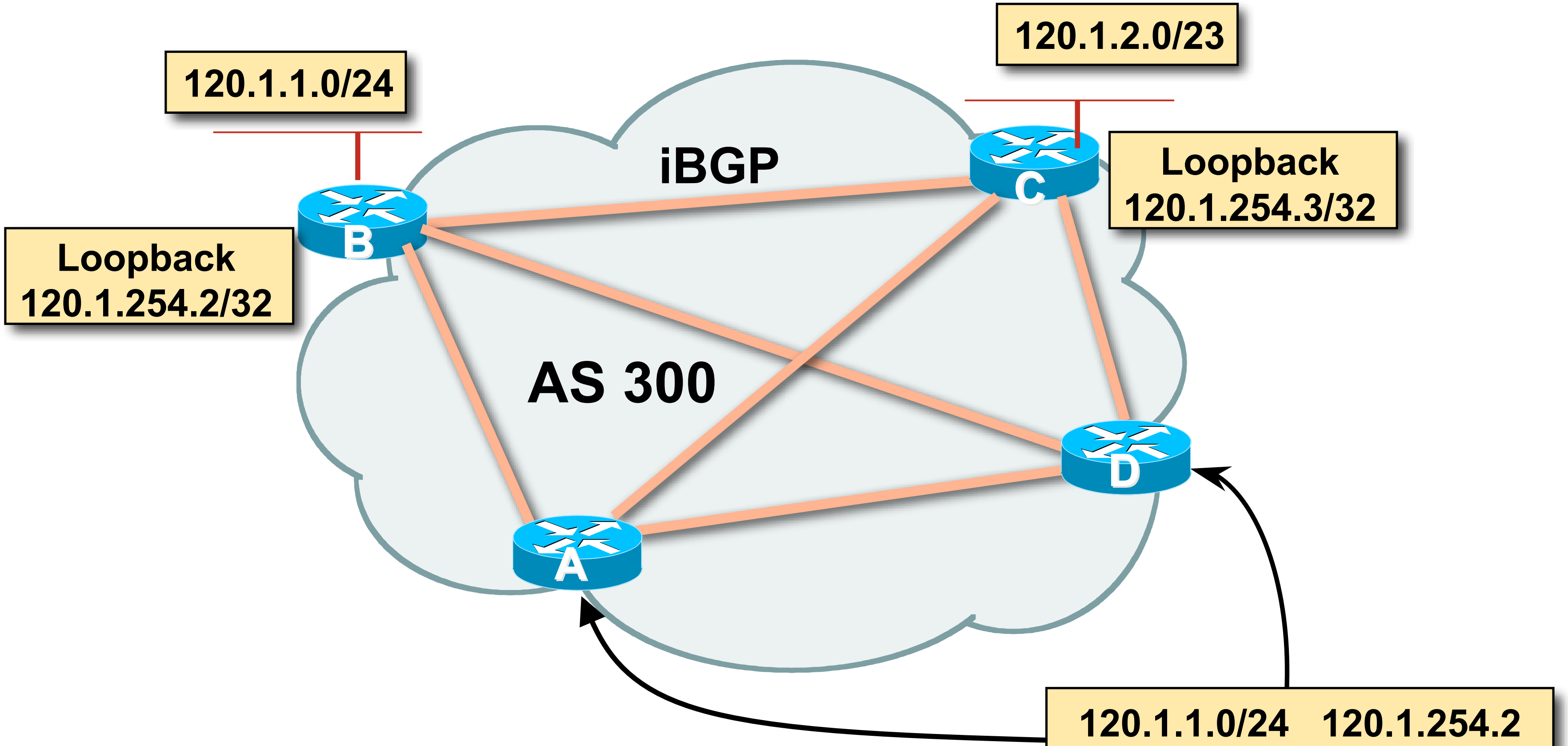
Next Hop



150.10.0.0/16	150.10.1.1
160.10.0.0/16	150.10.1.1

- eBGP – address of external neighbour
- iBGP – NEXT_HOP from eBGP

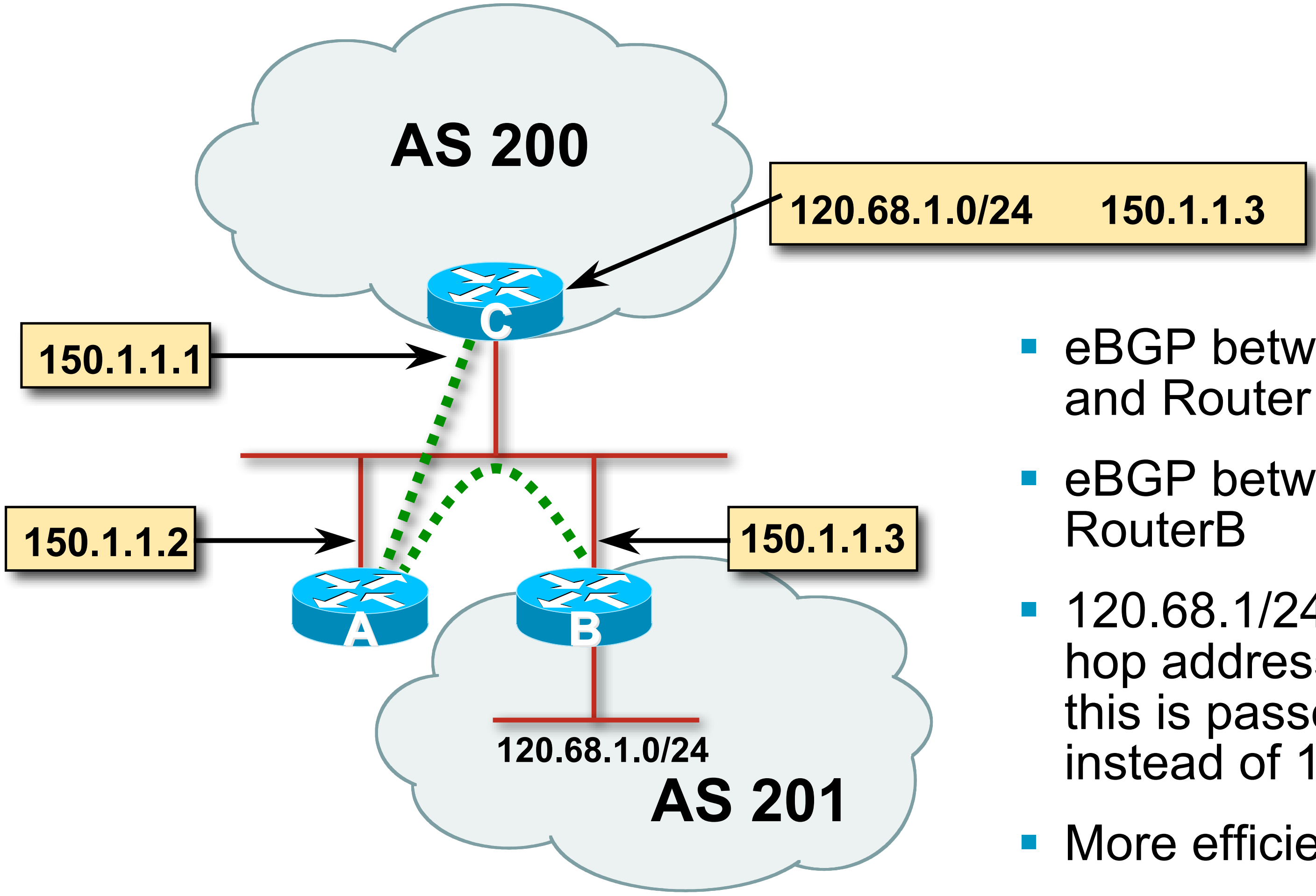
iBGP Next Hop



120.1.1.0/24	120.1.254.2
120.1.2.0/23	120.1.254.3

- 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
- Influences best path selection
- Three values: IGP, EGP, incomplete

  IGP – generated by BGP network statement

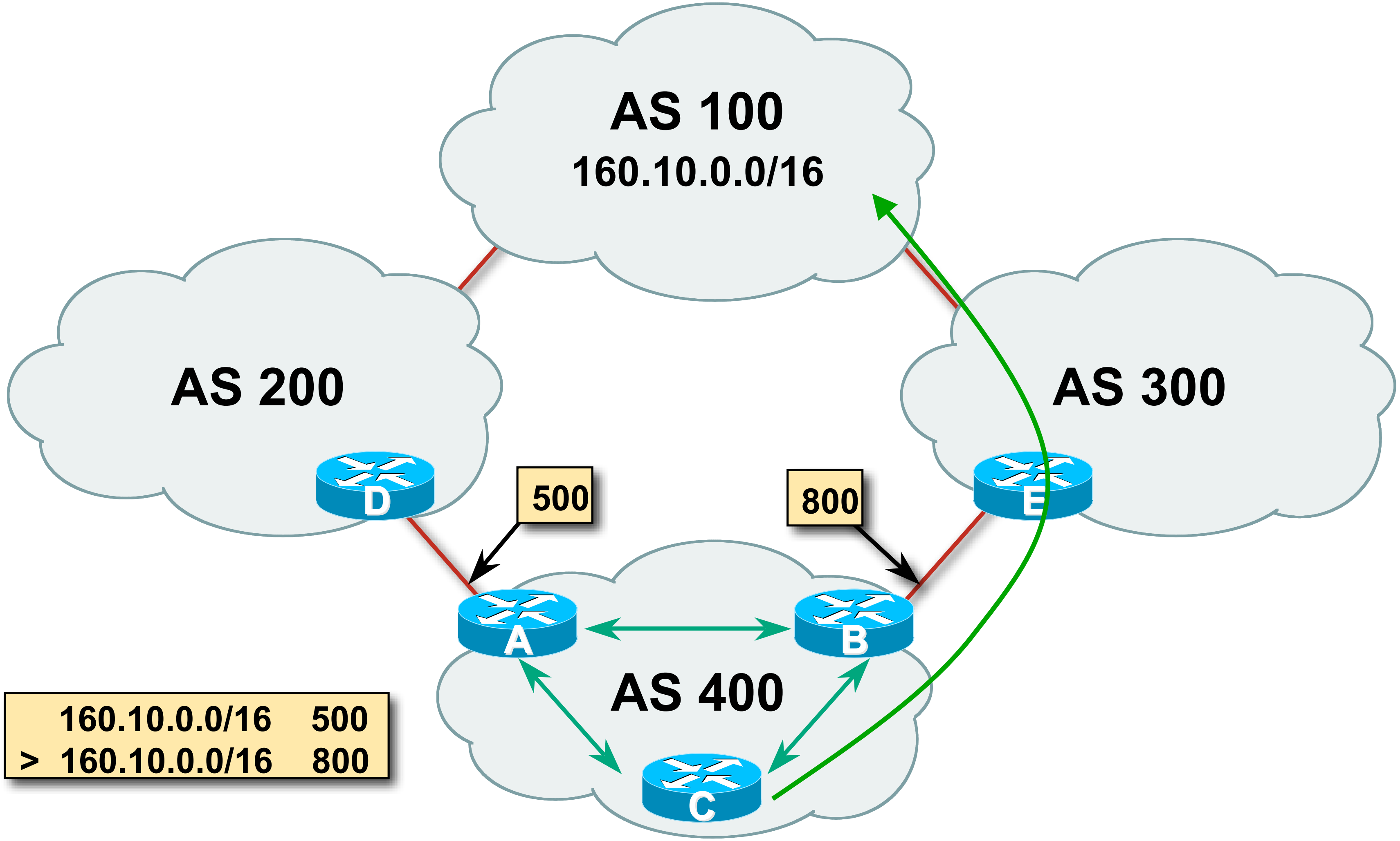
  EGP – generated by EGP

  incomplete – redistributed from another routing protocol

Aggregator

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

Local Preference



```
   160.10.0.0/16   500
>  160.10.0.0/16   800
```

Local Preference

- Local to an AS – non-transitive

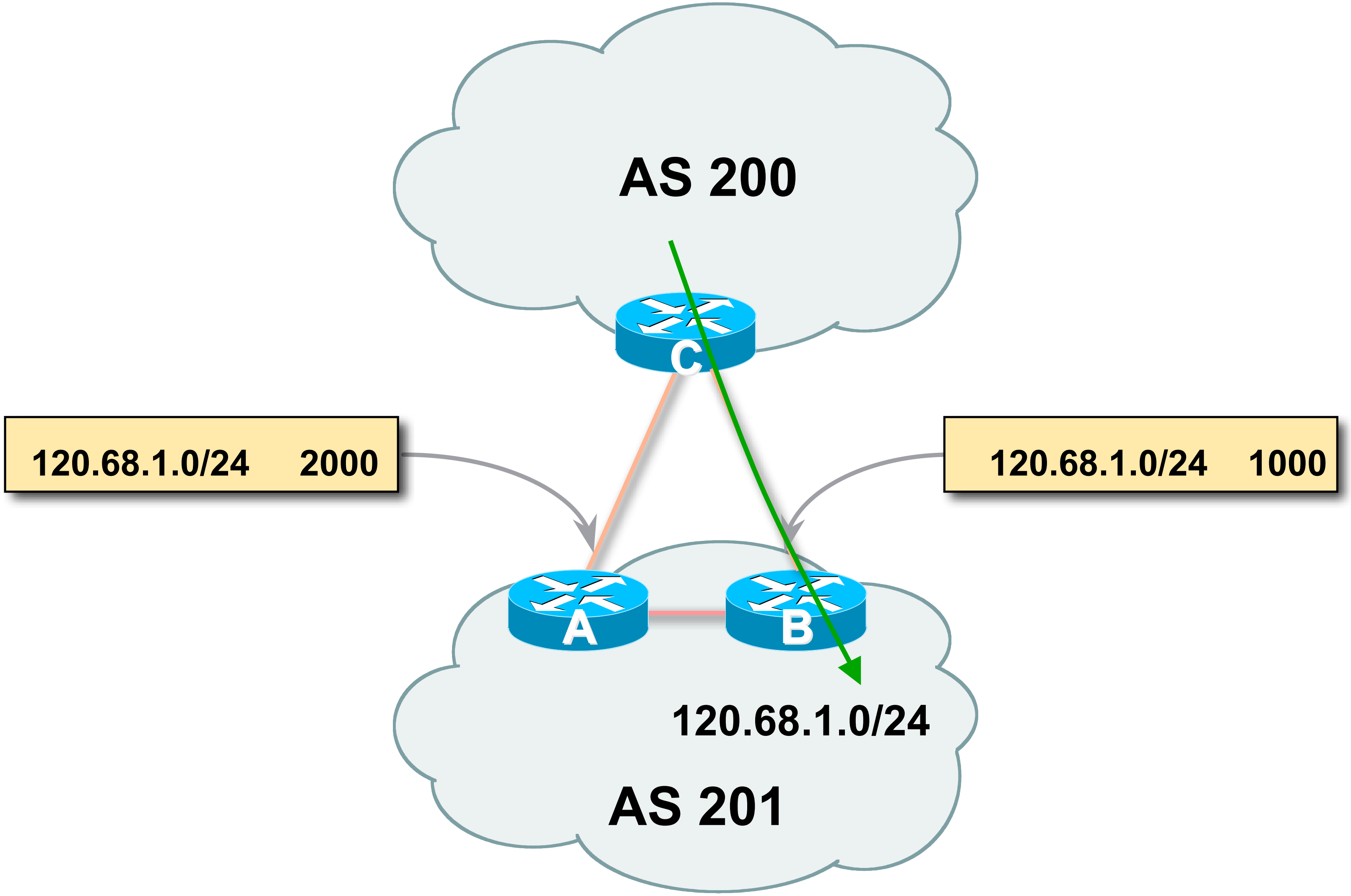
  Default local preference is 100 (IOS)

- Used to influence BGP path selection

  determines best path for *outbound* traffic

- Path with highest local preference wins

Multi-Exit Discriminator (MED)



Multi-Exit Discriminator

- Inter-AS – non-transitive & optional attribute
- Used to convey the relative preference of entry points

  determines best path for inbound traffic

- Comparable if paths are from same AS

  Implementations have a knob to allow comparisons of MEDs from different ASes

- Path with lowest MED wins
- Absence of MED attribute implies MED value of zero (RFC4271)

Multi-Exit Discriminator
"metric confusion"

- MED is non-transitive and optional attribute

  Some implementations send learned MEDs to iBGP peers by default, others do not

  Some implementations send MEDs to eBGP peers by default, others do not

- Default metric varies according to vendor implementation

  Original BGP spec (RFC1771) made no recommendation

  Some implementations said that absence of metric was equivalent to 0

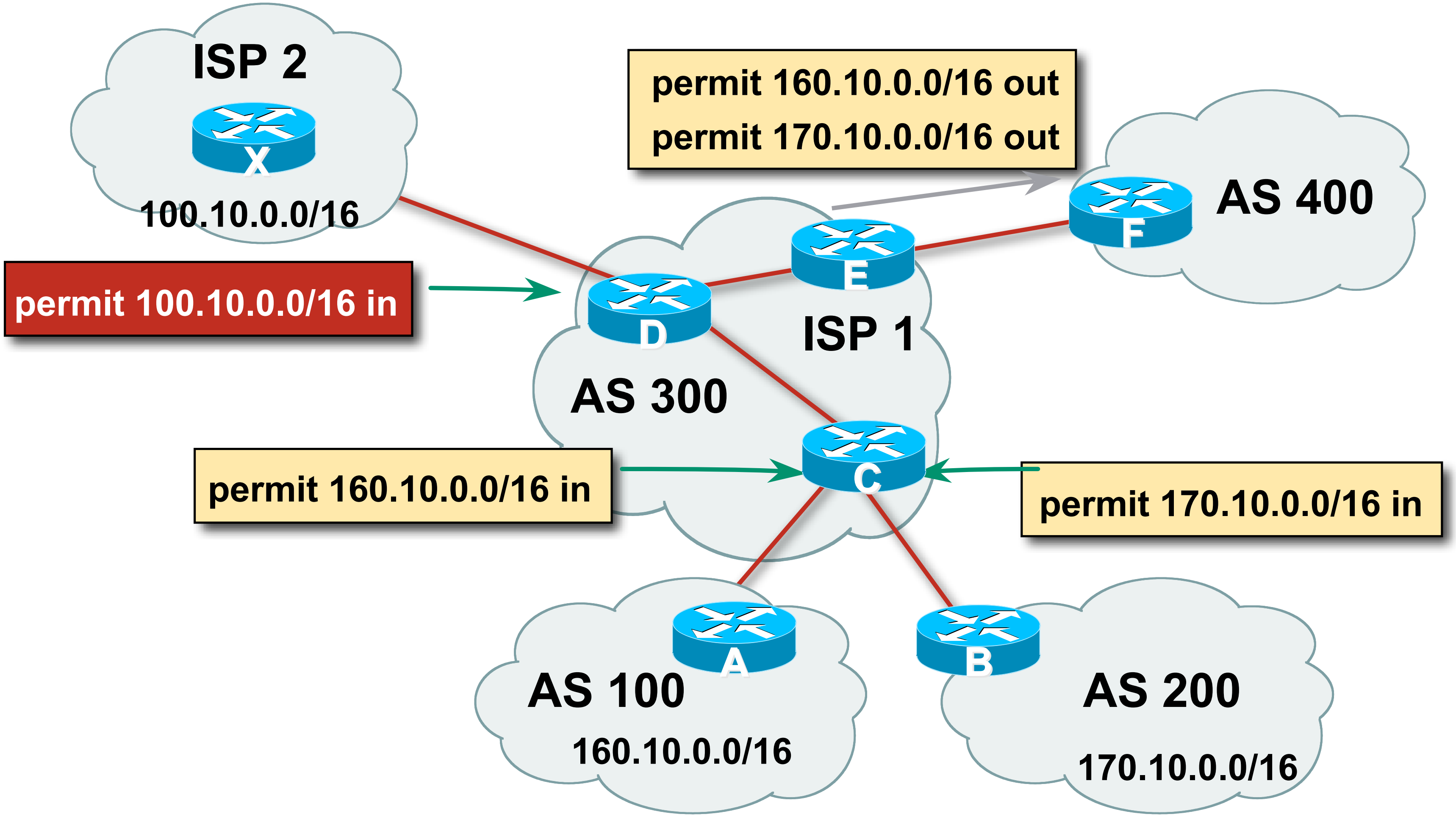
  Other implementations said that absence of metric was equivalent to 2^{32}-1 (highest possible) or 2^{32}-2

