Deploying BGP

Philip Smith <pfs@cisco.com>

NZNOG 2006

22-24 Mar 2006

Wellington

Presentation Slides

- Will be available on

  ftp://ftp-eng.cisco.com

  /pfs/seminars/NZNOG2006-BGP-part1+2.pdf

  And on the NZNOG 2006 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 this BGP thing?

Border Gateway Protocol

- Routing Protocol used to exchange routing information between networks

  exterior gateway protocol

- Described in RFC4271

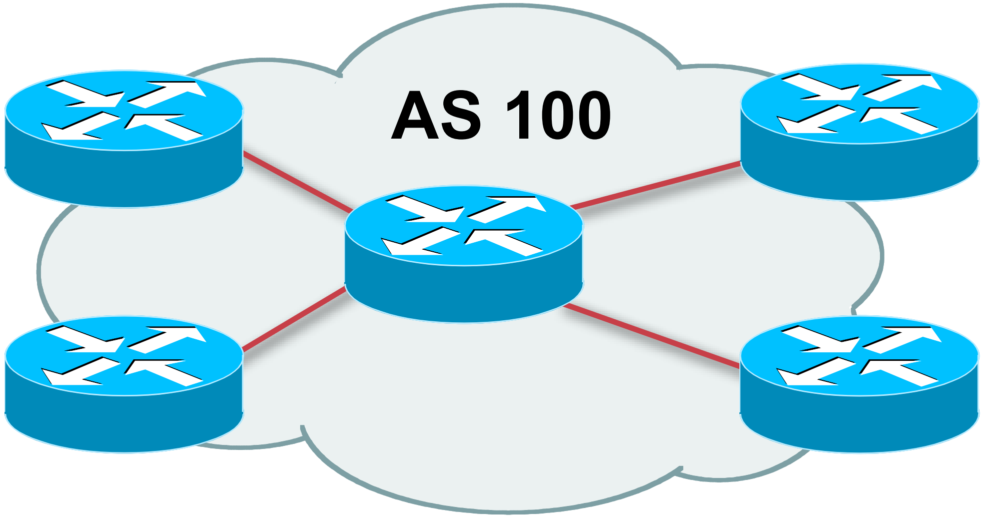
  RFC4276 gives an implementation report on BGP-4

  RFC4277 describes operational experiences using BGP-4

- The Autonomous System is BGP's fundamental operating unit

  It is used to uniquely identify networks with 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**

Autonomous System Number (ASN)

- An ASN is a 16 bit number

  1-64511 are assigned by the RIRs

  64512-65534 are for private use and should never appear on the Internet

  0 and 65535 are reserved

- 32 bit ASNs are coming soon

  www.ietf.org/internet-drafts/draft-ietf-idr-as4bytes-12.txt

  With AS 23456 reserved for the transition

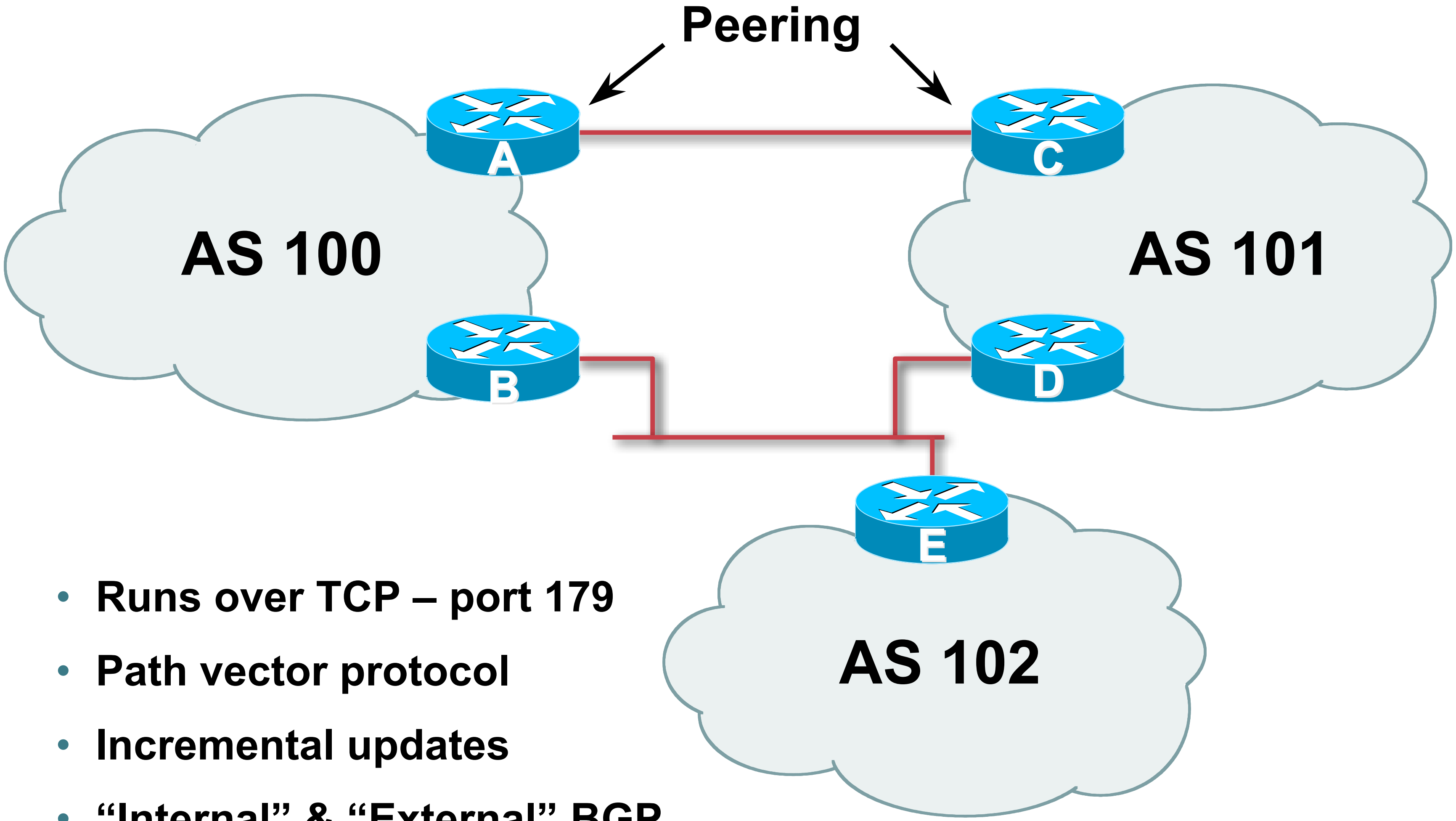
Autonomous System Number (ASN)