  Potential for "metric confusion"

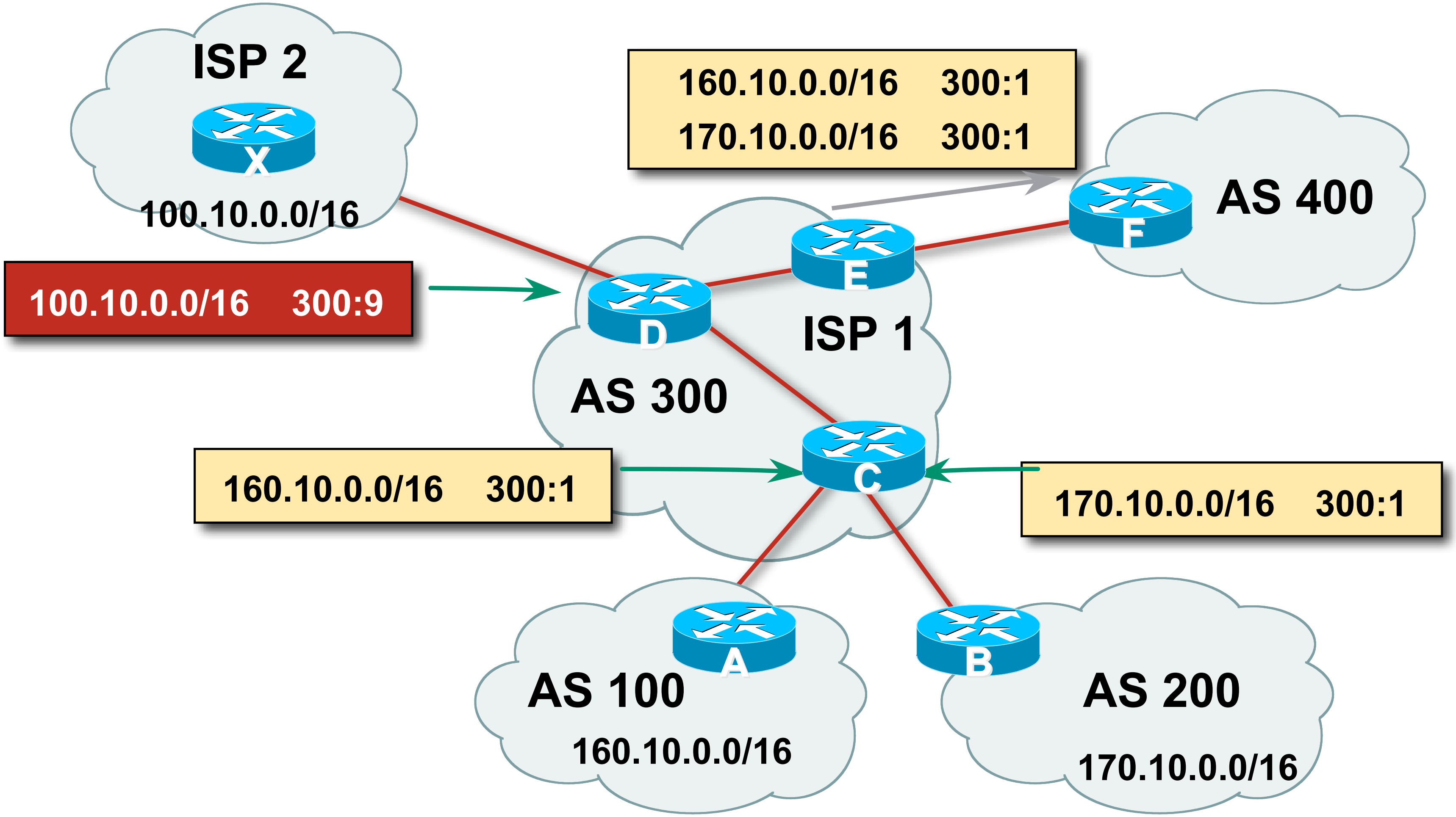
Community

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

Community Example (before)



Community Example (after)



Well-Known Communities

- Several well known communities

  www.iana.org/assignments/bgp-well-known-communities

- no-export 65535:65281

  do not advertise to any eBGP peers

- no-advertise 65535:65282

  do not advertise to any BGP peer

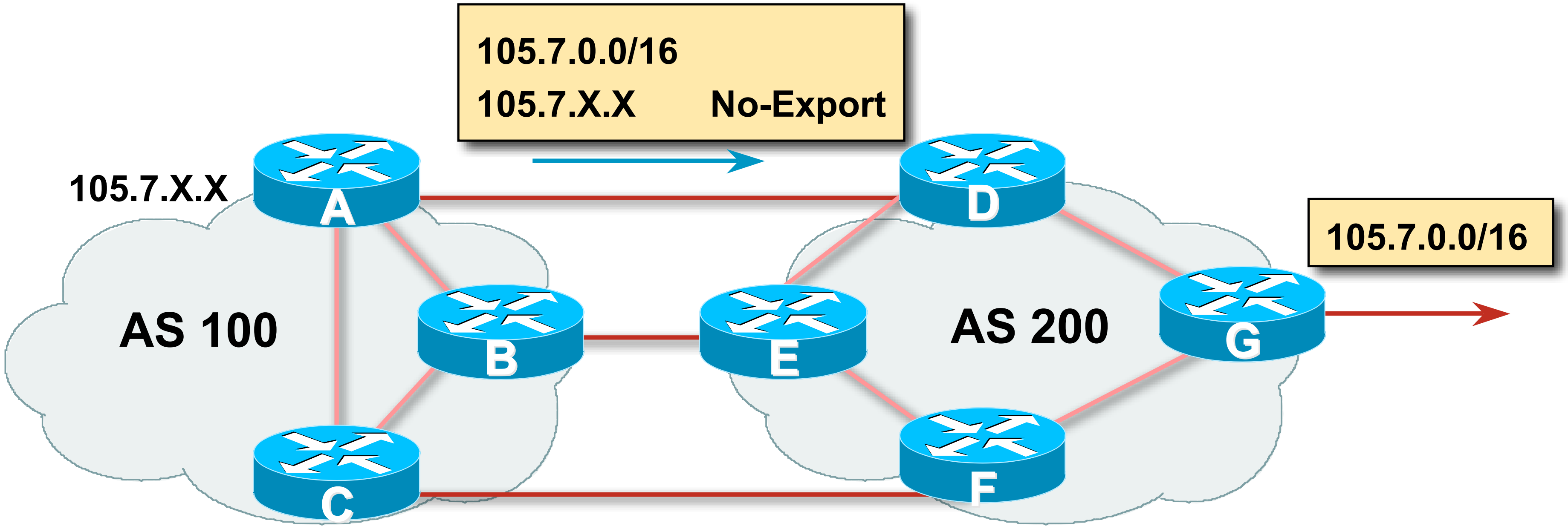
- no-export-subconfed 65535:65283

  do not advertise outside local AS (only used with confederations)

- no-peer 65535:65284

  do not advertise to bi-lateral peers (RFC3765)

No-Export Community

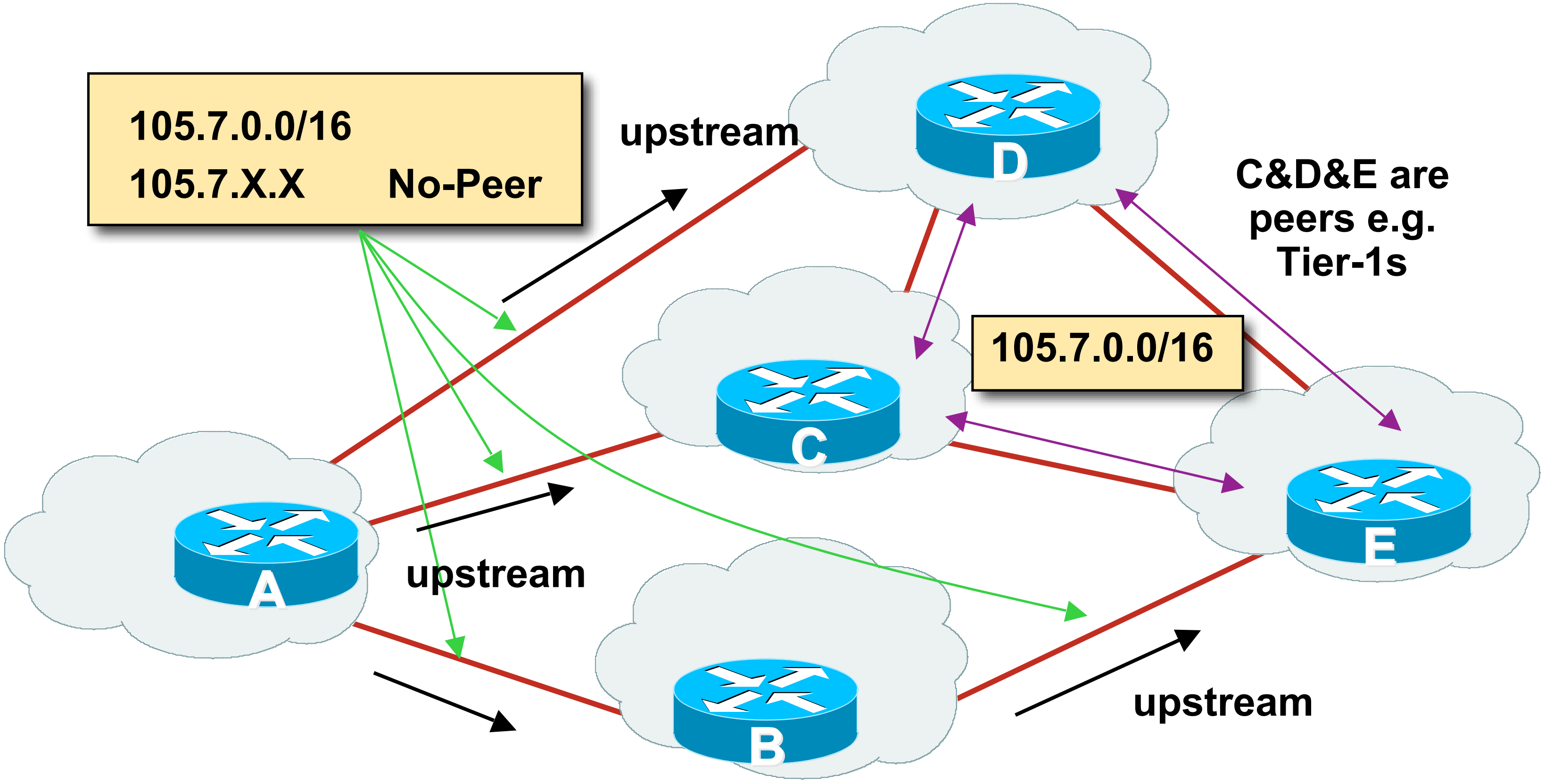


- AS100 announces aggregate and subprefixes

  Intention is to improve loadsharing by leaking subprefixes

- Subprefixes marked with no-export community
- Router G in AS200 does not announce prefixes with no-export community set

No-Peer Community



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

  They are only sent to upstream providers

Community
Implementation details

- Community is an optional attribute

  Some implementations send communities to iBGP peers by default, some do not

  Some implementations send communities to eBGP peers by default, some do not

- Being careless can lead to community “confusion”

  ISPs need consistent community policy within their own networks

  And they need to inform peers, upstreams and customers about their community expectations

BGP Path Selection Algorithm

Why Is This the Best Path?

BGP Path Selection Algorithm for IOS Part One

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

BGP Path Selection Algorithm for IOS Part Two

- Lowest origin code

  IGP < EGP < incomplete

- Lowest Multi-Exit Discriminator (MED)

  If bgp deterministic-med, order the paths before comparing

  (BGP spec does not specify in which order the paths should be compared. This means best path depends on order in which the paths are compared.)

  If bgp always-compare-med, then compare for all paths

  otherwise MED only considered if paths are from the same AS (default)

BGP Path Selection Algorithm for IOS Part Three

- Prefer eBGP path over iBGP path
- Path with lowest IGP metric to next-hop
- Lowest router-id (originator-id for reflected routes)
- Shortest Cluster-List

  Client **must** be aware of Route Reflector attributes!
- Lowest neighbour IP address

BGP Path Selection Algorithm

- In multi-vendor environments:

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

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

  Watch out for possible MED confusion

Applying Policy with BGP

Controlling Traffic Flow & Traffic Engineering

Applying Policy in BGP: Why?

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

Applying Policy in BGP: How?

- Policies are applied by:
  - Setting BGP attributes (local-pref, MED, AS-PATH, community), thereby influencing the path selection process
  - Advertising or Filtering prefixes
  - Advertising or Filtering prefixes according to ASN and AS-PATHs
  - Advertising or Filtering prefixes according to Community membership

Applying Policy with BGP: Tools

- Most implementations have tools to apply policies to BGP:

  Prefix manipulation/filtering

  AS-PATH manipulation/filtering

  Community Attribute setting and matching

- Implementations also have policy language which can do various match/set constructs on the attributes of chosen BGP routes

BGP Capabilities

Extending BGP

BGP Capabilities

- Documented in RFC2842
- Capabilities parameters passed in BGP open message
- Unknown or unsupported capabilities will result in NOTIFICATION message
- Codes:

  0 to 63 are assigned by IANA by IETF consensus

  64 to 127 are assigned by IANA “first come first served”

  128 to 255 are vendor specific

BGP Capabilities

Current capabilities are:

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

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

BGP Capabilities

- Multiprotocol extensions

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

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

  Another tutorial (or many!)

- Route refresh is a well known scaling technique – covered shortly
- 32-bit ASNs have recently arrived
- The other capabilities are still in development or not widely implemented or deployed yet

BGP for Internet Service Providers

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

BGP Scaling Techniques

BGP Scaling Techniques

- How does a service provider:
    - Scale the iBGP mesh beyond a few peers?
    - Implement new policy without causing flaps and route churning?
    - Keep the network stable, scalable, as well as simple?

BGP Scaling Techniques

- Route Refresh
- Route Reflectors
- Confederations

Dynamic Reconfiguration

Route Refresh

Route Refresh

- BGP peer reset required after every policy change
  - Because the router does not store prefixes which are rejected by policy
- Hard BGP peer reset:
  - Terminates BGP peering & Consumes CPU
  - Severely disrupts connectivity for all networks
- Soft BGP peer reset (or Route Refresh):
  - BGP peering remains active
  - Impacts only those prefixes affected by policy change

Route Refresh Capability

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

Dynamic Reconfiguration

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

Consider the impact to be equivalent to a router reboot

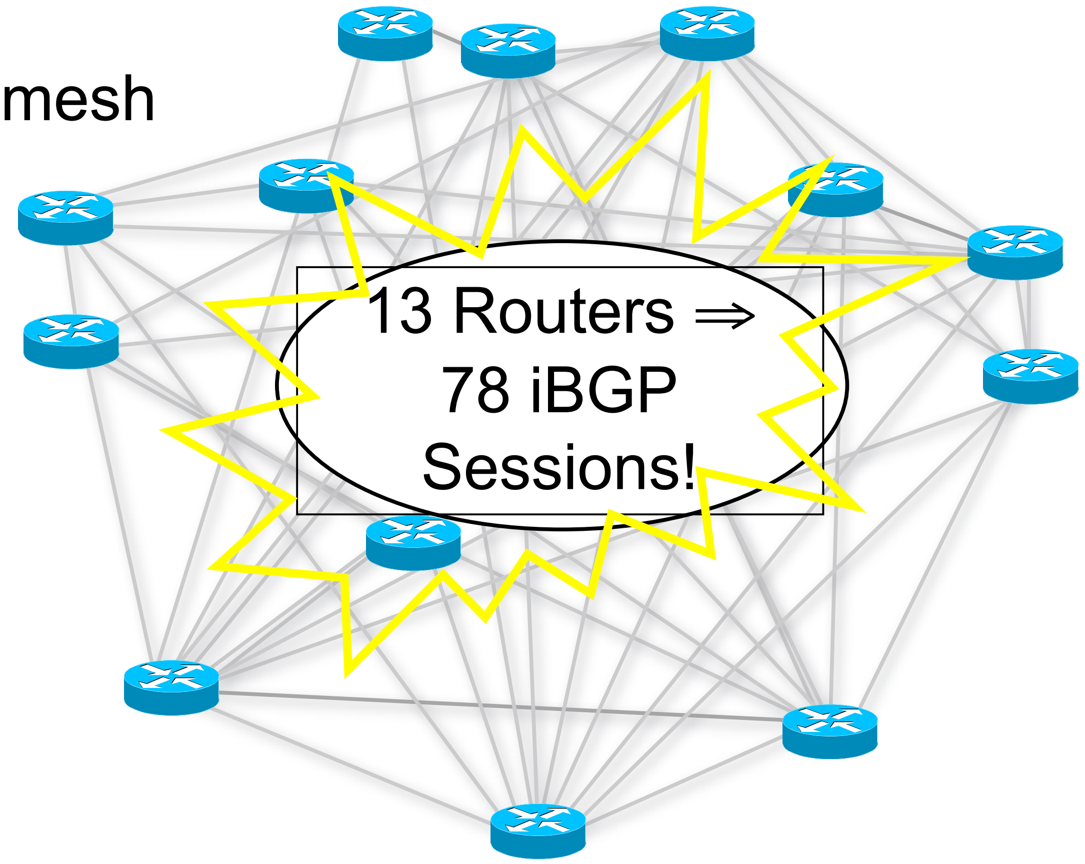
Route Reflectors

Scaling the iBGP mesh

Scaling iBGP mesh

- Avoid ½n(n-1) iBGP mesh

n=1000 ⇒ nearly half a million ibgp sessions!

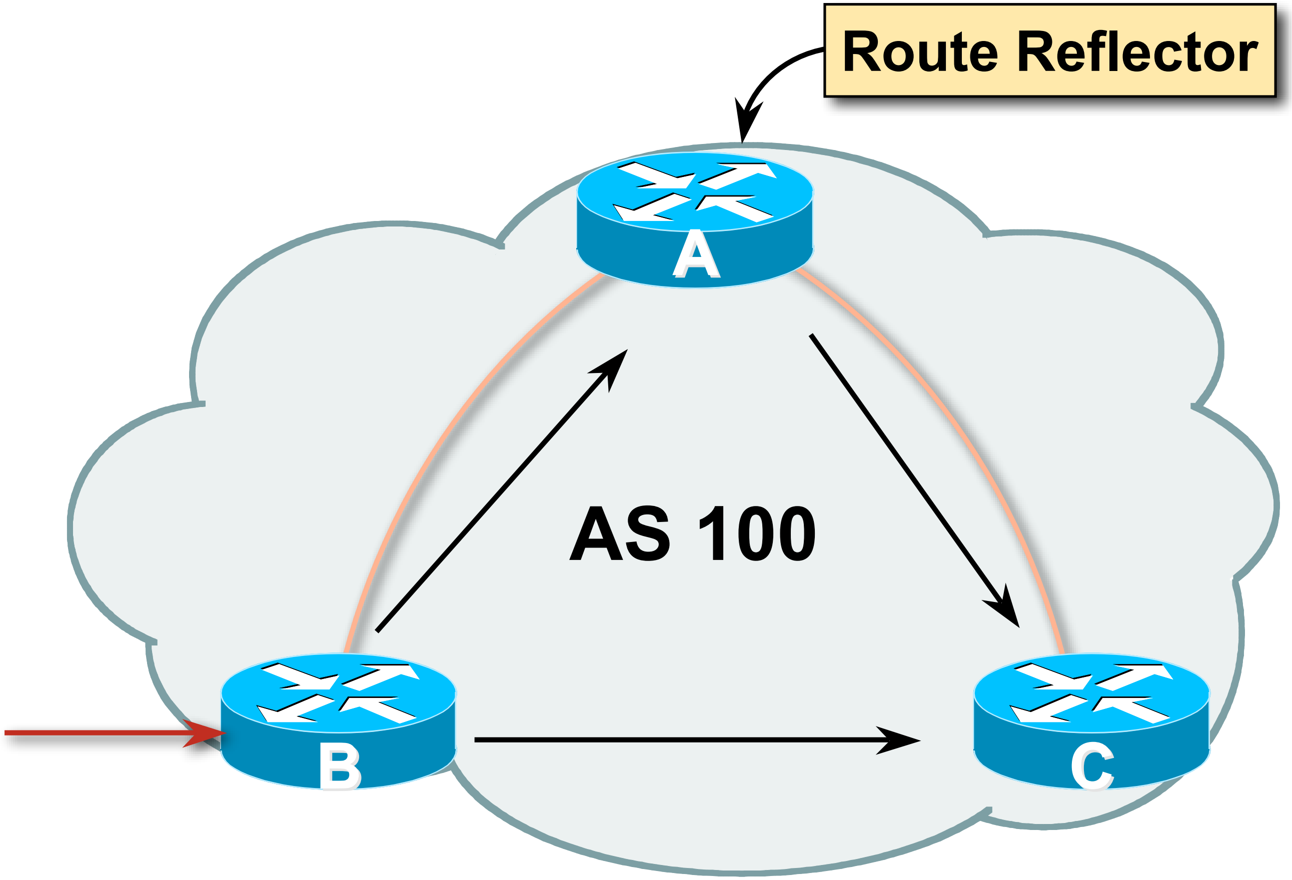


- Two solutions

  Route reflector – simpler to deploy and run

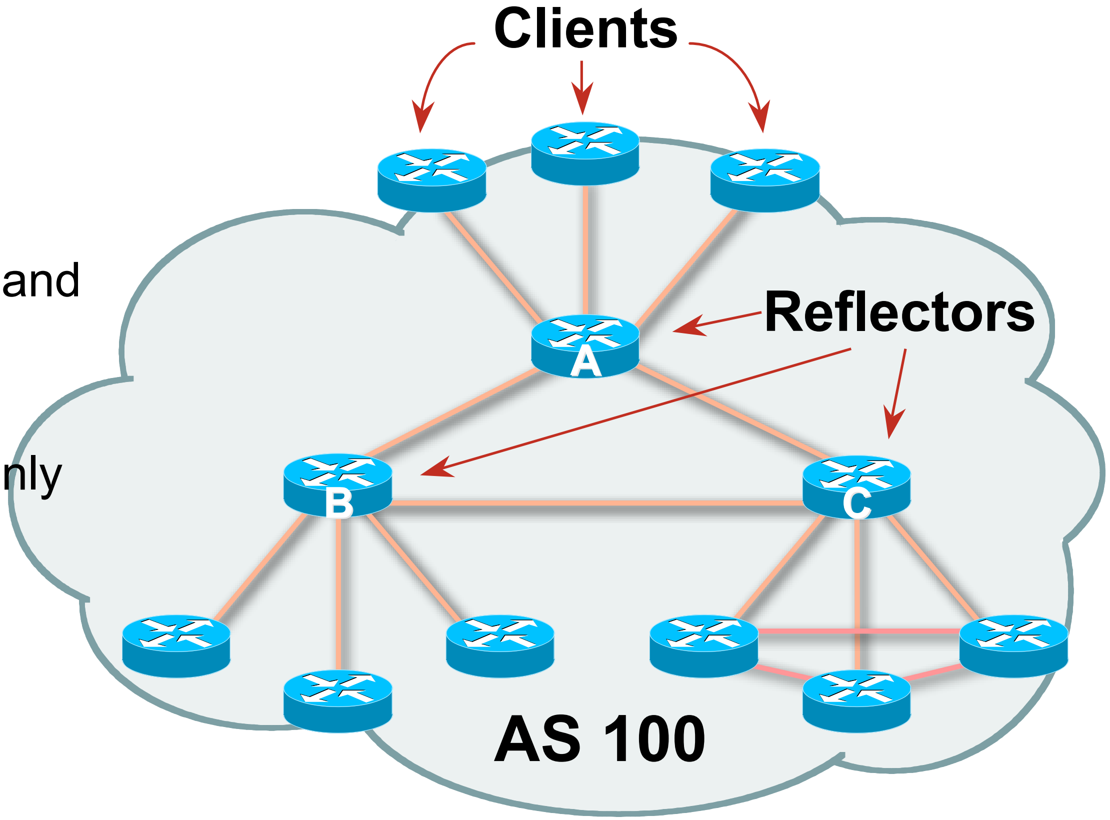
  Confederation – more complex, has corner case advantages

Route Reflector: Principle



Route Reflector

- Reflector receives path from clients and non-clients
- Selects best path
- If best path is from client, reflect to other clients and non-clients
- If best path is from non-client, reflect to clients only
- Non-meshed clients
- Described in RFC4456



Route Reflector: Topology

- Divide the backbone into multiple clusters
- At least one route reflector and few clients per cluster
- Route reflectors are fully meshed
- Clients in a cluster could be fully meshed
- Single IGP to carry next hop and local routes

Route Reflector: Loop Avoidance

- Originator_ID attribute

  Carries the RID of the originator of the route in the local AS (created by the RR)

- Cluster_list attribute

  The local cluster-id is added when the update is sent by the RR

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

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

Route Reflector: Redundancy

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

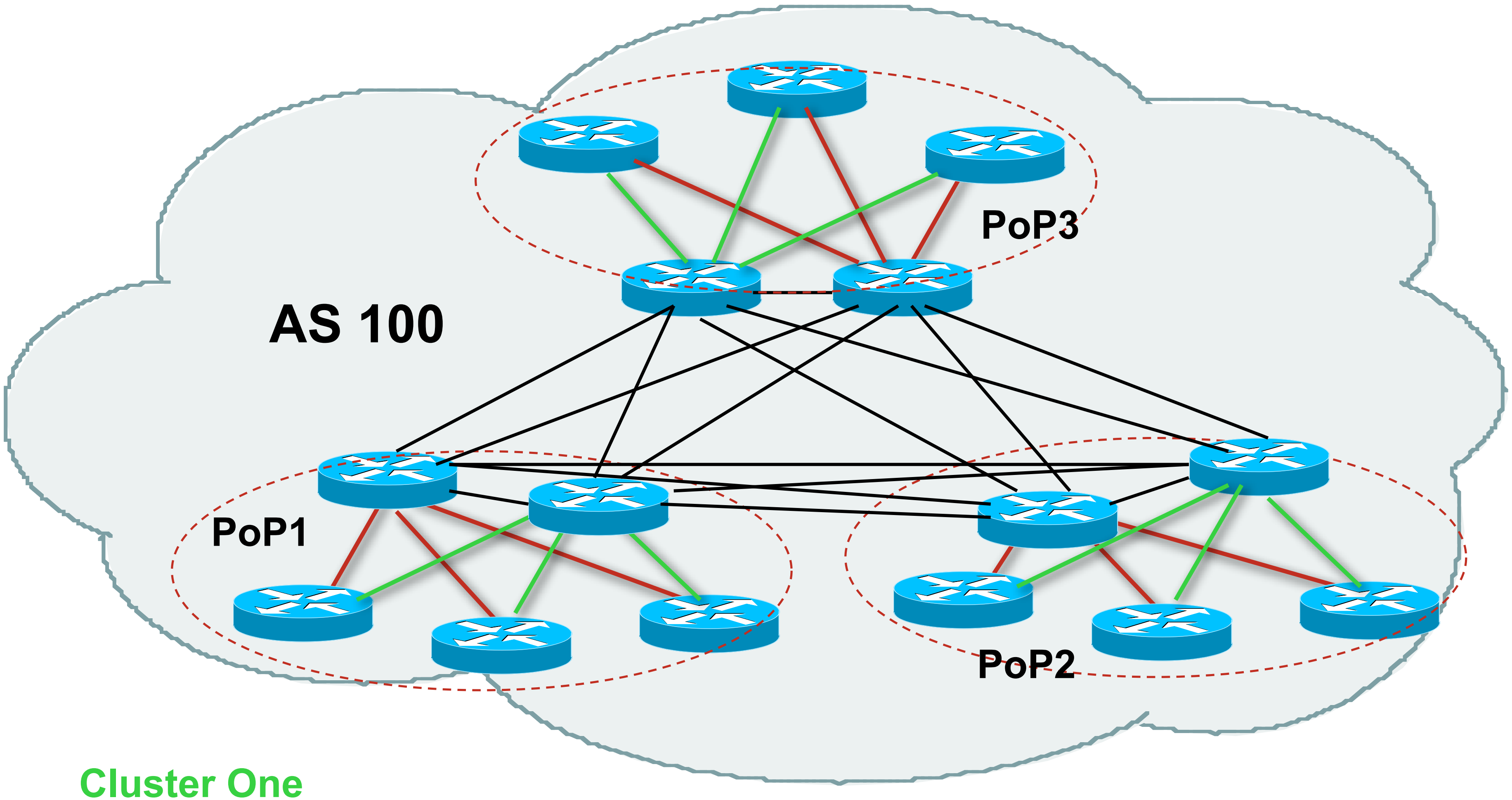
  All RRs in the cluster must have the same cluster-id (otherwise it is a different cluster)

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

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

  → Each client has two RRs = redundancy

Route Reflector: Redundancy



Cluster One

Cluster Two

Route Reflector: Benefits

- Solves iBGP mesh problem
- Packet forwarding is not affected
- Normal BGP speakers co-exist
- Multiple reflectors for redundancy
- Easy migration
- Multiple levels of route reflectors

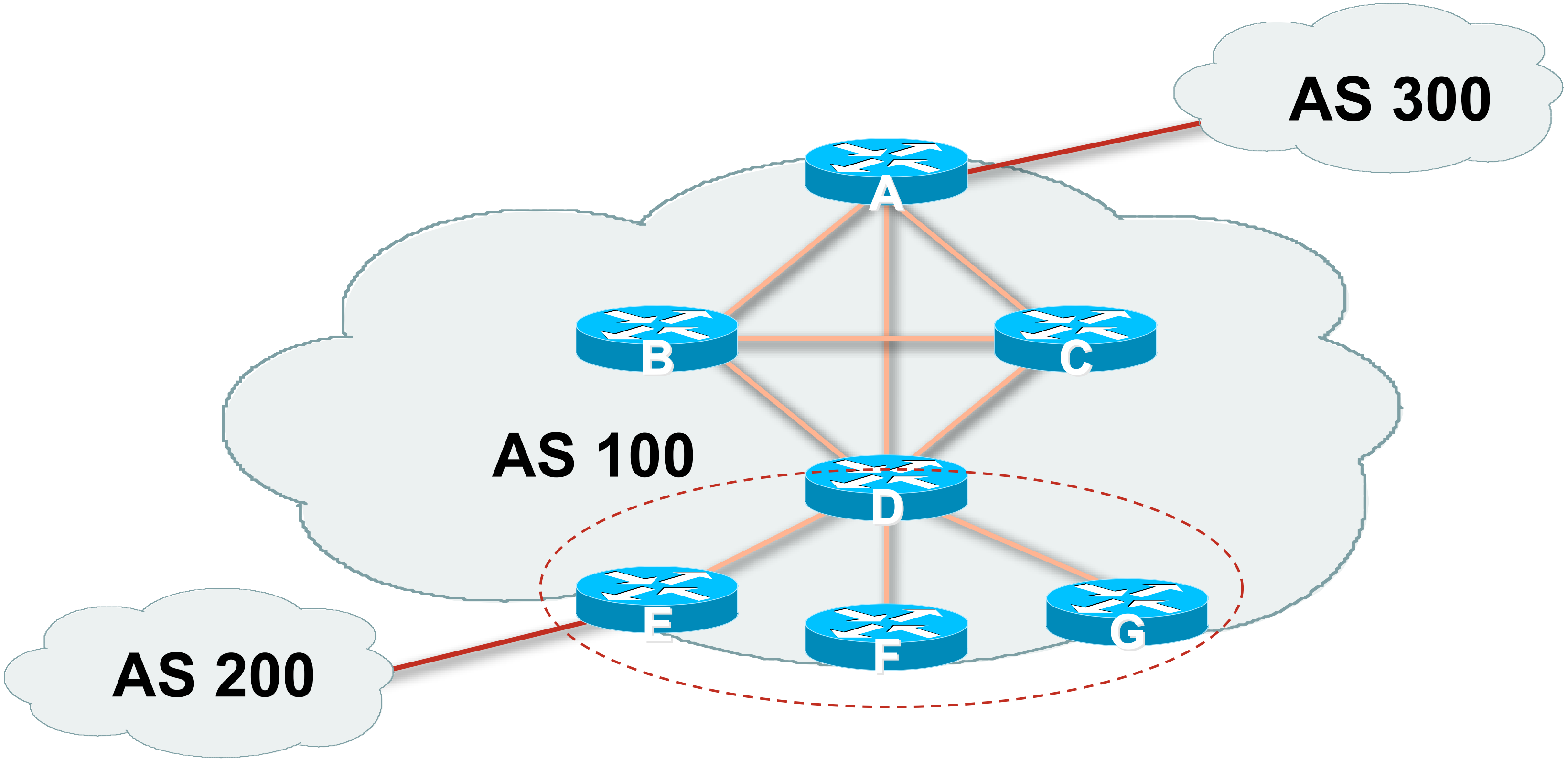
Route Reflector: Deployment

- Where to place the route reflectors?
  - Always follow the physical topology!
  - This will guarantee that the packet forwarding won’t be affected
- Typical ISP network:
  - PoP has two core routers
  - Core routers are RR for the PoP
  - Two overlaid clusters

Route Reflector: Migration

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

Route Reflector: Migration



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

BGP Confederations

Confederations

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

Confederations (Cont.)