- ASNs are distributed by the Regional Internet Registries
- Also available from upstream ISPs who are members of one of the RIRs

  Current ASN allocations up to 39935 have been made to the RIRs

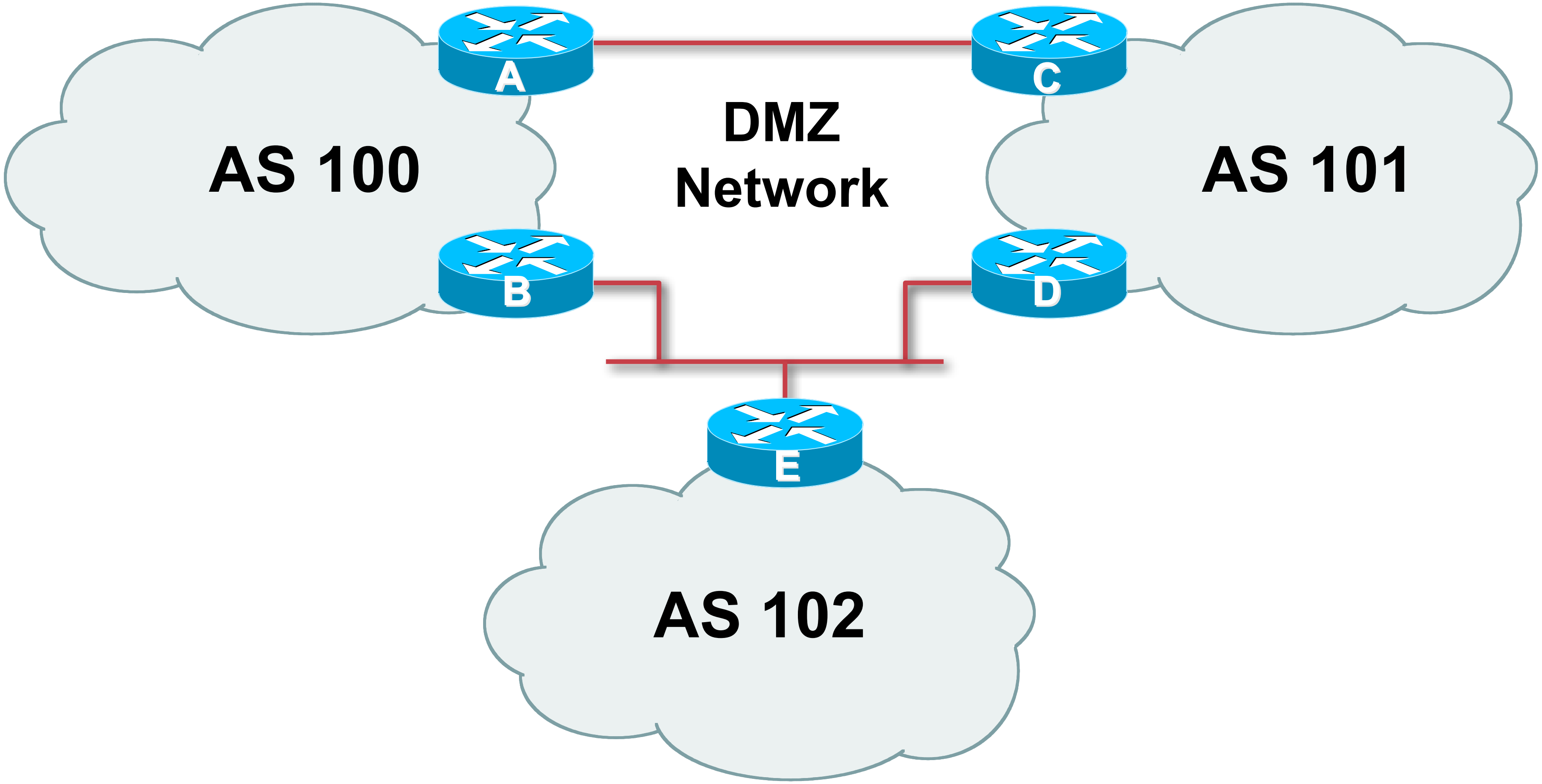
  Of these, around 21000 are visible on the Internet

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 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

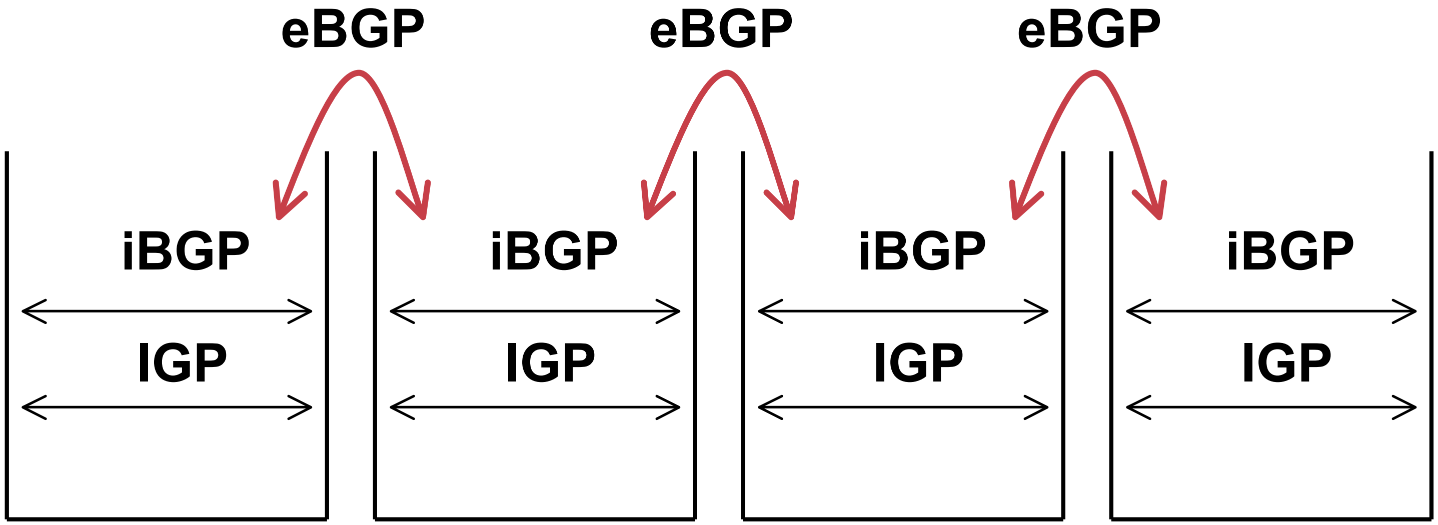
  ISP’s customer prefixes
- eBGP used to