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

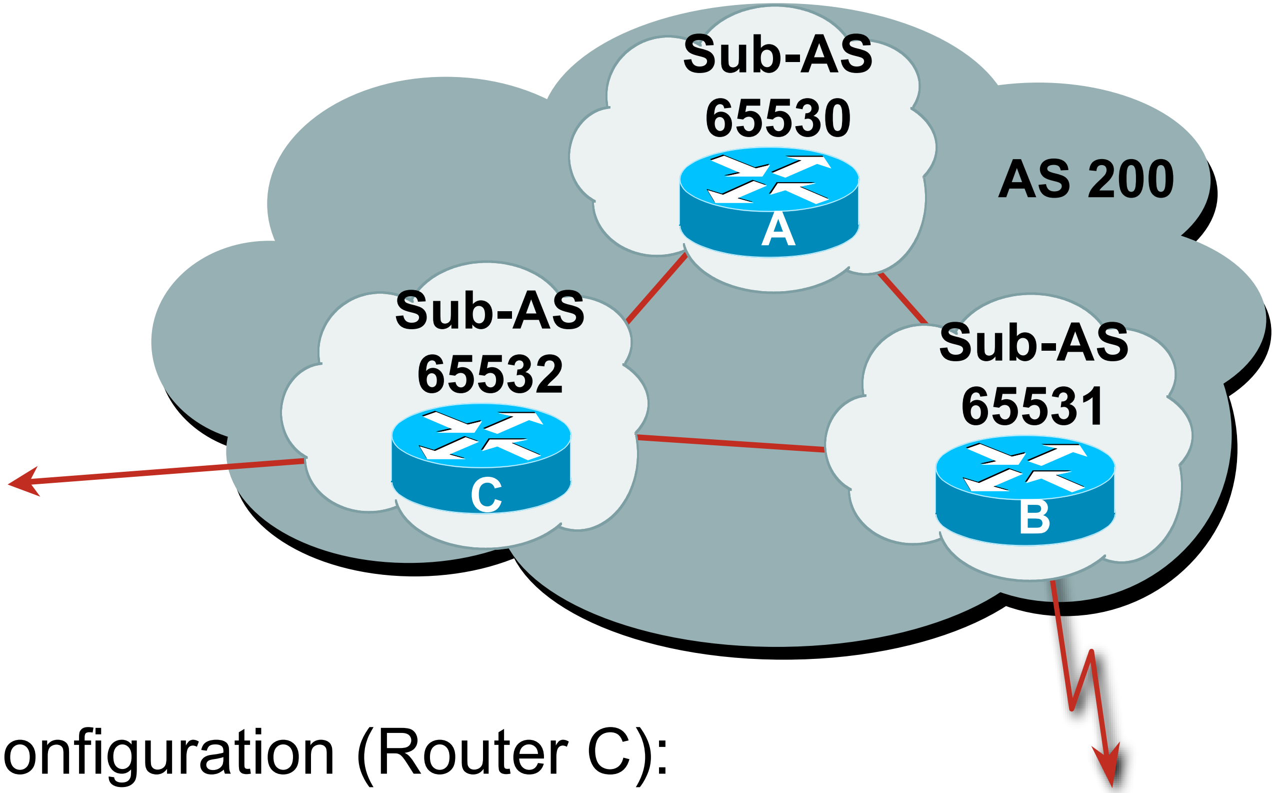
  Each sub-AS uses a number from the private AS range (64512-65534)

- iBGP speakers in each sub-AS are fully meshed

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

  Can also use Route-Reflector within sub-AS

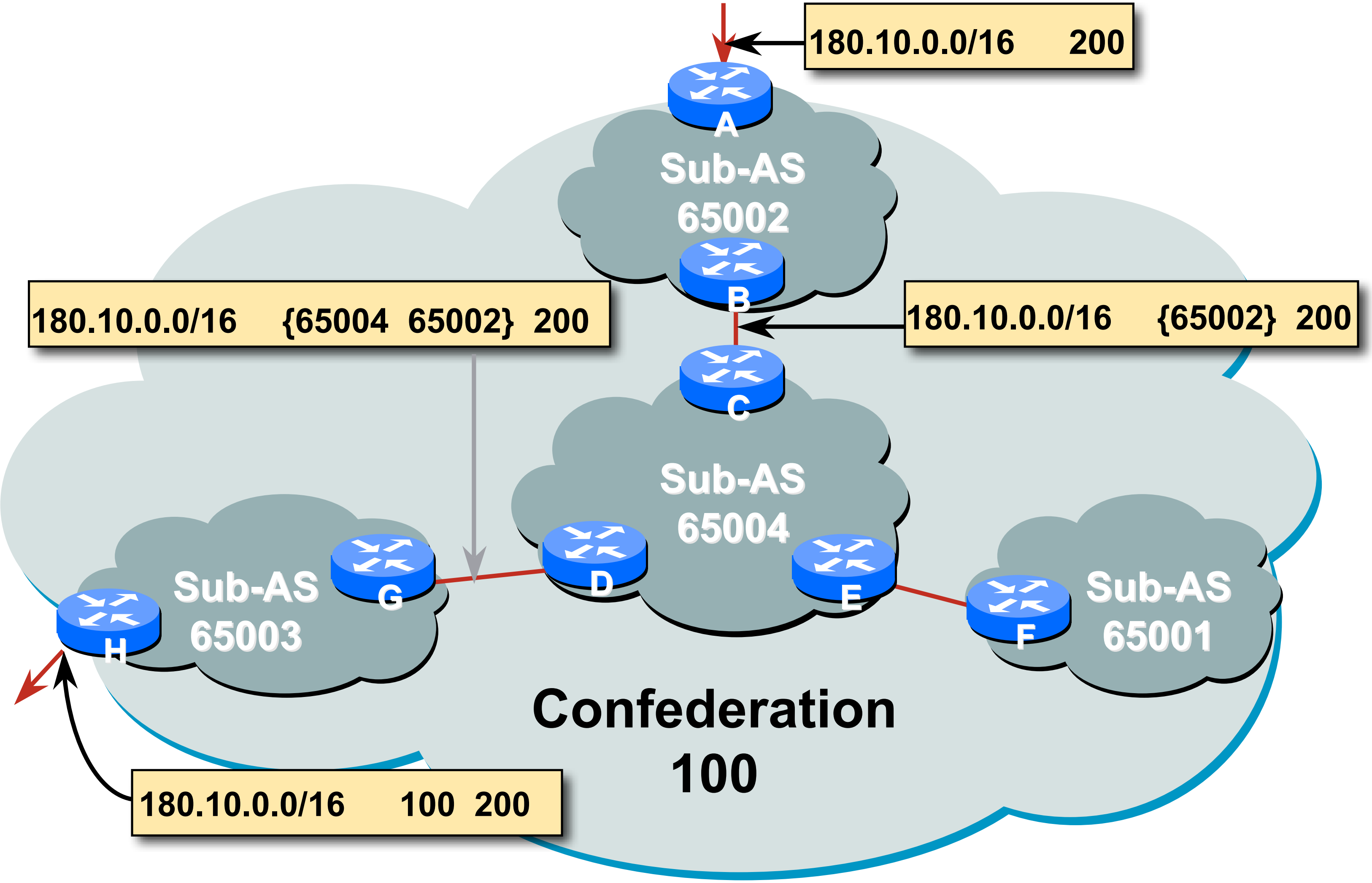
Confederations



- Configuration (Router C):

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

Confederations: AS-Sequence



180.10.0.0/16 200

180.10.0.0/16 {65002} 200

180.10.0.0/16 {65004 65002} 200

180.10.0.0/16 100 200

Route Propagation Decisions

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

RRs or Confederations

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

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

More points about Confederations

- Can ease "absorbing" other ISPs into you ISP – e.g., if one ISP buys another

    Or can use AS masquerading feature available in some implementations to do a similar thing

- Can use route-reflectors with confederation sub-AS to reduce the sub-AS iBGP mesh

Route Flap Damping

Network Stability for the 1990s

Network Instability for the 21st Century!

Route Flap Damping

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

Route Flap Damping

- Route flap
  - Going up and down of path or change in attribute
    - BGP WITHDRAW followed by UPDATE = 1 flap
    - eBGP neighbour going down/up is NOT a flap
  - Ripples through the entire Internet
  - Wastes CPU
- Damping aims to reduce scope of route flap propagation

Route Flap Damping (continued)

- Requirements

  Fast convergence for normal route changes

  History predicts future behaviour

  Suppress oscillating routes

  Advertise stable routes

- Implementation described in RFC 2439

Operation

- Add penalty (1000) for each flap

  Change in attribute gets penalty of 500

- Exponentially decay penalty

  half life determines decay rate

- Penalty above suppress-limit

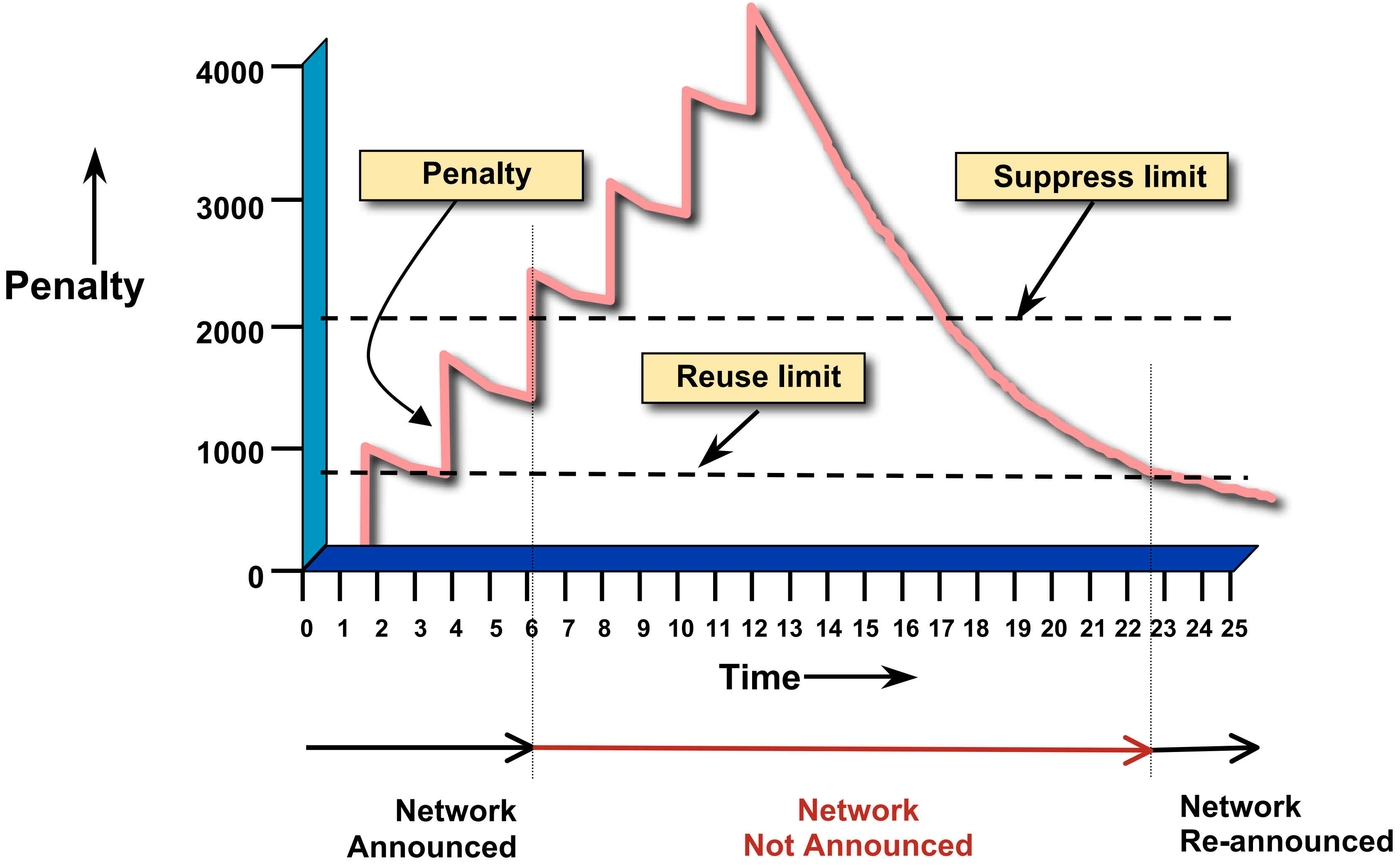
  do not advertise route to BGP peers

- Penalty decayed below reuse-limit

  re-advertise route to BGP peers

  penalty reset to zero when it is half of reuse-limit

Operation



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

Operation

- Only applied to inbound announcements from eBGP peers
- Alternate paths still usable
- Controllable by at least:

  Half-life

  reuse-limit

  suppress-limit

  maximum suppress time

Configuration

- Implementations allow various policy control with flap damping

  Fixed damping, same rate applied to all prefixes

  Variable damping, different rates applied to different ranges of prefixes and prefix lengths

Route Flap Damping History

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

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

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

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

Serious Problems:

- "Route Flap Damping Exacerbates Internet Routing Convergence“

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

- “What is the sound of one route flapping?”

  Tim Griffin, June 2002

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

- “Happy Packets”

  Closely related work by Randy Bush et al

Problem 1:

- One path flaps:

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

  Those peers see change in best path, flap counter incremented

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

Problem 2:

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

Solution:

- Do **NOT** use Route Flap Damping whatever you do!
- RFD will unnecessarily impair access

  to your network and

  to the Internet
- More information contained in RIPE Routing Working Group recommendations:

  www.ripe.net/ripe/docs/ripe-378.[pdf,html,txt]

BGP for Internet Service Providers

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

Deploying BGP in an ISP Network

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

Deploying BGP

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

The role of IGP and iBGP

Ships in the night?

Or

Good foundations?

BGP versus OSPF/ISIS

- Internal Routing Protocols (IGPs)

  examples are ISIS and OSPF

  used for carrying infrastructure addresses

  NOT used for carrying Internet prefixes or customer prefixes

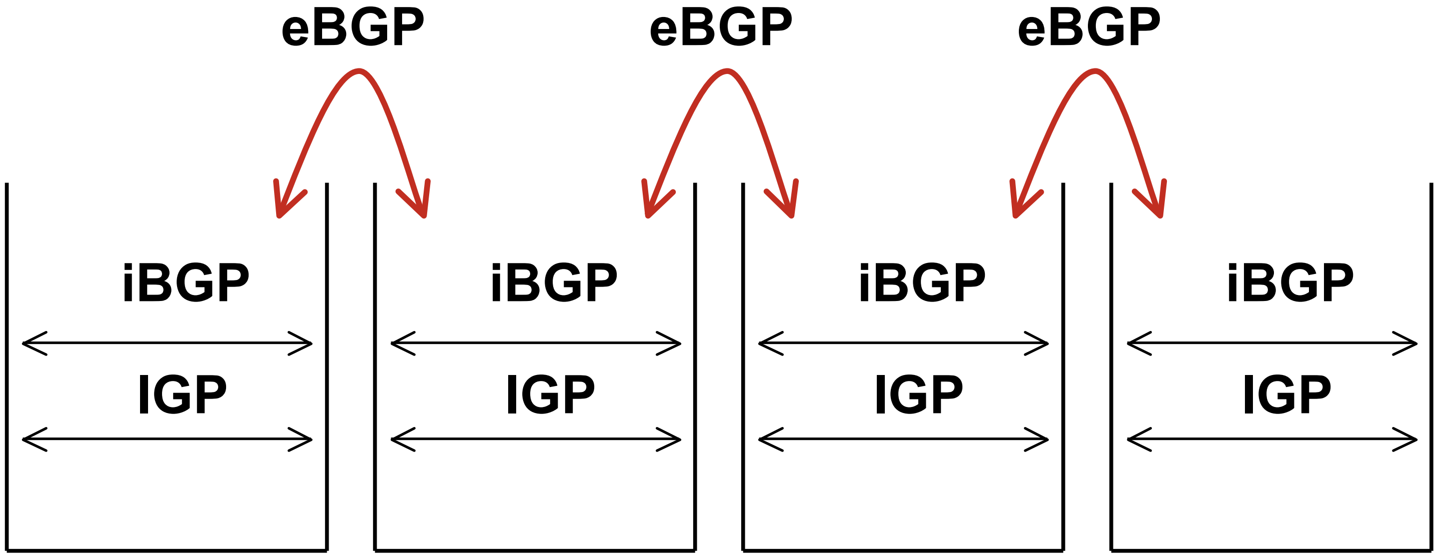
  design goal is to minimise number of prefixes in IGP to aid scalability and rapid convergence

BGP versus OSPF/ISIS

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

BGP/IGP model used in ISP networks

- Model representation



BGP versus OSPF/ISIS

- DO NOT:

  distribute BGP prefixes into an IGP

  distribute IGP routes into BGP

  use an IGP to carry customer prefixes

- YOUR NETWORK WILL NOT SCALE

Injecting prefixes into iBGP

- Use iBGP to carry customer prefixes

    Don’t ever use IGP

- Point static route to customer interface
- Enter network into BGP process

    Ensure that implementation options are used so that the prefix always remains in iBGP, regardless of state of interface

    i.e. avoid iBGP flaps caused by interface flaps

Aggregation

Quality or Quantity?

Aggregation

- Aggregation means announcing the address block received from the RIR to the other ASes connected to your network
- Subprefixes of this aggregate *may* be:
  - Used internally in the ISP network
  - Announced to other ASes to aid with multihoming
- Unfortunately too many people are still thinking about class Cs, resulting in a proliferation of /24s in the Internet routing table

Aggregation

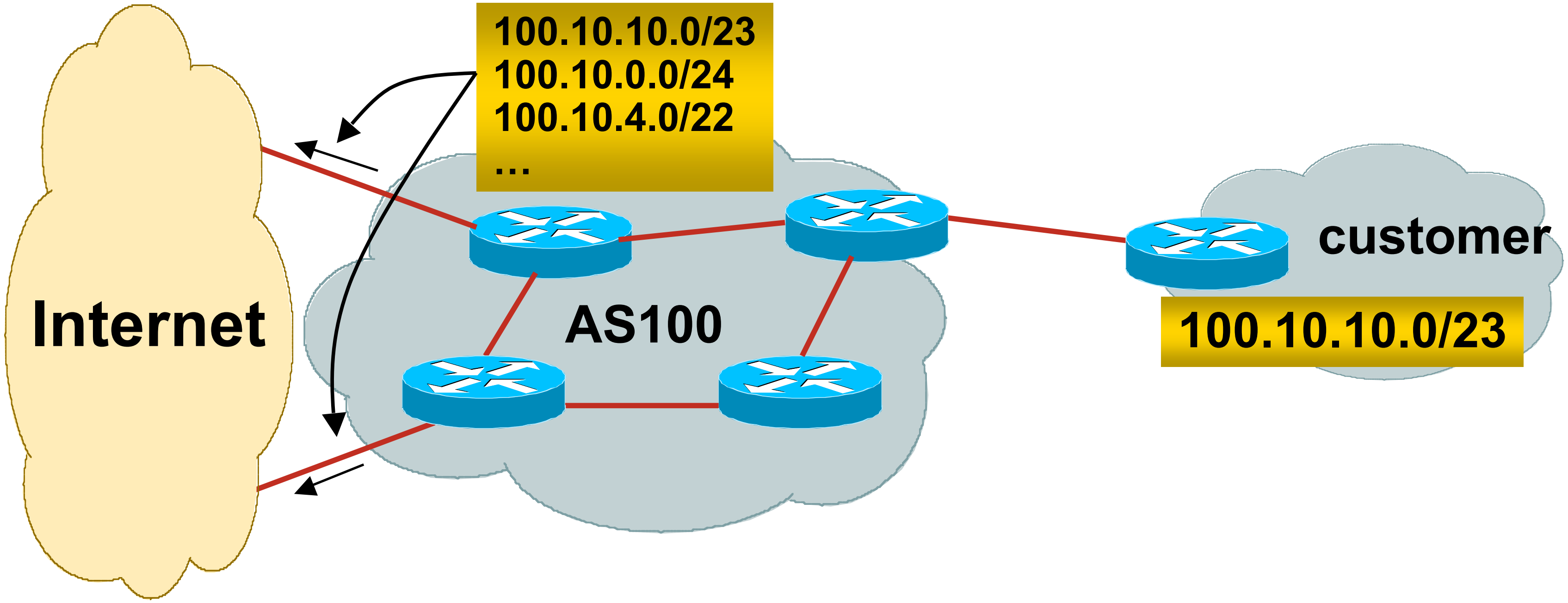
- Address block should be announced to the Internet as an aggregate
- Subprefixes of address block should NOT be announced to Internet unless special circumstances (more later)
- Aggregate should be generated internally

  Not on the network borders!

Announcing an Aggregate

- ISPs who don’t and won’t aggregate are held in poor regard by community
- Registries publish their minimum allocation size
  - Anything from a /20 to a /22 depending on RIR
  - Different sizes for different address blocks
- No real reason to see anything longer than a /22 prefix in the Internet
  - BUT there are currently >124000 /24s!