  exchange prefixes with other ASes

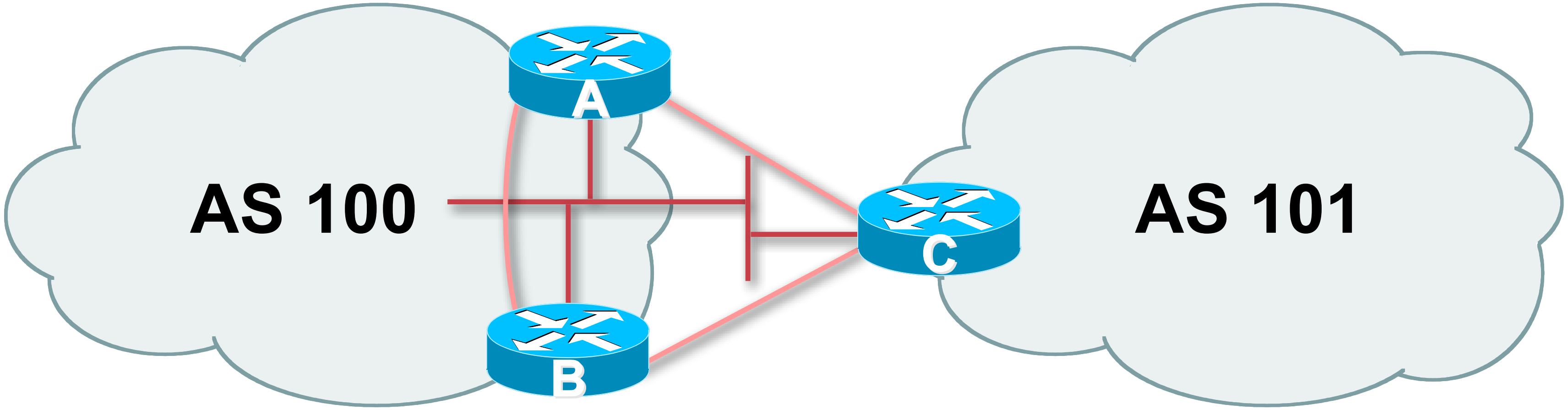
  implement routing policy

BGP/IGP model used in ISP networks

- **Model representation**



External BGP Peering (eBGP)

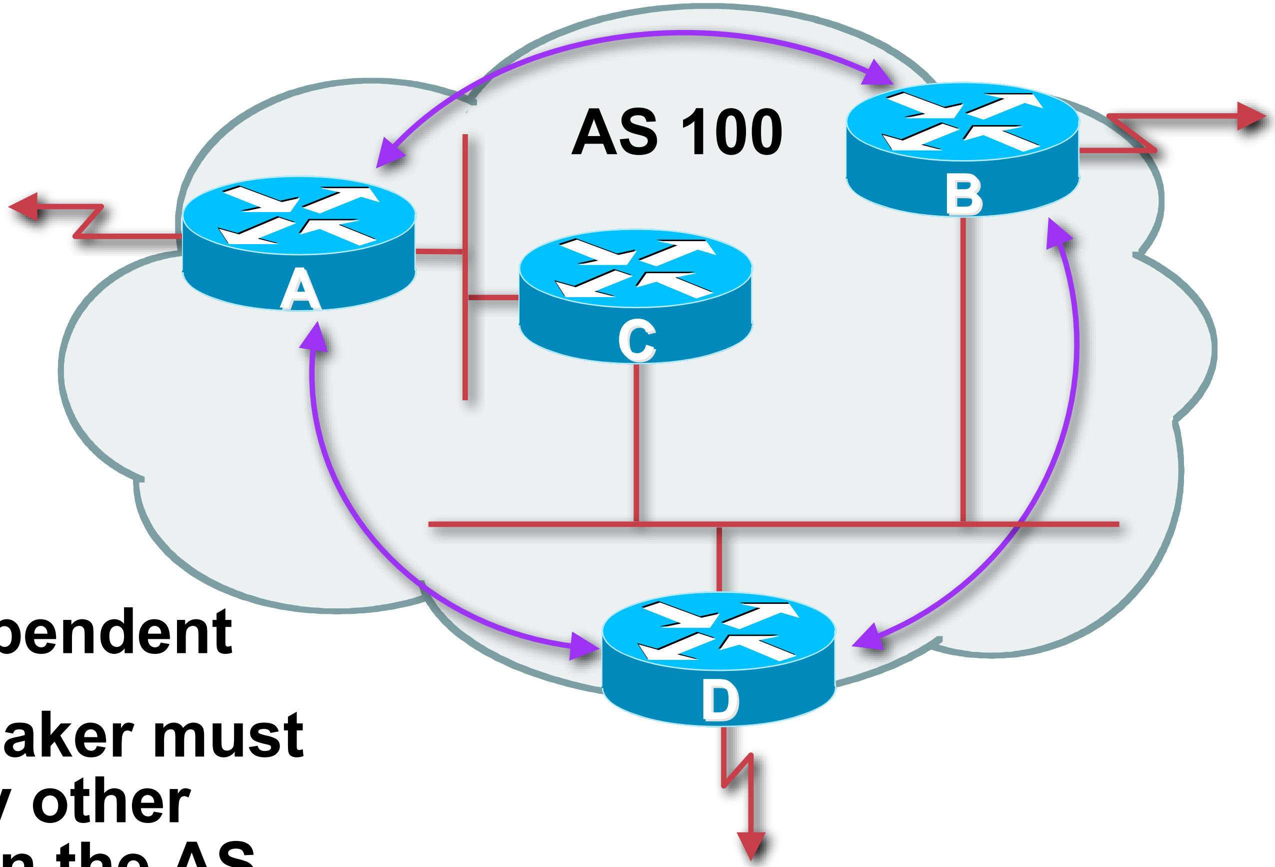


- **Between BGP speakers in different AS**
- **Should be directly connected**
- **Never run an IGP between eBGP peers**

Internal BGP (iBGP)

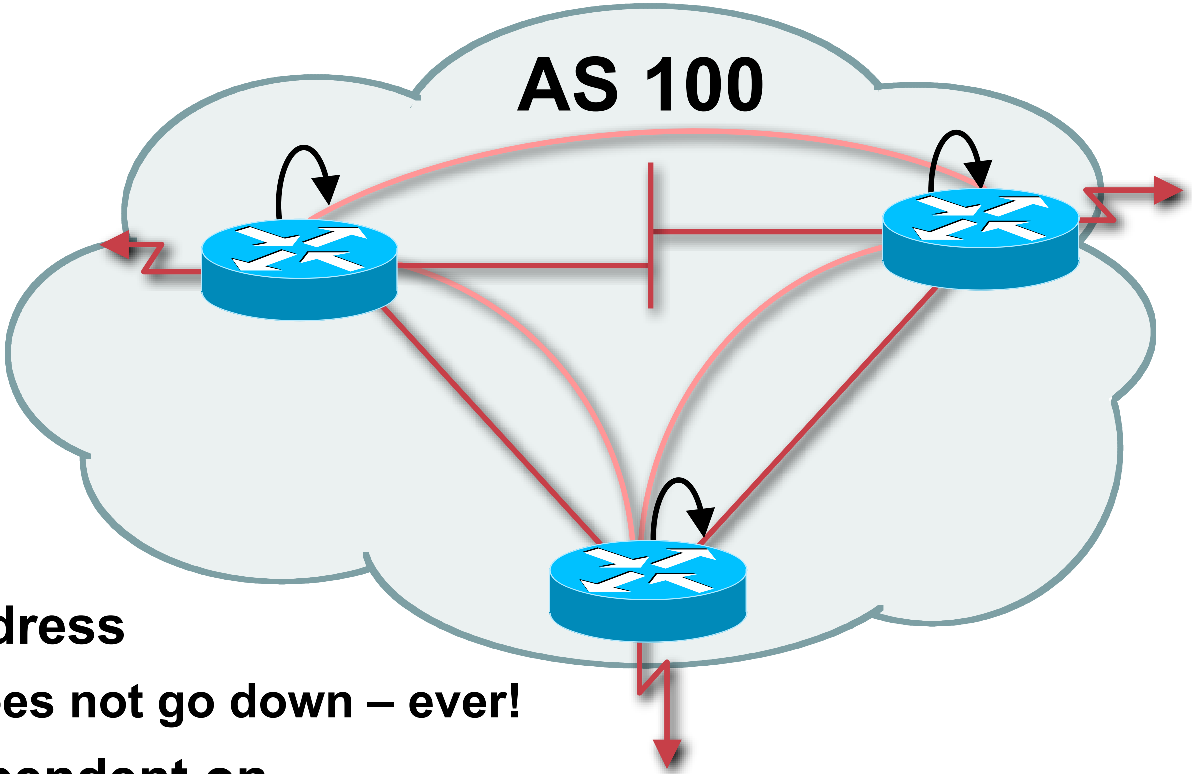
- BGP peer within the same AS
- Not required to be directly connected
  - IGP takes care of inter-BGP speaker connectivity
- iBGP speakers need to be fully meshed
  - they originate connected networks
  - they do not pass on prefixes learned from other iBGP speakers

Internal BGP Peering (iBGP)



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

Peering to loopback addresses



- **Peer with loop-back address**

  Loop-back interface does not go down – ever!