Aggregation – Example



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

Aggregation – Bad Example

- Customer link goes down

  Their /23 network becomes unreachable

  /23 is withdrawn from AS100’s iBGP

- Their ISP doesn’t aggregate its /19 network block

  /23 network withdrawal announced to peers

  starts rippling through the Internet

  added load on all Internet backbone routers as network is removed from routing table

- Customer link returns

  Their /23 network is now visible to their ISP

  Their /23 network is re-advertised to peers

  Starts rippling through Internet

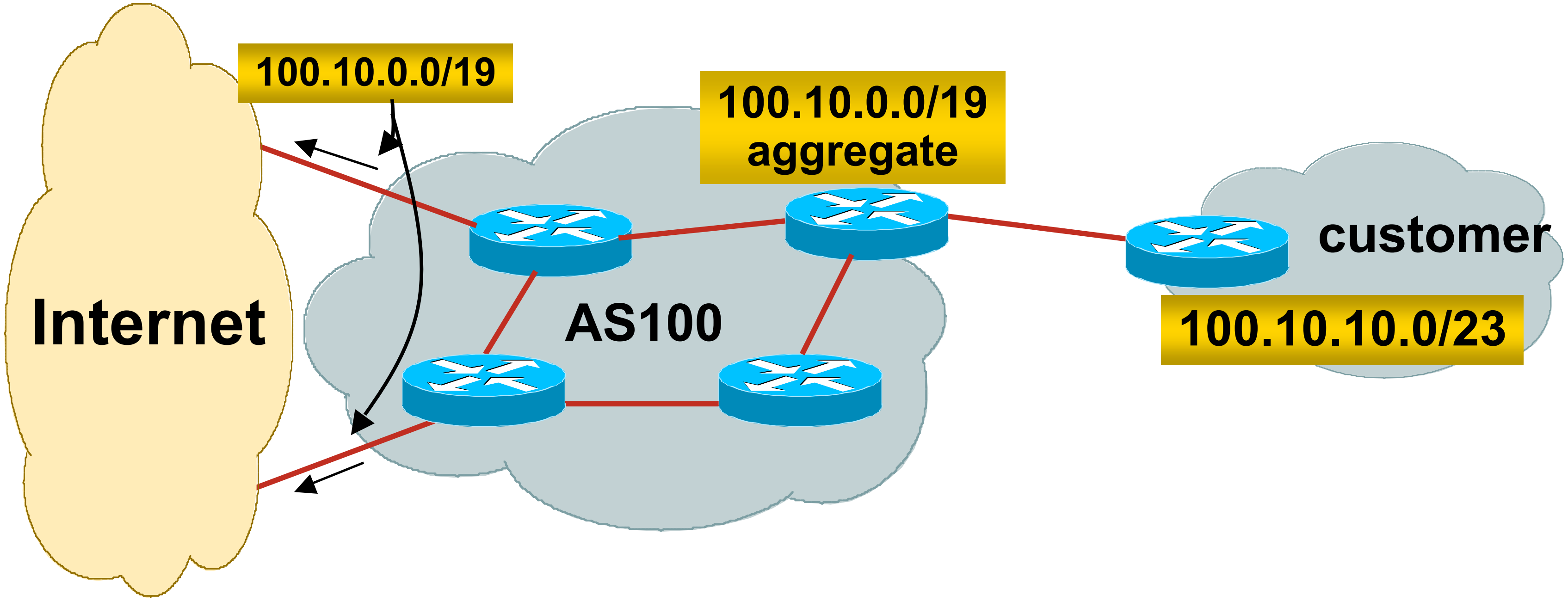
  Load on Internet backbone routers as network is reinserted into routing table

  Some ISP’s suppress the flaps

  Internet may take 10-20 min or longer to be visible

  Where is the Quality of Service???

Aggregation – Example



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

Aggregation – Good Example

- Customer link goes down
  - their /23 network becomes unreachable
  - /23 is withdrawn from AS100's iBGP
- /19 aggregate is still being announced
  - no BGP hold down problems
  - no BGP propagation delays
  - no damping by other ISPs

- Customer link returns
- Their /23 network is visible again
  - The /23 is re-injected into AS100's iBGP
- The whole Internet becomes visible immediately
- Customer has Quality of Service perception

Aggregation – Summary

- Good example is what everyone should do!
  - Adds to Internet stability
  - Reduces size of routing table
  - Reduces routing churn
  - Improves Internet QoS for everyone
- Bad example is what too many still do!
  - Why? Lack of knowledge?
  - Laziness?

The Internet Today (January 2008)

- Current Internet Routing Table Statistics

BGP Routing Table Entries	243144
Prefixes after maximum aggregation	124223
Unique prefixes in Internet	118705
Prefixes smaller than registry alloc	114484
/24s announced	127680
only 5745 /24s are from 192.0.0.0/8	
ASes in use	27085

“The New Swamp”

- Swamp space is name used for areas of poor aggregation
  - The original swamp was 192.0.0.0/8 from the former class C block
    - Name given just after the deployment of CIDR
  - The new swamp is creeping across all parts of the Internet
    - Not just RIR space, but “legacy” space too

"The New Swamp"
RIR Space – February 1999

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

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

“The New Swamp” RIR Space – February 2007

RIR blocks contribute 192490 prefixes or 90% of the Internet Routing Table

Block	Networks	Block	Networks	Block	Networks	Block	Networks
24/8	2930	77/8	1214	118/8	3	203/8	10459
41/8	288	78/8	8	119/8	3	204/8	5569
58/8	1097	79/8	2	120/8	3	205/8	2892
59/8	1152	80/8	2053	121/8	426	206/8	3857
60/8	604	81/8	1695	122/8	698	207/8	4331
61/8	2589	82/8	1564	123/8	534	208/8	4258
62/8	2193	83/8	1172	124/8	1340	209/8	5540
63/8	2967	84/8	1269	125/8	1554	210/8	4759
64/8	5501	85/8	1891	126/8	41	211/8	2733
65/8	3917	86/8	800	189/8	169	212/8	2900
66/8	6575	87/8	1157	190/8	1077	213/8	3052
67/8	2015	88/8	847	192/8	6927	216/8	6930
68/8	2770	89/8	1970	193/8	5704	217/8	2615
69/8	3702	90/8	105	194/8	4652	218/8	1561
70/8	1693	91/8	577	195/8	4279	219/8	1197
71/8	1188	96/8	8	196/8	1600	220/8	1988
72/8	2878	97/8	1	198/8	4748	221/8	894
73/8	273	98/8	3	199/8	4184	222/8	1241
74/8	1483	99/8	0	200/8	7482		
75/8	483	116/8	3	201/8	2927		
76/8	194	117/8	3	202/8	10529		

“The New Swamp” Summary

- RIR space shows creeping deaggregation

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

  So their existing 87 /8s will eventually cause 435000 prefix announcements

- Food for thought:

  Remaining 42 unallocated /8s and the 87 RIR /8s combined will cause:

  645000 prefixes with 5000 prefixes per /8 density

  774000 prefixes with 6000 prefixes per /8 density

  Plus 12% due to “non RIR space deaggregation”

  → Routing Table size of 866880 prefixes

“The New Swamp” Summary

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

BGP Report (bgp.potaroo.net)

- 199336 total announcements in October 2006
- 129795 prefixes

  After aggregating including full AS PATH info

  i.e. including each ASN’s traffic engineering

  35% saving possible
- 109034 prefixes

  After aggregating by Origin AS

  i.e. ignoring each ASN’s traffic engineering

  10% saving possible

Deaggregation: The Excuses

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

Efforts to improve aggregation

- The CIDR Report

  Initiated and operated for many years by Tony Bates

  Now combined with Geoff Huston's routing analysis

  www.cidr-report.org

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

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

- RIPE Routing WG aggregation recommendation

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

Efforts to Improve Aggregation
The CIDR Report

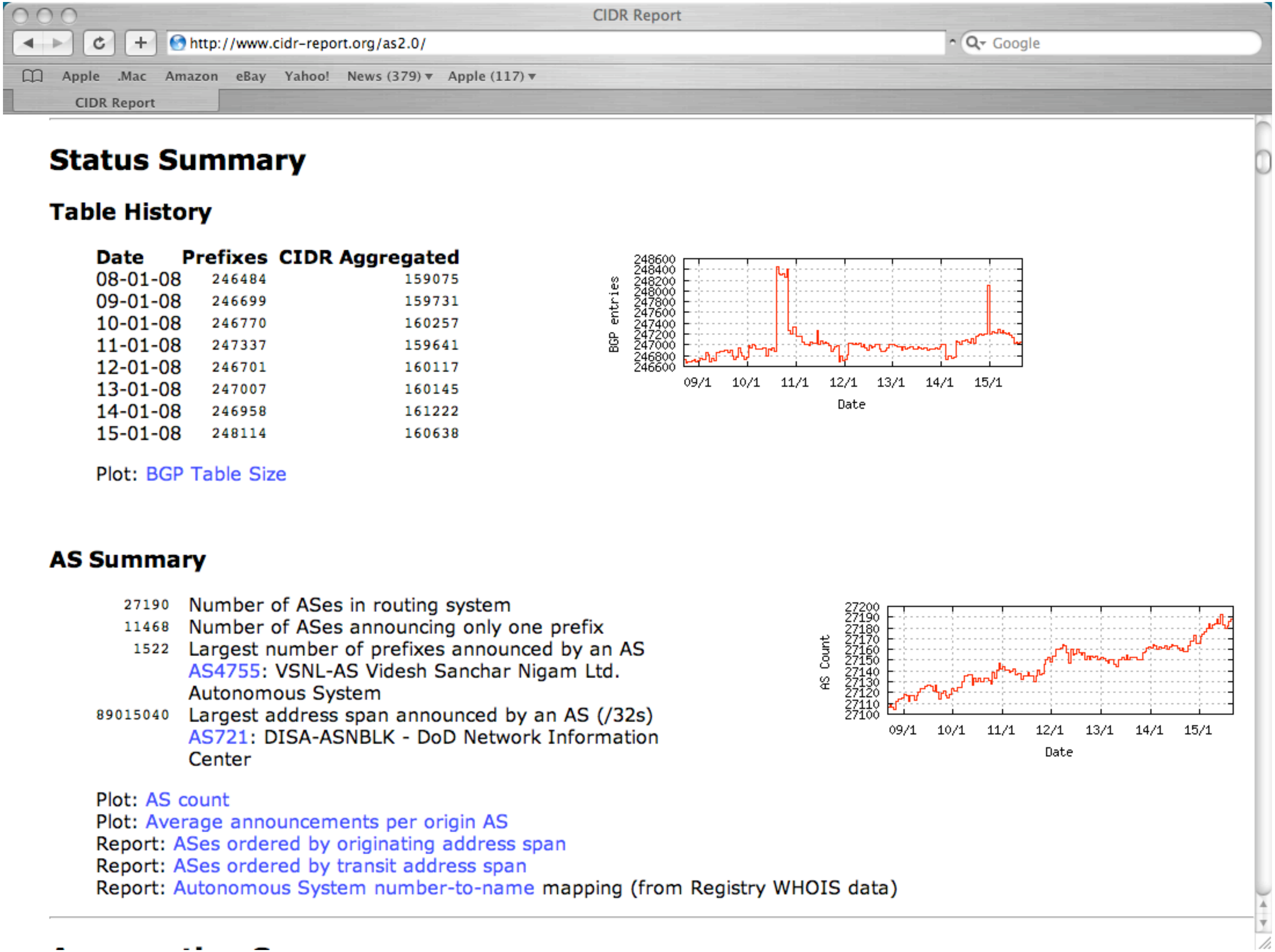
- Also computes the size of the routing table assuming ISPs performed optimal aggregation
- Website allows searches and computations of aggregation to be made on a per AS basis

  Flexible and powerful tool to aid ISPs

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

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

  Very effectively challenges the traffic engineering excuse



Status Summary

Table History

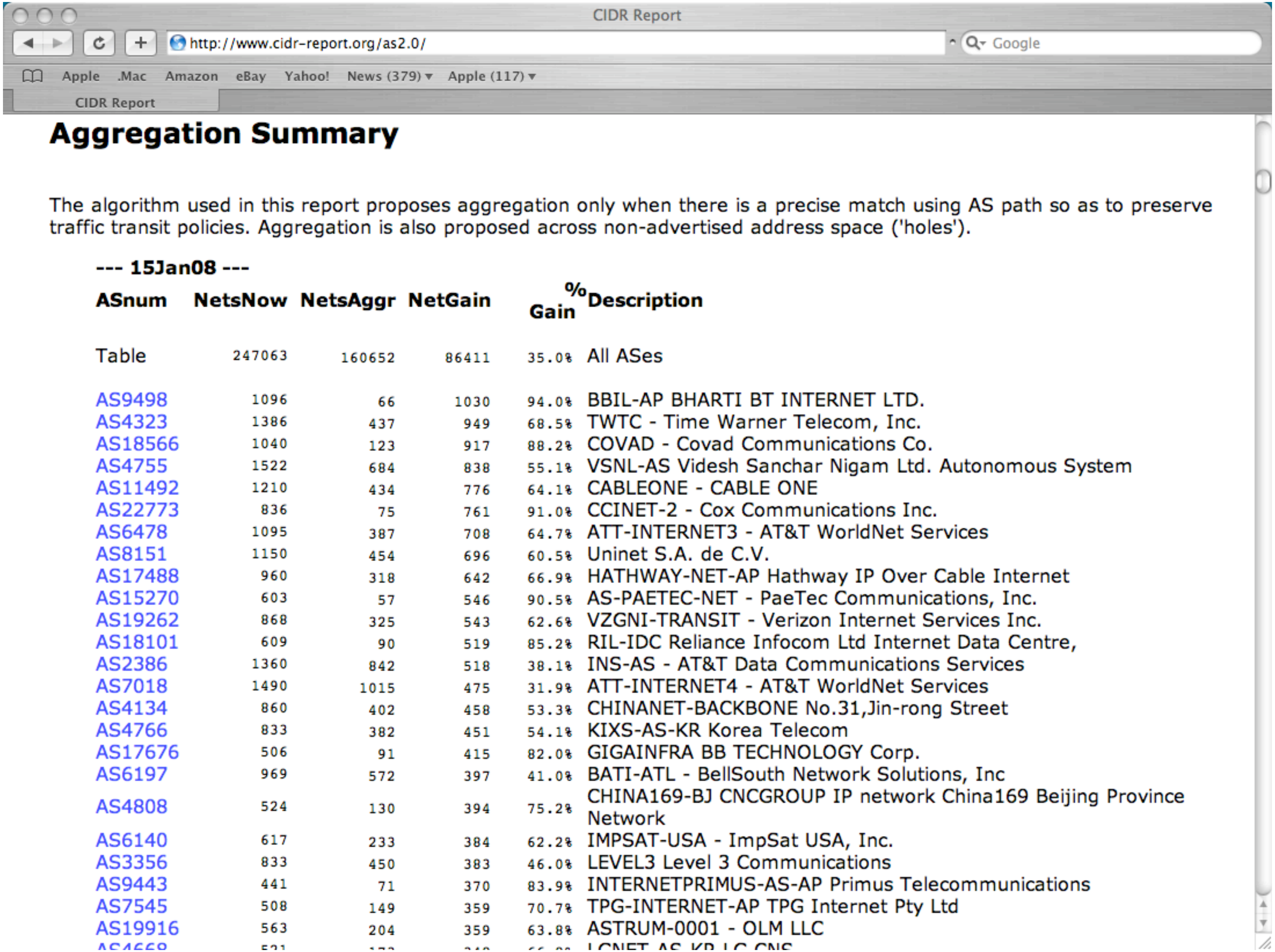
Date	Prefixes	CIDR Aggregated
08-01-08	246484	159075
09-01-08	246699	159731
10-01-08	246770	160257
11-01-08	247337	159641
12-01-08	246701	160117
13-01-08	247007	160145
14-01-08	246958	161222
15-01-08	248114	160638

Plot: BGP Table Size

AS Summary

27190	Number of ASes in routing system
11468	Number of ASes announcing only one prefix
1522	Largest number of prefixes announced by an AS AS4755: VSNL-AS Videsh Sanchar Nigam Ltd. Autonomous System
89015040	Largest address span announced by an AS (/32s) AS721: DISA-ASNBLK - DoD Network Information Center

Plot: AS count
Plot: Average announcements per origin AS
Report: ASes ordered by originating address span
Report: ASes ordered by transit address span
Report: Autonomous System number-to-name mapping (from Registry WHOIS data)

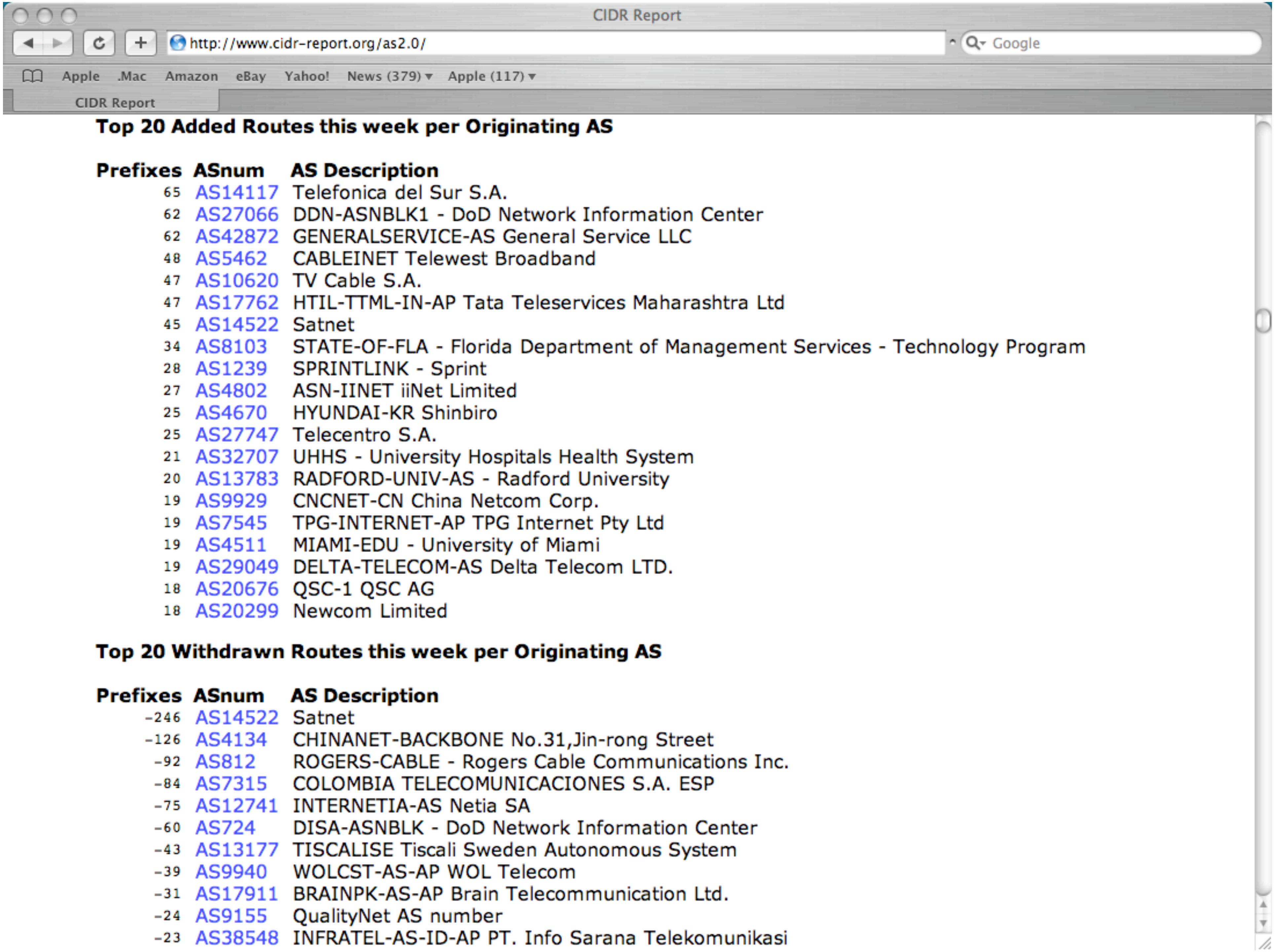


Aggregation Summary

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').