- **iBGP session is not dependent on**

  State of a single interface

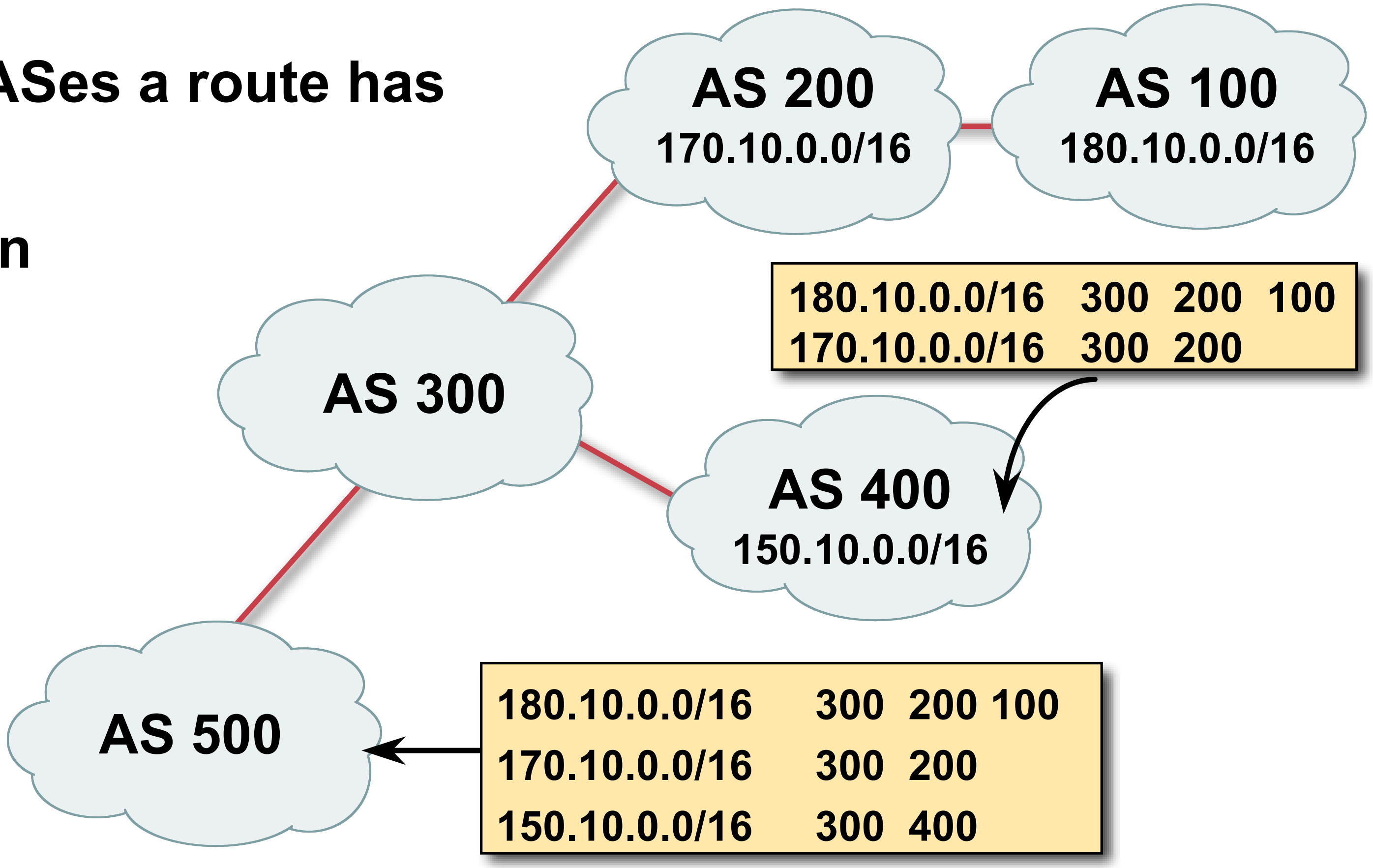
  Physical topology

BGP Attributes

Information about BGP

AS-Path

- Sequence of ASes a route has traversed
- Loop detection
- Apply policy



AS 200
170.10.0.0/16

AS 100
180.10.0.0/16

AS 300