--- 15Jan08 ---

ASnum	NetsNow	NetsAggr	NetGain	% Gain	Description
Table	247063	160652	86411	35.0%	All ASes
AS9498	1096	66	1030	94.0%	BBIL-AP BHARTI BT INTERNET LTD.
AS4323	1386	437	949	68.5%	TWTC - Time Warner Telecom, Inc.
AS18566	1040	123	917	88.2%	COVAD - Covad Communications Co.
AS4755	1522	684	838	55.1%	VSNL-AS Videsh Sanchar Nigam Ltd. Autonomous System
AS11492	1210	434	776	64.1%	CABLEONE - CABLE ONE
AS22773	836	75	761	91.0%	CCINET-2 - Cox Communications Inc.
AS6478	1095	387	708	64.7%	ATT-INTERNET3 - AT&T WorldNet Services
AS8151	1150	454	696	60.5%	Uninet S.A. de C.V.
AS17488	960	318	642	66.9%	HATHWAY-NET-AP Hathway IP Over Cable Internet
AS15270	603	57	546	90.5%	AS-PAETEC-NET - PaeTec Communications, Inc.
AS19262	868	325	543	62.6%	VZGNI-TRANSIT - Verizon Internet Services Inc.
AS18101	609	90	519	85.2%	RIL-IDC Reliance Infocom Ltd Internet Data Centre,
AS2386	1360	842	518	38.1%	INS-AS - AT&T Data Communications Services
AS7018	1490	1015	475	31.9%	ATT-INTERNET4 - AT&T WorldNet Services
AS4134	860	402	458	53.3%	CHINANET-BACKBONE No.31,Jin-rong Street
AS4766	833	382	451	54.1%	KIXS-AS-KR Korea Telecom
AS17676	506	91	415	82.0%	GIGAINFRA BB TECHNOLOGY Corp.
AS6197	969	572	397	41.0%	BATI-ATL - BellSouth Network Solutions, Inc
AS4808	524	130	394	75.2%	CHINA169-BJ CNCGROUP IP network China169 Beijing Province Network
AS6140	617	233	384	62.2%	IMPSAT-USA - ImpSat USA, Inc.
AS3356	833	450	383	46.0%	LEVEL3 Level 3 Communications
AS9443	441	71	370	83.9%	INTERNETPRIMUS-AS-AP Primus Telecommunications
AS7545	508	149	359	70.7%	TPG-INTERNET-AP TPG Internet Pty Ltd
AS19916	563	204	359	63.8%	ASTRUM-0001 - OLM LLC
AS4668	[illegible]	[illegible]	[illegible]	[illegible]	LGNET-AS-KR LG CNS

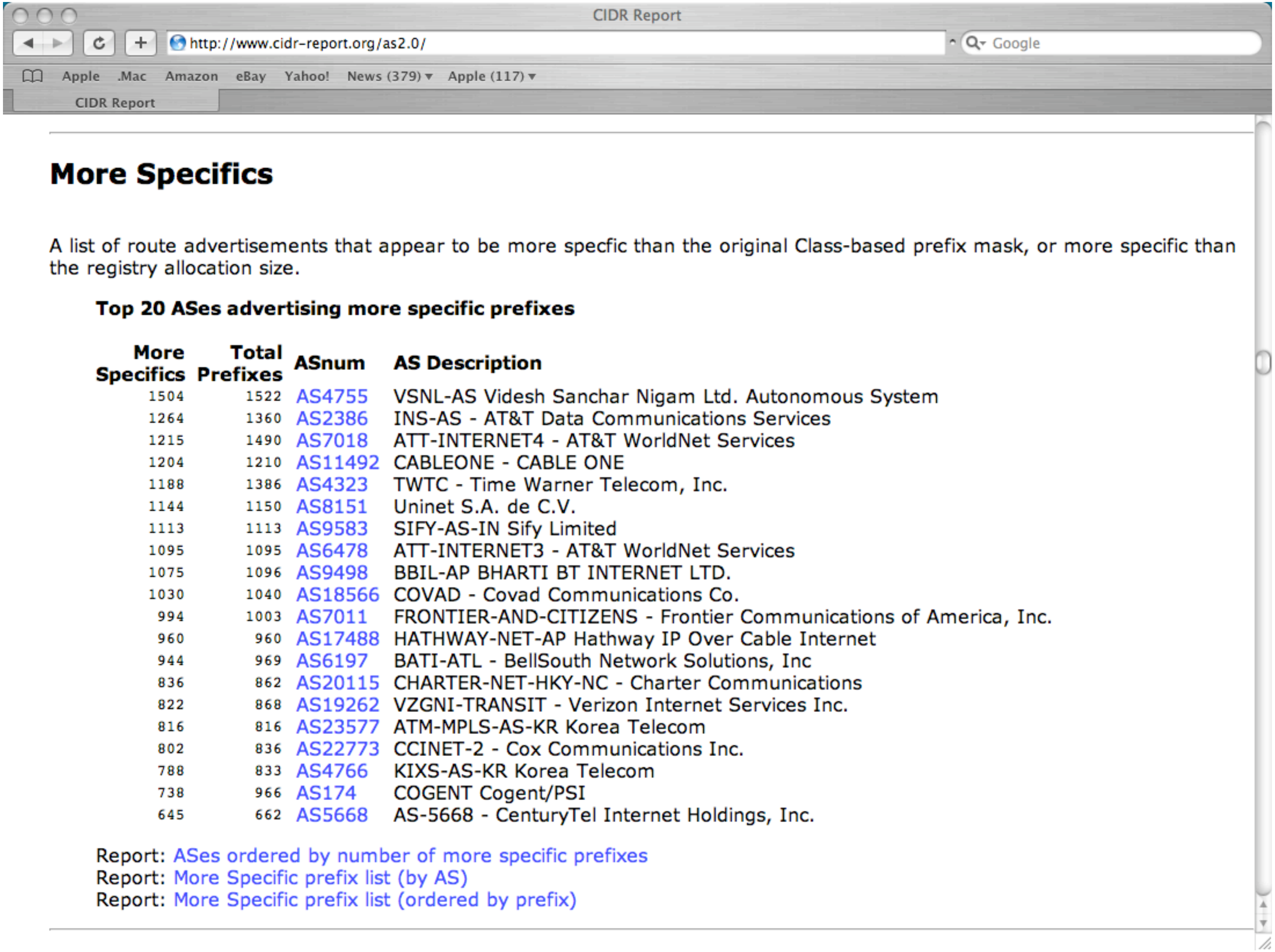


Top 20 Added Routes this week per Originating AS

Prefixes	ASnum	AS Description
65	AS14117	Telefonica del Sur S.A.
62	AS27066	DDN-ASNBLK1 - DoD Network Information Center
62	AS42872	GENERALSERVICE-AS General Service LLC
48	AS5462	CABLEINET Telewest Broadband
47	AS10620	TV Cable S.A.
47	AS17762	HTIL-TTML-IN-AP Tata Teleservices Maharashtra Ltd
45	AS14522	Satnet
34	AS8103	STATE-OF-FLA - Florida Department of Management Services - Technology Program
28	AS1239	SPRINTLINK - Sprint
27	AS4802	ASN-IINET iiNet Limited
25	AS4670	HYUNDAI-KR Shinbiro
25	AS27747	Telecentro S.A.
21	AS32707	UHHS - University Hospitals Health System
20	AS13783	RADFORD-UNIV-AS - Radford University
19	AS9929	CNCNET-CN China Netcom Corp.
19	AS7545	TPG-INTERNET-AP TPG Internet Pty Ltd
19	AS4511	MIAMI-EDU - University of Miami
19	AS29049	DELTA-TELECOM-AS Delta Telecom LTD.
18	AS20676	QSC-1 QSC AG
18	AS20299	Newcom Limited

Top 20 Withdrawn Routes this week per Originating AS

Prefixes	ASnum	AS Description
-246	AS14522	Satnet
-126	AS4134	CHINANET-BACKBONE No.31,Jin-rong Street
-92	AS812	ROGERS-CABLE - Rogers Cable Communications Inc.
-84	AS7315	COLOMBIA TELECOMUNICACIONES S.A. ESP
-75	AS12741	INTERNETIA-AS Netia SA
-60	AS724	DISA-ASNBLK - DoD Network Information Center
-43	AS13177	TISCALISE Tiscali Sweden Autonomous System
-39	AS9940	WOLCST-AS-AP WOL Telecom
-31	AS17911	BRAINPK-AS-AP Brain Telecommunication Ltd.
-24	AS9155	QualityNet AS number
-23	AS38548	INFRATEL-AS-ID-AP PT. Info Sarana Telekomunikasi



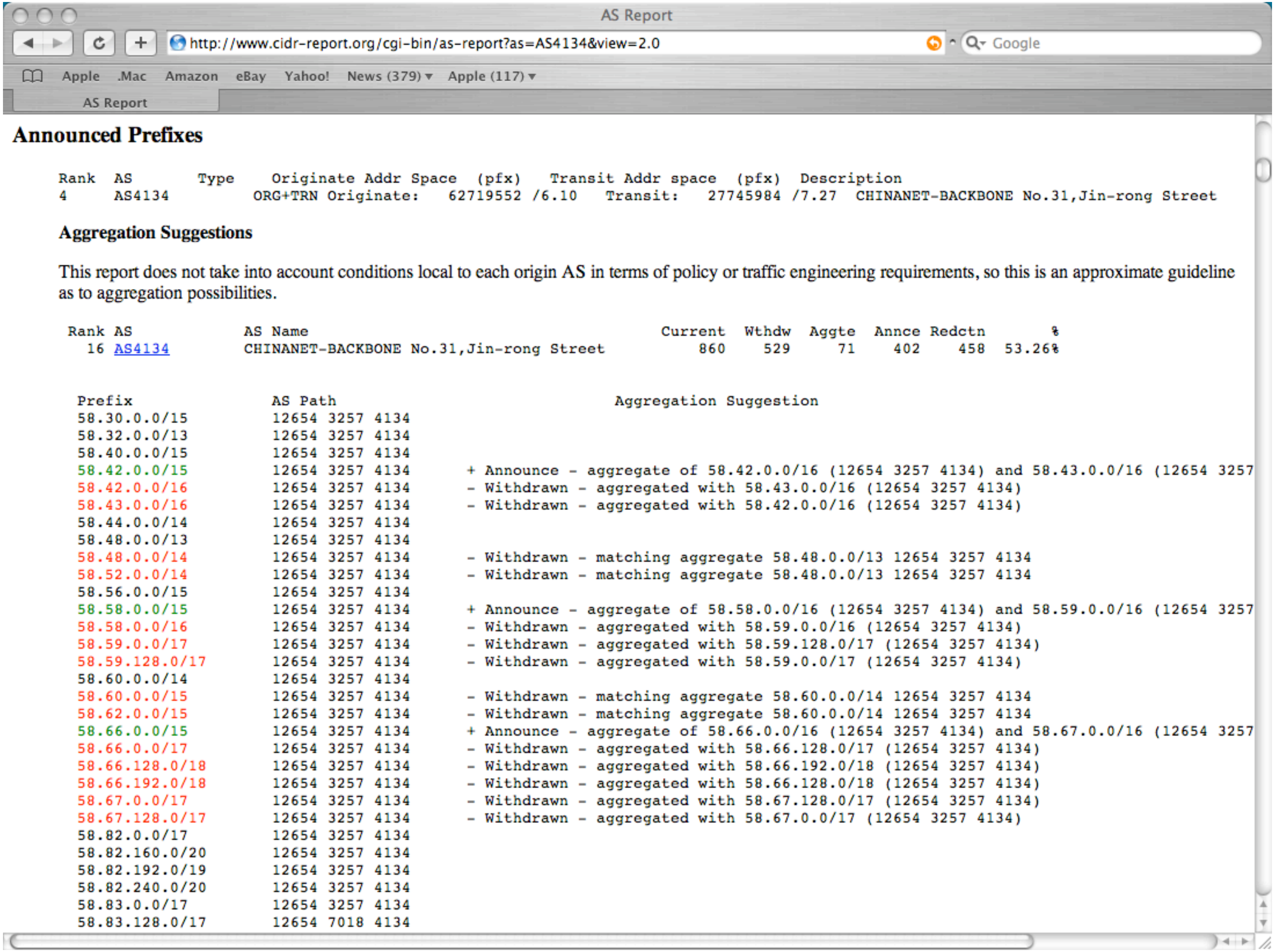
More Specifics

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
1504	1522	AS4755	VSNL-AS Videsh Sanchar Nigam Ltd. Autonomous System
1264	1360	AS2386	INS-AS - AT&T Data Communications Services
1215	1490	AS7018	ATT-INTERNET4 - AT&T WorldNet Services
1204	1210	AS11492	CABLEONE - CABLE ONE
1188	1386	AS4323	TWTC - Time Warner Telecom, Inc.
1144	1150	AS8151	Uninet S.A. de C.V.
1113	1113	AS9583	SIFY-AS-IN Sify Limited
1095	1095	AS6478	ATT-INTERNET3 - AT&T WorldNet Services
1075	1096	AS9498	BBIL-AP BHARTI BT INTERNET LTD.
1030	1040	AS18566	COVAD - Covad Communications Co.
994	1003	AS7011	FRONTIER-AND-CITIZENS - Frontier Communications of America, Inc.
960	960	AS17488	HATHWAY-NET-AP Hathway IP Over Cable Internet
944	969	AS6197	BATI-ATL - BellSouth Network Solutions, Inc
836	862	AS20115	CHARTER-NET-HKY-NC - Charter Communications
822	868	AS19262	VZGNI-TRANSIT - Verizon Internet Services Inc.
816	816	AS23577	ATM-MPLS-AS-KR Korea Telecom
802	836	AS22773	CCINET-2 - Cox Communications Inc.
788	833	AS4766	KIXS-AS-KR Korea Telecom
738	966	AS174	COGENT Cogent/PSI
645	662	AS5668	AS-5668 - CenturyTel Internet Holdings, Inc.

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



Announced Prefixes

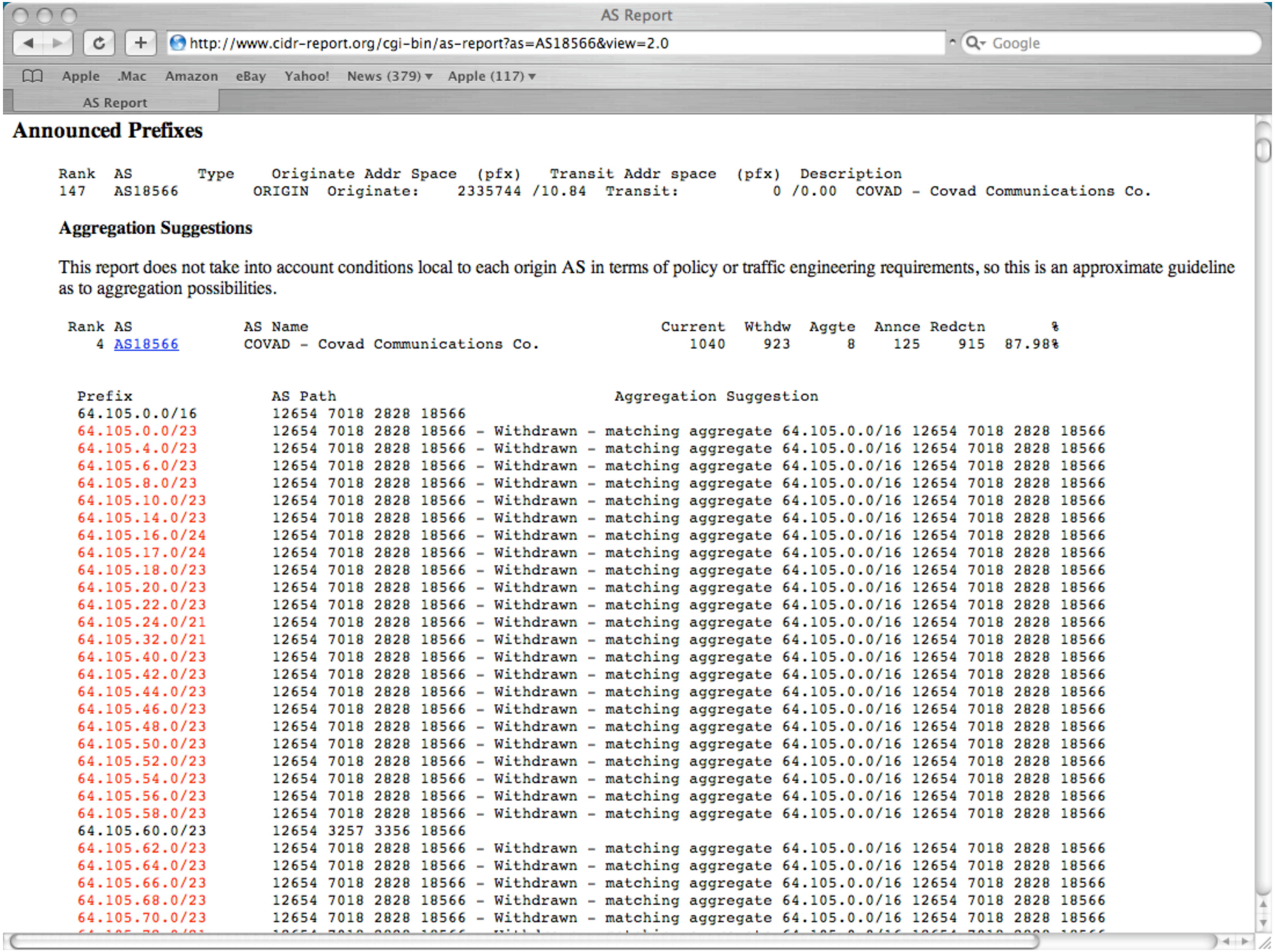
```
Rank  AS       Type    Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
4     AS4134         ORG+TRN Originate:   62719552 /6.10   Transit:   27745984 /7.27  CHINANET-BACKBONE No.31,Jin-rong Street
```