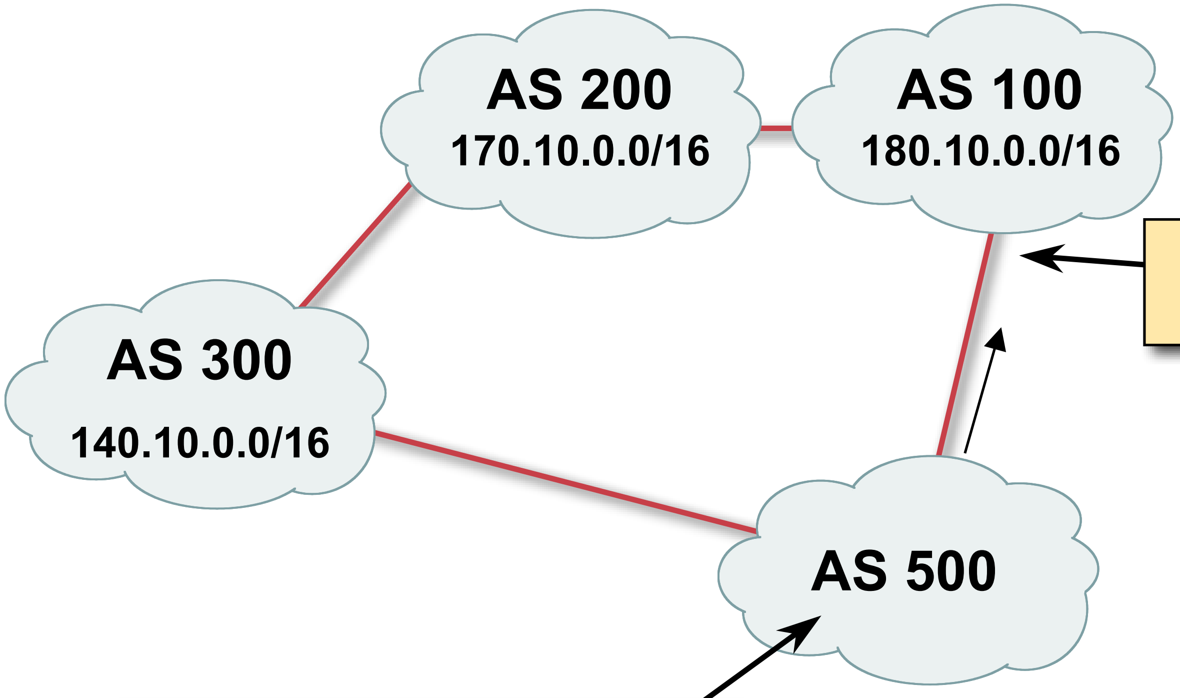
AS 400
150.10.0.0/16

AS 500

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 loop detection

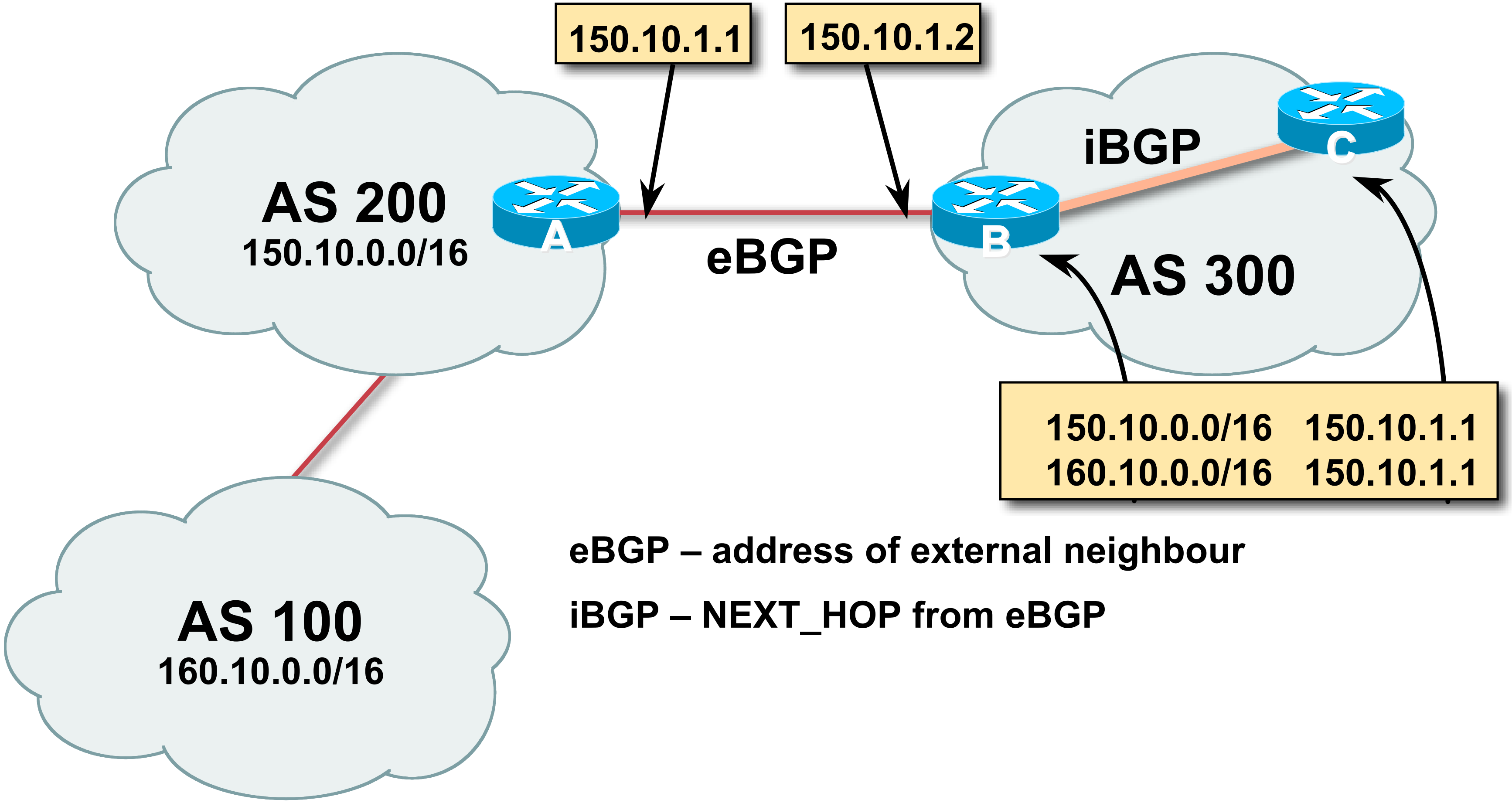


140.10.0.0/16	500 300
170.10.0.0/16	500 300 200

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 attribute – this is loop detection in action

Next Hop

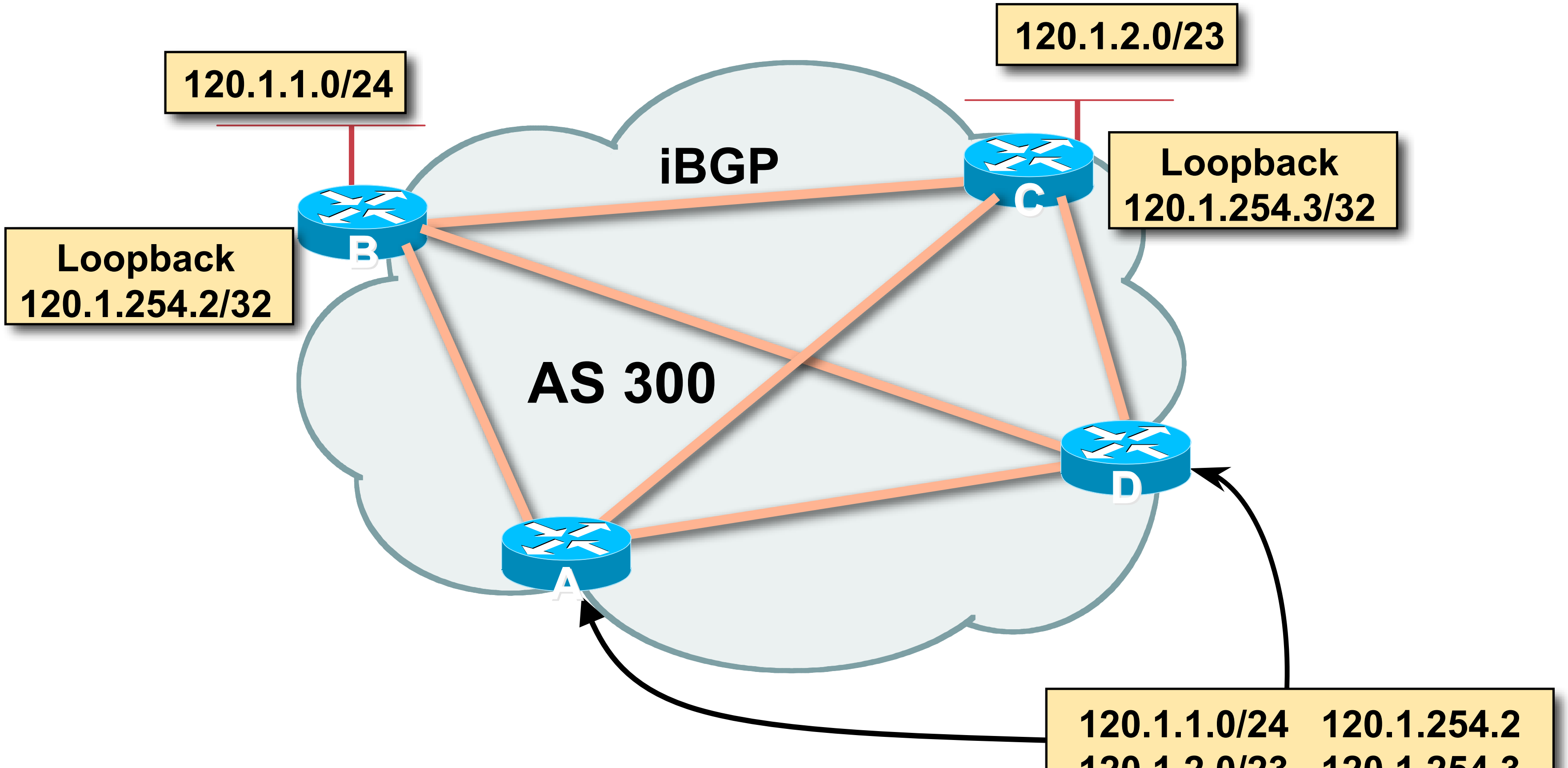


150.10.0.0/16	150.10.1.1
160.10.0.0/16	150.10.1.1

eBGP – address of external neighbour

iBGP – NEXT_HOP from eBGP

iBGP Next Hop

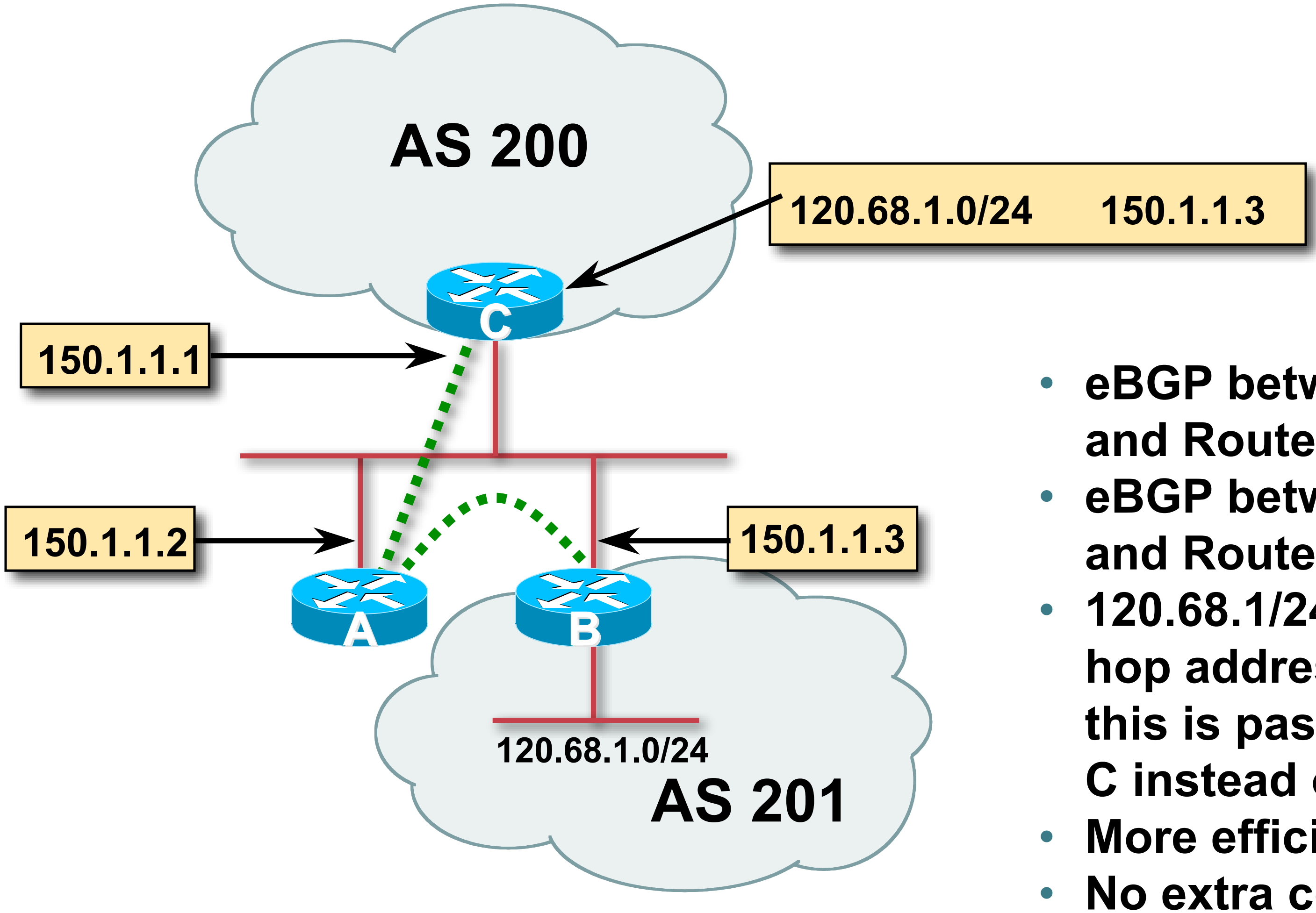


120.1.1.0/24	120.1.254.2
120.1.2.0/23	120.1.254.3

Next hop is ibgp router loopback address

Recursive route look-up

Third Party Next Hop



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

Next Hop (summary)

- IGP should carry route to next hops
- Recursive route look-up
- Unlinks BGP from actual physical topology
- Allows IGP to make intelligent forwarding decision

Origin

- Conveys the origin of the prefix
- “Historical” attribute
- Influences best path selection
- Three values: IGP, EGP, incomplete