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      %
   16 AS4134       CHINANET-BACKBONE No.31,Jin-rong Street      860    529     71    402    458  53.26%


  Prefix             AS Path                            Aggregation Suggestion
  58.30.0.0/15       12654 3257 4134
  58.32.0.0/13       12654 3257 4134
  58.40.0.0/15       12654 3257 4134
  58.42.0.0/15       12654 3257 4134     + Announce - aggregate of 58.42.0.0/16 (12654 3257 4134) and 58.43.0.0/16 (12654 3257
  58.42.0.0/16       12654 3257 4134     - Withdrawn - aggregated with 58.43.0.0/16 (12654 3257 4134)
  58.43.0.0/16       12654 3257 4134     - Withdrawn - aggregated with 58.42.0.0/16 (12654 3257 4134)
  58.44.0.0/14       12654 3257 4134
  58.48.0.0/13       12654 3257 4134
  58.48.0.0/14       12654 3257 4134     - Withdrawn - matching aggregate 58.48.0.0/13 12654 3257 4134
  58.52.0.0/14       12654 3257 4134     - Withdrawn - matching aggregate 58.48.0.0/13 12654 3257 4134
  58.56.0.0/15       12654 3257 4134
  58.58.0.0/15       12654 3257 4134     + Announce - aggregate of 58.58.0.0/16 (12654 3257 4134) and 58.59.0.0/16 (12654 3257
  58.58.0.0/16       12654 3257 4134     - Withdrawn - aggregated with 58.59.0.0/16 (12654 3257 4134)
  58.59.0.0/17       12654 3257 4134     - Withdrawn - aggregated with 58.59.128.0/17 (12654 3257 4134)
  58.59.128.0/17     12654 3257 4134     - Withdrawn - aggregated with 58.59.0.0/17 (12654 3257 4134)
  58.60.0.0/14       12654 3257 4134
  58.60.0.0/15       12654 3257 4134     - Withdrawn - matching aggregate 58.60.0.0/14 12654 3257 4134
  58.62.0.0/15       12654 3257 4134     - Withdrawn - matching aggregate 58.60.0.0/14 12654 3257 4134
  58.66.0.0/15       12654 3257 4134     + Announce - aggregate of 58.66.0.0/16 (12654 3257 4134) and 58.67.0.0/16 (12654 3257
  58.66.0.0/17       12654 3257 4134     - Withdrawn - aggregated with 58.66.128.0/17 (12654 3257 4134)
  58.66.128.0/18     12654 3257 4134     - Withdrawn - aggregated with 58.66.192.0/18 (12654 3257 4134)
  58.66.192.0/18     12654 3257 4134     - Withdrawn - aggregated with 58.66.128.0/18 (12654 3257 4134)
  58.67.0.0/17       12654 3257 4134     - Withdrawn - aggregated with 58.67.128.0/17 (12654 3257 4134)
  58.67.128.0/17     12654 3257 4134     - Withdrawn - aggregated with 58.67.0.0/17 (12654 3257 4134)
  58.82.0.0/17       12654 3257 4134
  58.82.160.0/20     12654 3257 4134
  58.82.192.0/19     12654 3257 4134
  58.82.240.0/20     12654 3257 4134
  58.83.0.0/17       12654 3257 4134
  58.83.128.0/17     12654 7018 4134
```


AS Report
http://www.cidr-report.org/cgi-bin/as-report?as=AS18566&view=2.0
Google
Apple .Mac Amazon eBay Yahoo! News (379) ▾ Apple (117) ▾
AS Report

Announced Prefixes

```
Rank  AS        Type    Originate Addr Space  (pfx)  Transit Addr space  (pfx)  Description
147   AS18566         ORIGIN  Originate:    2335744 /10.84  Transit:          0 /0.00  COVAD - Covad Communications Co.
```

Aggregation Suggestions

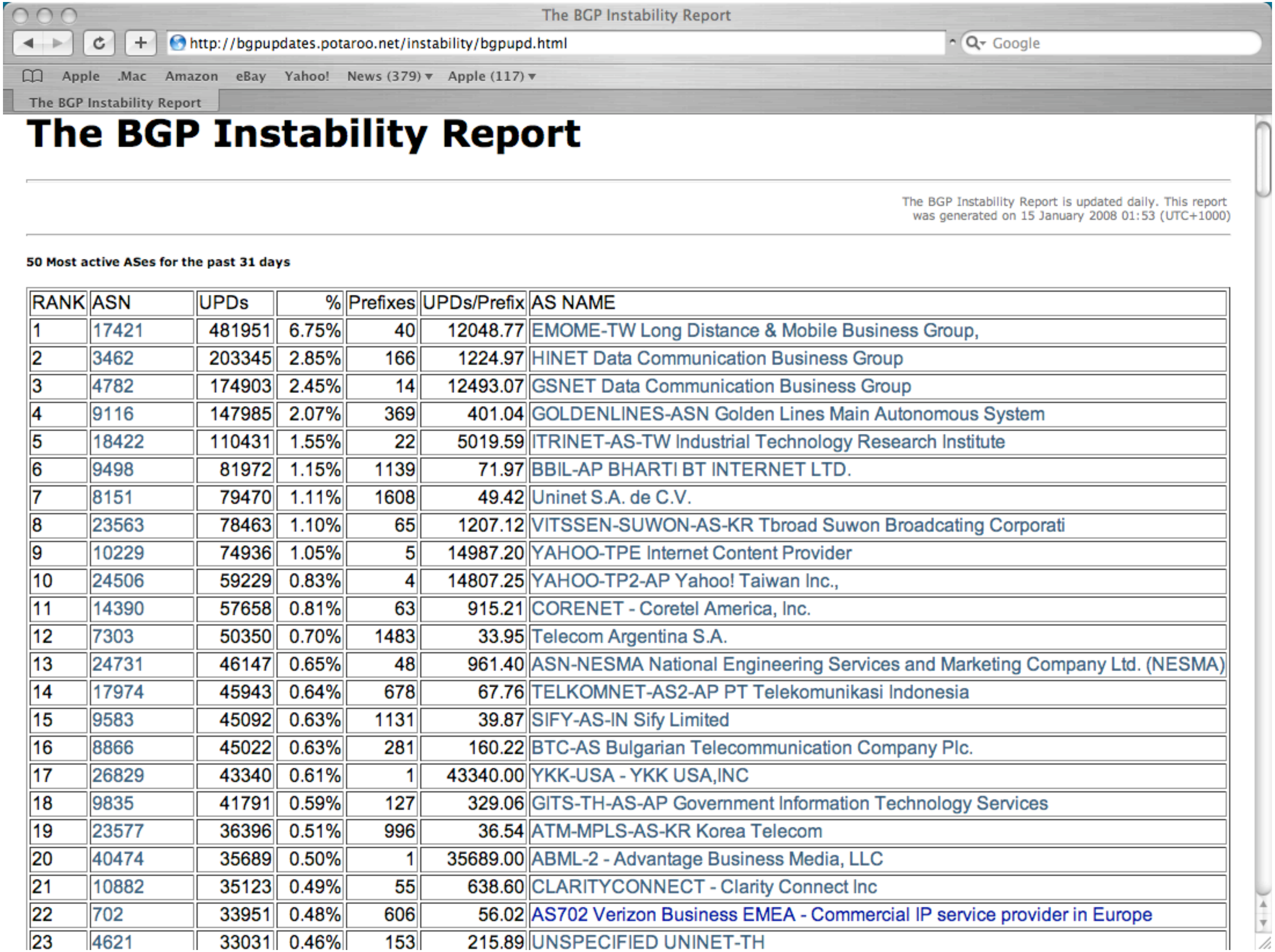
This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

```
 Rank AS           AS Name                                       Current  Wthdw  Aggte  Annce Redctn      %
    4 AS18566      COVAD - Covad Communications Co.                 1040    923      8    125    915  87.98%


  Prefix              AS Path                             Aggregation Suggestion
  64.105.0.0/16       12654 7018 2828 18566
  64.105.0.0/23       12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.4.0/23       12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.6.0/23       12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.8.0/23       12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.10.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.14.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.16.0/24      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.17.0/24      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.18.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.20.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.22.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.24.0/21      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.32.0/21      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.40.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.42.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.44.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.46.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.48.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.50.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.52.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.54.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.56.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.58.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.60.0/23      12654 3257 3356 18566
  64.105.62.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.64.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.66.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.68.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
  64.105.70.0/23      12654 7018 2828 18566 - Withdrawn - matching aggregate 64.105.0.0/16 12654 7018 2828 18566
```

Importance of Aggregation

- Size of routing table
  - Memory is no longer a problem
  - Routers can be specified to carry 1 million prefixes
- Convergence of the Routing System
  - This is a problem
  - Bigger table takes longer for CPU to process
  - BGP updates take longer to deal with
  - BGP Instability Report tracks routing system update activity
  - **http://bgpupdates.potaroo.net/instability/bgpupd.html**



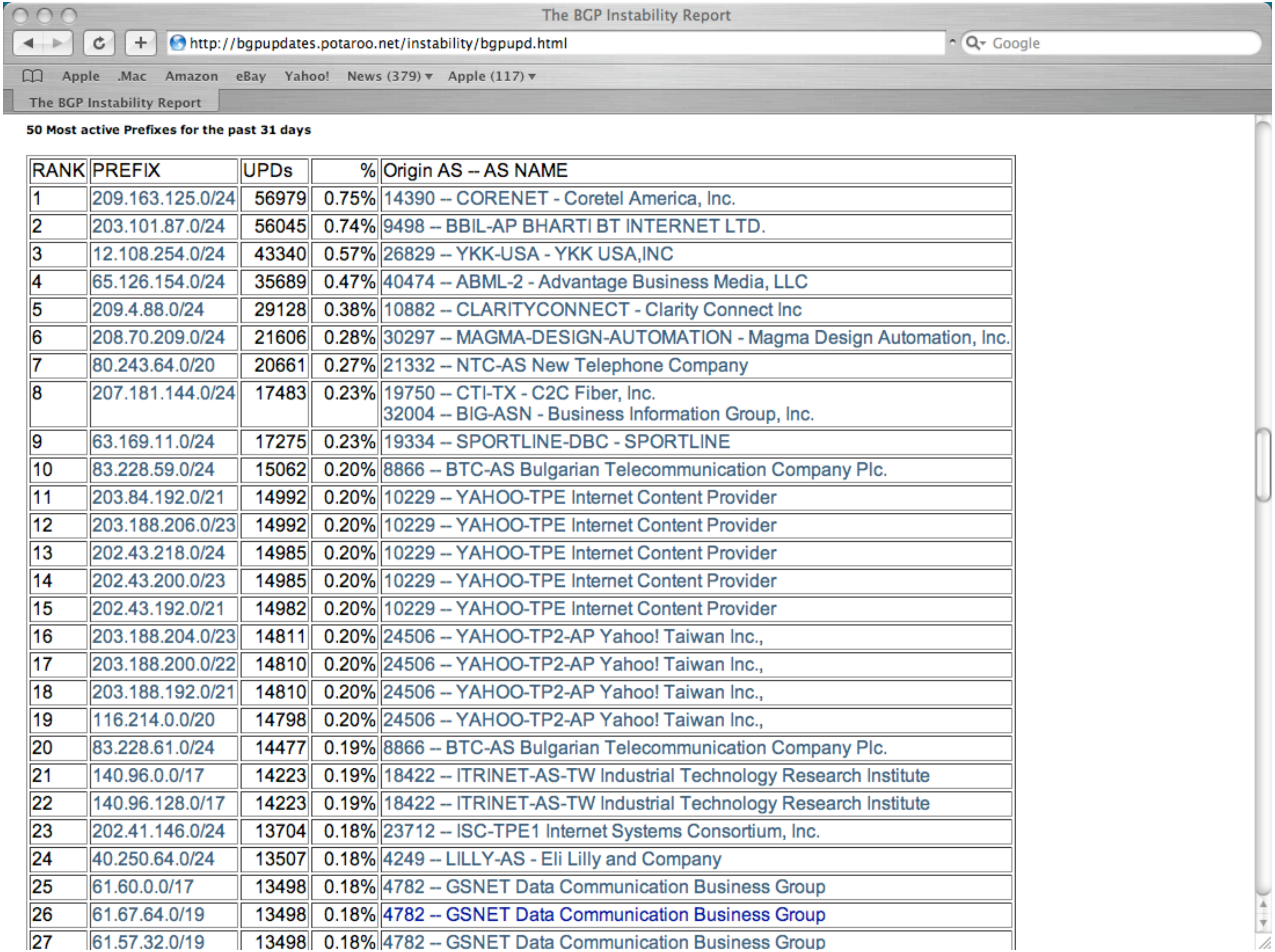
The BGP Instability Report
http://bgpupdates.potaroo.net/instability/bgpupd.html
Google
Apple
.Mac
Amazon
eBay
Yahoo!
News (379)
Apple (117)
The BGP Instability Report

The BGP Instability Report

The BGP Instability Report is updated daily. This report was generated on 15 January 2008 01:53 (UTC+1000)

50 Most active ASes for the past 31 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	17421	481951	6.75%	40	12048.77	EMOME-TW Long Distance & Mobile Business Group,
2	3462	203345	2.85%	166	1224.97	HINET Data Communication Business Group
3	4782	174903	2.45%	14	12493.07	GSNET Data Communication Business Group
4	9116	147985	2.07%	369	401.04	GOLDENLINES-ASN Golden Lines Main Autonomous System
5	18422	110431	1.55%	22	5019.59	ITRINET-AS-TW Industrial Technology Research Institute
6	9498	81972	1.15%	1139	71.97	BBIL-AP BHARTI BT INTERNET LTD.
7	8151	79470	1.11%	1608	49.42	Uninet S.A. de C.V.
8	23563	78463	1.10%	65	1207.12	VITSSEN-SUWON-AS-KR Tbroad Suwon Broadcating Corporati
9	10229	74936	1.05%	5	14987.20	YAHOO-TPE Internet Content Provider
10	24506	59229	0.83%	4	14807.25	YAHOO-TP2-AP Yahoo! Taiwan Inc.,
11	14390	57658	0.81%	63	915.21	CORENET - Coretel America, Inc.
12	7303	50350	0.70%	1483	33.95	Telecom Argentina S.A.
13	24731	46147	0.65%	48	961.40	ASN-NESMA National Engineering Services and Marketing Company Ltd. (NESMA)
14	17974	45943	0.64%	678	67.76	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
15	9583	45092	0.63%	1131	39.87	SIFY-AS-IN Sify Limited
16	8866	45022	0.63%	281	160.22	BTC-AS Bulgarian Telecommunication Company Plc.
17	26829	43340	0.61%	1	43340.00	YKK-USA - YKK USA,INC
18	9835	41791	0.59%	127	329.06	GITS-TH-AS-AP Government Information Technology Services
19	23577	36396	0.51%	996	36.54	ATM-MPLS-AS-KR Korea Telecom
20	40474	35689	0.50%	1	35689.00	ABML-2 - Advantage Business Media, LLC
21	10882	35123	0.49%	55	638.60	CLARITYCONNECT - Clarity Connect Inc
22	702	33951	0.48%	606	56.02	AS702 Verizon Business EMEA - Commercial IP service provider in Europe
23	4621	33031	0.46%	153	215.89	UNSPECIFIED UNINET-TH

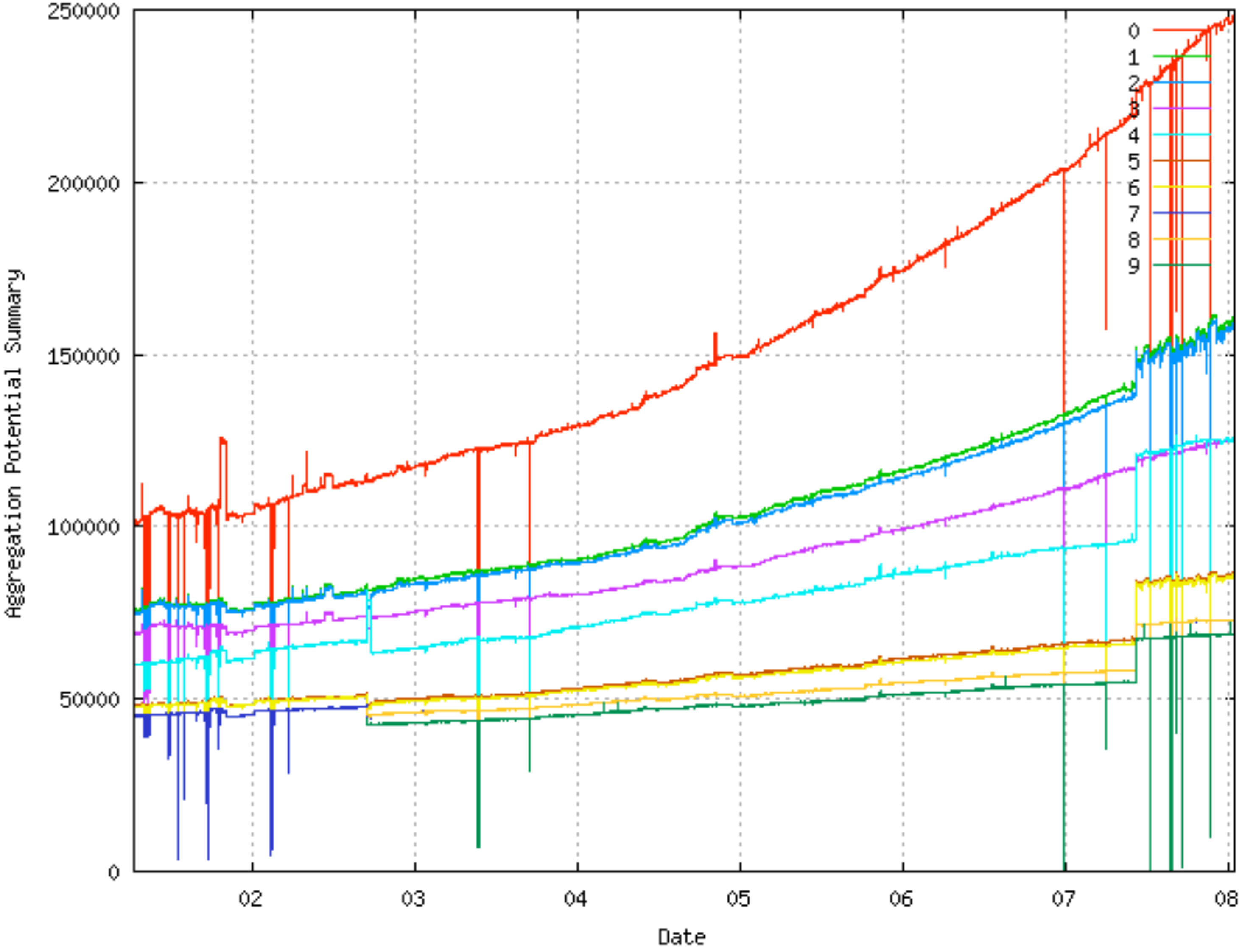


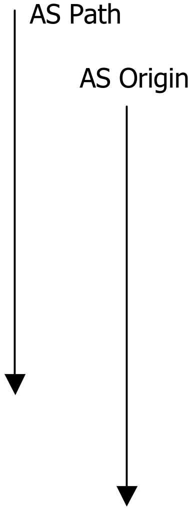
The BGP Instability Report
http://bgpupdates.potaroo.net/instability/bgpupd.html
Google
Apple
.Mac
Amazon
eBay
Yahoo!
News (379)
Apple (117)
The BGP Instability Report

50 Most active Prefixes for the past 31 days

RANK	PREFIX	UPDs	%	Origin AS -- AS NAME
1	209.163.125.0/24	56979	0.75%	14390 -- CORENET - Coretel America, Inc.
2	203.101.87.0/24	56045	0.74%	9498 -- BBIL-AP BHARTI BT INTERNET LTD.
3	12.108.254.0/24	43340	0.57%	26829 -- YKK-USA - YKK USA,INC
4	65.126.154.0/24	35689	0.47%	40474 -- ABML-2 - Advantage Business Media, LLC
5	209.4.88.0/24	29128	0.38%	10882 -- CLARITYCONNECT - Clarity Connect Inc
6	208.70.209.0/24	21606	0.28%	30297 -- MAGMA-DESIGN-AUTOMATION - Magma Design Automation, Inc.
7	80.243.64.0/20	20661	0.27%	21332 -- NTC-AS New Telephone Company
8	207.181.144.0/24	17483	0.23%	19750 -- CTI-TX - C2C Fiber, Inc. 32004 -- BIG-ASN - Business Information Group, Inc.
9	63.169.11.0/24	17275	0.23%	19334 -- SPORTLINE-DBC - SPORTLINE
10	83.228.59.0/24	15062	0.20%	8866 -- BTC-AS Bulgarian Telecommunication Company Plc.
11	203.84.192.0/21	14992	0.20%	10229 -- YAHOO-TPE Internet Content Provider
12	203.188.206.0/23	14992	0.20%	10229 -- YAHOO-TPE Internet Content Provider
13	202.43.218.0/24	14985	0.20%	10229 -- YAHOO-TPE Internet Content Provider
14	202.43.200.0/23	14985	0.20%	10229 -- YAHOO-TPE Internet Content Provider
15	202.43.192.0/21	14982	0.20%	10229 -- YAHOO-TPE Internet Content Provider
16	203.188.204.0/23	14811	0.20%	24506 -- YAHOO-TP2-AP Yahoo! Taiwan Inc.,
17	203.188.200.0/22	14810	0.20%	24506 -- YAHOO-TP2-AP Yahoo! Taiwan Inc.,
18	203.188.192.0/21	14810	0.20%	24506 -- YAHOO-TP2-AP Yahoo! Taiwan Inc.,
19	116.214.0.0/20	14798	0.20%	24506 -- YAHOO-TP2-AP Yahoo! Taiwan Inc.,
20	83.228.61.0/24	14477	0.19%	8866 -- BTC-AS Bulgarian Telecommunication Company Plc.
21	140.96.0.0/17	14223	0.19%	18422 -- ITRINET-AS-TW Industrial Technology Research Institute
22	140.96.128.0/17	14223	0.19%	18422 -- ITRINET-AS-TW Industrial Technology Research Institute
23	202.41.146.0/24	13704	0.18%	23712 -- ISC-TPE1 Internet Systems Consortium, Inc.
24	40.250.64.0/24	13507	0.18%	4249 -- LILLY-AS - Eli Lilly and Company
25	61.60.0.0/17	13498	0.18%	4782 -- GSNET Data Communication Business Group
26	61.67.64.0/19	13498	0.18%	4782 -- GSNET Data Communication Business Group
27	61.57.32.0/19	13498	0.18%	4782 -- GSNET Data Communication Business Group

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





Aggregation Summary

- Aggregation on the Internet could be MUCH better
  - 35% saving on Internet routing table size is quite feasible
  - Tools are available
    - Commands on the routers are not hard
    - CIDR-Report webpage

Receiving Prefixes

Receiving Prefixes

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

Receiving Prefixes: From Customers

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

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

  The tool: whois

Receiving Prefixes: From Customers

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

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

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

Receiving Prefixes: From Customers

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

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

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

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

Receiving Prefixes: From Peers

- A peer is an ISP with whom you agree to exchange prefixes you originate into the Internet routing table
  - Prefixes you accept from a peer are only those they have indicated they will announce
  - Prefixes you announce to your peer are only those you have indicated you will announce

Receiving Prefixes:
From Peers

- Agreeing what each will announce to the other:

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

  OR

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

  www.isc.org/sw/IRRToolSet/

Receiving Prefixes: From Upstream/Transit Provider

- Upstream/Transit Provider is an ISP who you pay to give you transit to the WHOLE Internet
- Receiving prefixes from them is not desirable unless really necessary

  special circumstances – see later
- Ask upstream/transit provider to either:

  originate a default-route

  OR

  announce one prefix you can use as default

Receiving Prefixes: From Upstream/Transit Provider

- If necessary to receive prefixes from any provider, care is required
  - don’t accept RFC1918 *etc* prefixes
    - ftp://ftp.rfc-editor.org/in-notes/rfc3330.txt
  - don’t accept your own prefixes
  - don’t accept default (unless you need it)
  - don’t accept prefixes longer than /24
- Check Team Cymru’s bogon pages
  - http://www.cymru.com/Bogons
  - http://www.cymru.com/BGP/bogon-rs.html – bogon route server

Receiving Prefixes

- Paying attention to prefixes received from customers, peers and transit providers assists with:
  - The integrity of the local network
  - The integrity of the Internet
- Responsibility of all ISPs to be good Internet citizens

Preparing the network

Before we begin…

Preparing the Network

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

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

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

  Ask an ISP who is a registry member, or

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

Preparing the Network
Initial Assumptions

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

Preparing the Network
First Step: IGP

- Decide on an IGP: OSPF or ISIS ☺
- Assign loopback interfaces and /32 address to each router which will run the IGP

  Loopback is used for OSPF and BGP router id anchor

  Used for iBGP and route origination
- Deploy IGP (e.g. OSPF)

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

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

  Smallest distance wins

Preparing the Network
IGP (cont)

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

    Router loopbacks go in IGP

    WAN point to point links go in IGP

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

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

Preparing the Network
IGP (cont)

- Routes which don't go into the IGP include:

  Dynamic assignment pools (DSL/Cable/Dial)

  Customer point to point link addressing

  (using next-hop-self in iBGP ensures that these do NOT need to be in IGP)

  Static/Hosting LANs

  Customer assigned address space

  Anything else not listed in the previous slide

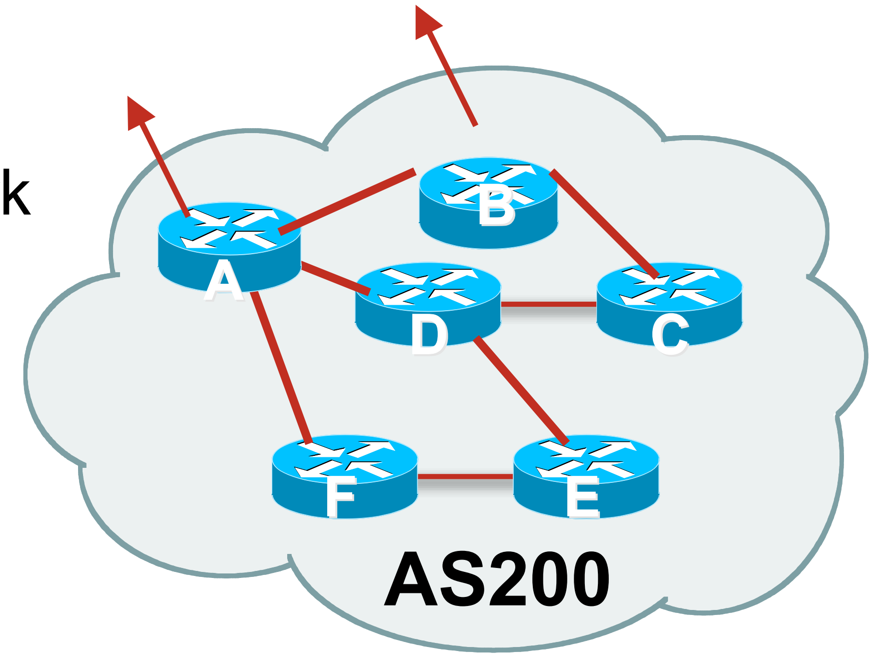
Preparing the Network
Second Step: iBGP

- Second step is to configure the local network to use iBGP
- iBGP can run on

  all routers, or

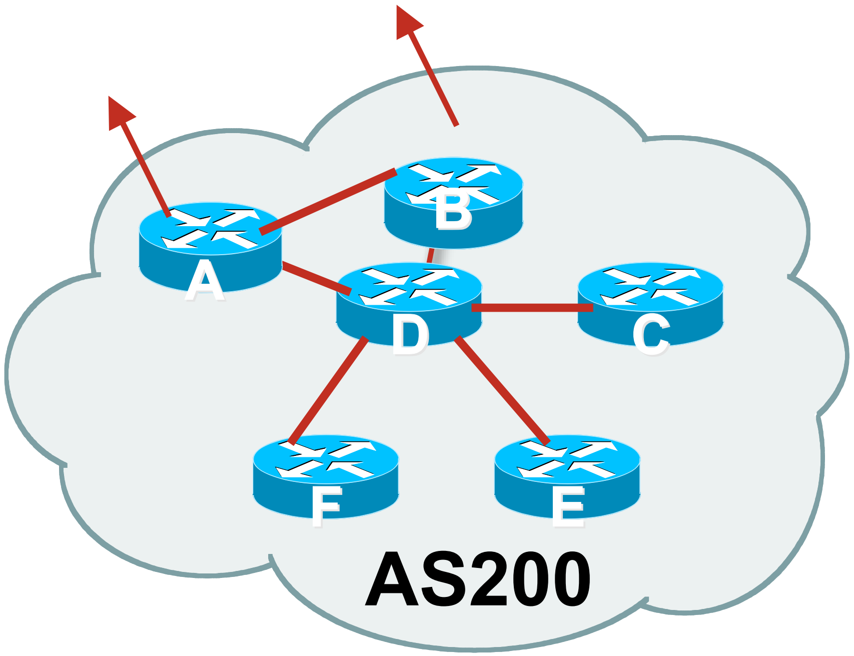
  a subset of routers, or

  just on the upstream edge
- *iBGP must run on all routers which are in the transit path between external connections*



Preparing the Network
Second Step: iBGP (Transit Path)

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



Preparing the Network
Layers

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

Preparing the Network Aggregation Layer

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

Preparing the Network Distribution Layer

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

Preparing the Network
Core Layer

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

Preparing the Network
iBGP Implementation

Decide on:

- Best iBGP policy

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

- iBGP scaling technique

  Community policy?

  Route-reflectors?

  Techniques such as peer groups and peer templates?

Preparing the Network
iBGP Implementation

- Then deploy iBGP:

  Step 1: Introduce iBGP mesh on chosen routers

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

  Step 2: Install “customer” prefixes into iBGP

  Check! Does the network still work?

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

  Check! Does the network still work?

  Step 4: Deployment of eBGP follows

Preparing the Network
iBGP Implementation

Install "customer" prefixes into iBGP?

- Customer assigned address space

  Network statement/static route combination

  Use unique community to identify customer assignments

- Customer facing point-to-point links

  Redistribute connected through filters which only permit point-to-point link addresses to enter iBGP

  Use a unique community to identify point-to-point link addresses (these are only required for your monitoring system)

- Dynamic assignment pools & local LANs

  Simple network statement will do this

  Use unique community to identify these networks

Preparing the Network
iBGP Implementation

Carefully remove static routes?

- Work on one router at a time:

  Check that static route for a particular destination is also learned by the iBGP

  If so, remove it

  If not, establish why and fix the problem

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

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

Preparing the Network
Completion

- Previous steps are NOT flag day steps

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

  Step One on Week One

  Step Two on Week Two

  Step Three on Week Three

  And so on

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

Preparing the Network
Example Two

- The network is not running any BGP at the moment

  single statically routed connection to upstream ISP

- The network is running an IGP though

  All internal routing information is in the IGP

  By IGP, OSPF or ISIS is assumed

Preparing the Network
IGP

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

Preparing the Network
iBGP

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

Preparing the Network
iBGP Implementation

- Then deploy iBGP:

  Step 1: Introduce iBGP mesh on chosen routers

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

  Step 2: Install “customer” prefixes into iBGP

  Check! Does the network still work?

  Step 3: Reduce BGP distance to be less than the IGP

  (so that iBGP routes take priority)

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

  Check! Does the network still work?

  Step 5: Restore BGP distance to less than IGP

  Step 6: Deployment of eBGP follows

Preparing the Network
iBGP implementation

Install "customer" prefixes into iBGP?

- Customer assigned address space

  Network statement/static route combination

  Use unique community to identify customer assignments

- Customer facing point-to-point links

  Redistribute connected through filters which only permit point-to-point link addresses to enter iBGP

  Use a unique community to identify point-to-point link addresses (these are only required for your monitoring system)

- Dynamic assignment pools & local LANs

  Simple network statement will do this

  Use unique community to identify these networks

Preparing the Network
iBGP implementation

Carefully remove "customer" routes from IGP?

- Work on one router at a time:

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

  If so, remove it from the IGP

  If not, establish why and fix the problem

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

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

Preparing the Network
Completion

- Previous steps are NOT flag day steps

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

  Step One on Week One

  Step Two on Week Two

  Step Three on Week Three

  And so on

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

Preparing the Network
Configuration Summary

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

Configuration Tips

Of passwords, tricks and templates

iBGP and IGPs Reminder!

- Make sure loopback is configured on router

  iBGP between loopbacks, NOT real interfaces

- Make sure IGP carries loopback /32 address
- Consider the DMZ nets:

  Use unnumbered interfaces?

  Use next-hop-self on iBGP neighbours

  Or carry the DMZ /30s in the iBGP

  Basically keep the DMZ nets out of the IGP!

iBGP: Next-hop-self

- BGP speaker announces external network to iBGP peers using router’s local address (loopback) as next-hop
- Used by many ISPs on edge routers
    - Preferable to carrying DMZ /30 addresses in the IGP
    - Reduces size of IGP to just core infrastructure
    - Alternative to using unnumbered interfaces
    - Helps scale network
    - Many ISPs consider this “best practice”

Limiting AS Path Length

- Some BGP implementations have problems with long AS_PATHS
  - Memory corruption
  - Memory fragmentation
- Even using AS_PATH prepends, it is not normal to see more than 20 ASes in a typical AS_PATH in the Internet today
  - The Internet is around 5 ASes deep on average
  - Largest AS_PATH is usually 16-20 ASNs

Limiting AS Path Length

- Some announcements have ridiculous lengths of AS-paths:

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

  This example is an error in one IPv6 implementation

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

  This example shows 20 prepends (for no obvious reason)

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

BGP TTL "hack"

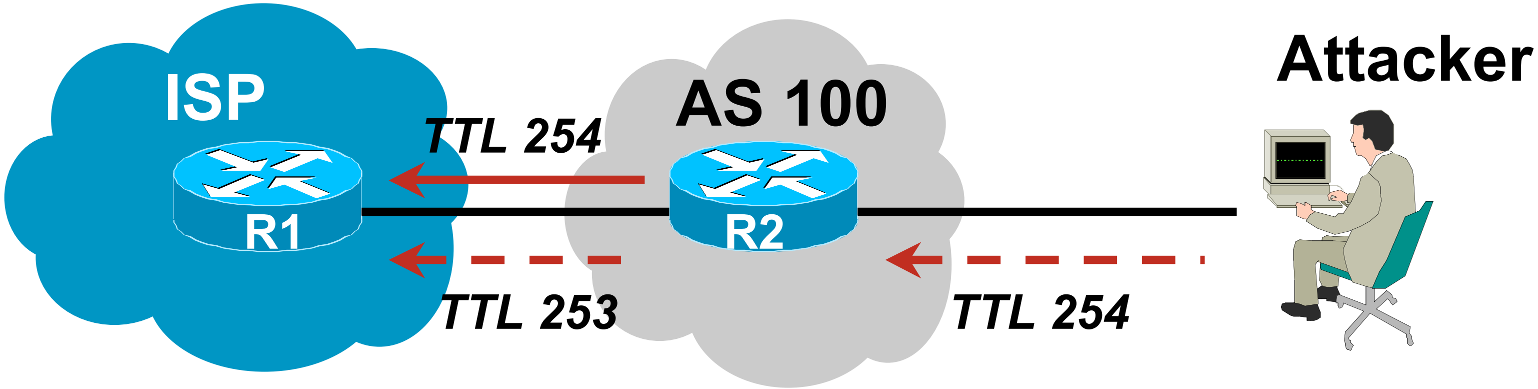
- Implement RFC5082 on BGP peerings

  (Generalised TTL Security Mechanism)

  Neighbour sets TTL to 255

  Local router expects TTL of incoming BGP packets to be 254

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



BGP TTL “hack”

- TTL Hack:

  Both neighbours must agree to use the feature

  TTL check is much easier to perform than MD5

  (Called BTSH – BGP TTL Security Hack)

- Provides “security” for BGP sessions

  In addition to packet filters of course

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

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

Templates

- Good practice to configure templates for everything

  Vendor defaults tend not to be optimal or even very useful for ISPs

  ISPs create their own defaults by using configuration templates

- eBGP and iBGP examples follow

  Also see Project Cymru's BGP templates

  www.cymru.com/Documents

iBGP Template Example

- iBGP between loopbacks!
- Next-hop-self

  Keep DMZ and external point-to-point out of IGP
- Always send communities in iBGP

  Otherwise accidents will happen
- Hardwire BGP to version 4

  Yes, this is being paranoid!

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>

NZNOG'08

23rd-25th January 2008

Dunedin