  IGP – generated by BGP network statement

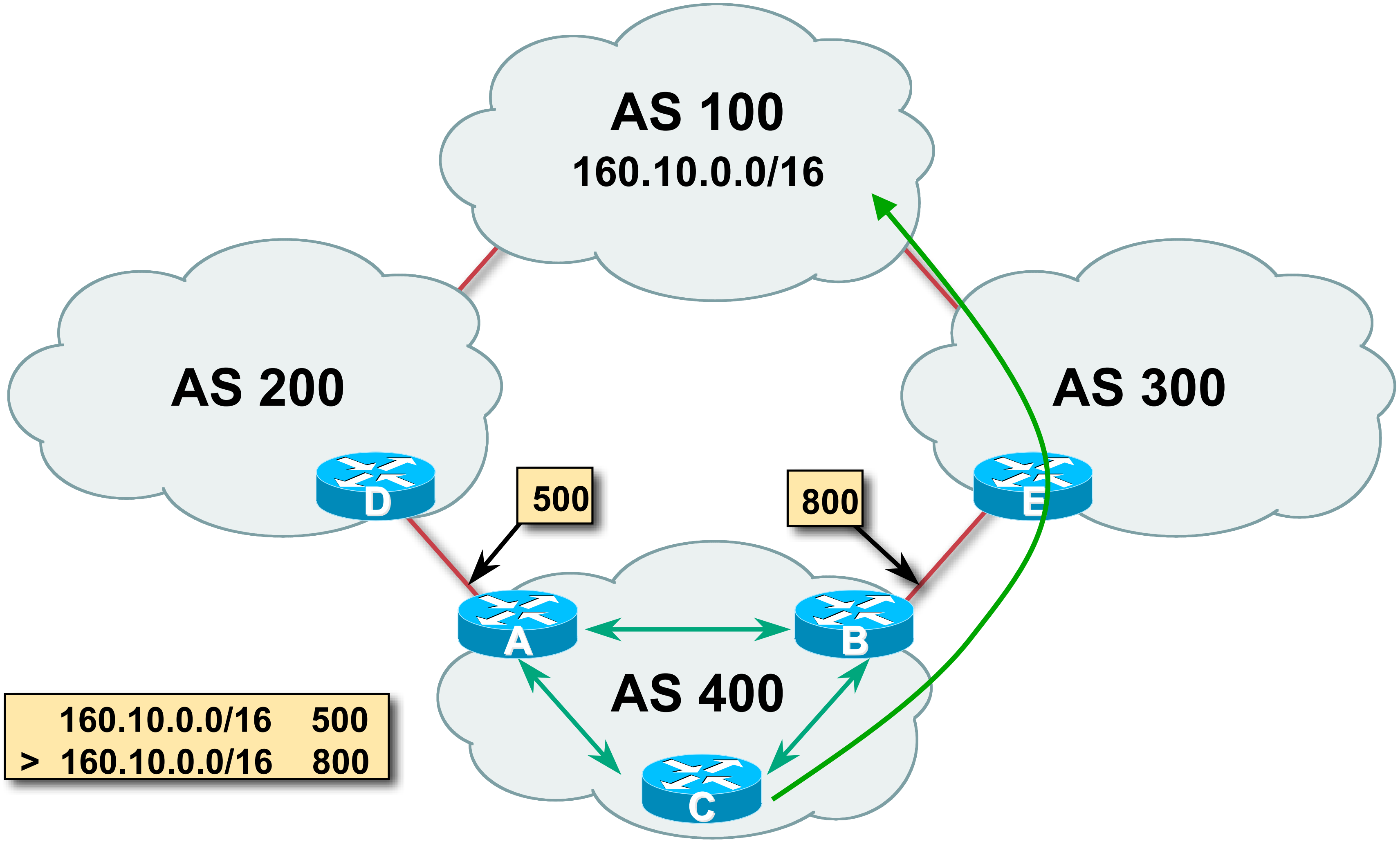
  EGP – generated by EGP

  incomplete – redistributed from another routing protocol

Aggregator

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

Local Preference



AS 100

160.10.0.0/16

AS 200

AS 300

AS 400

500

800

D

E

A

B

C

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

Local Preference

- **Local to an AS – non-transitive**

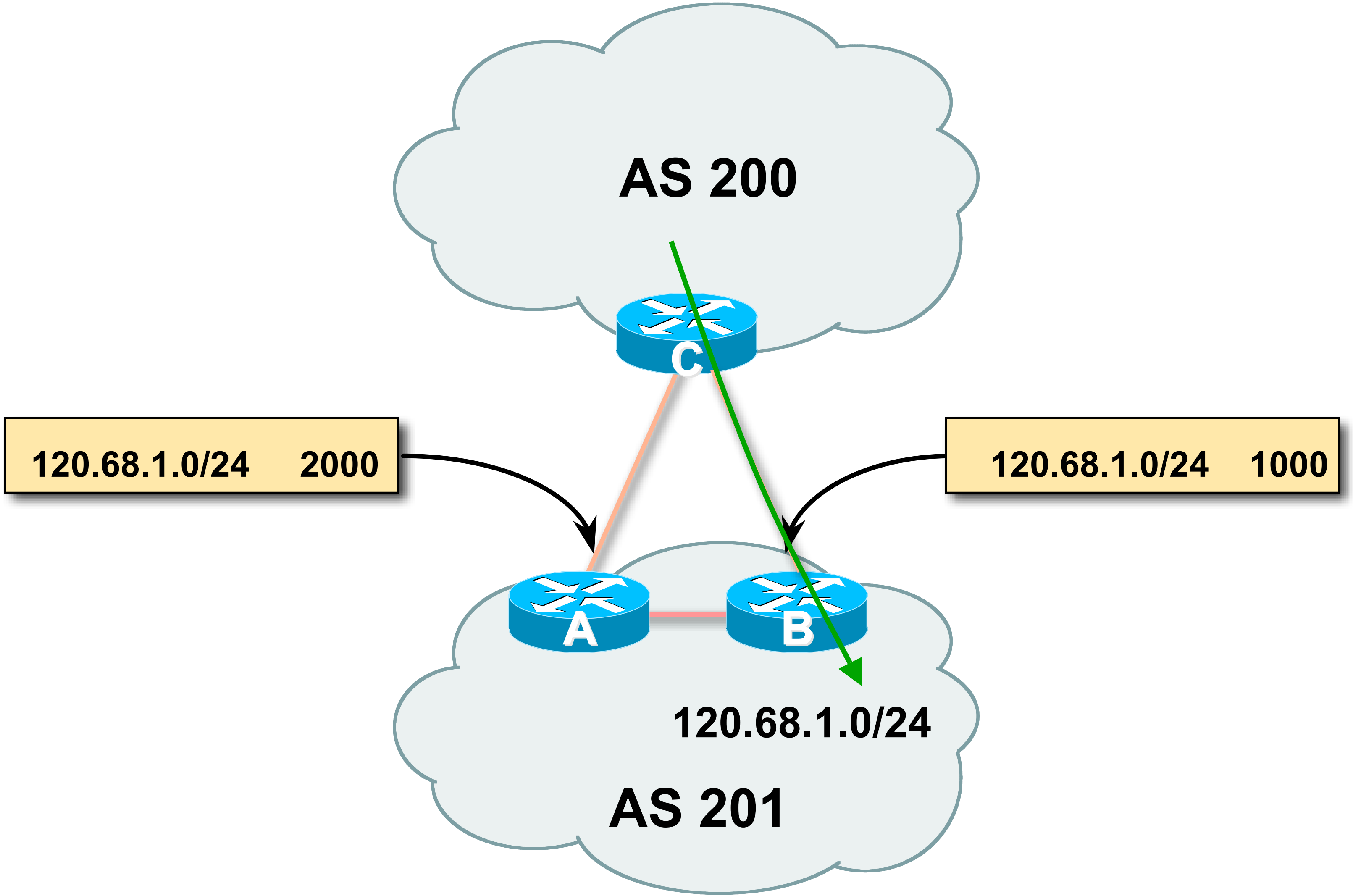
  Default local preference is 100 (IOS)

- **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

  bgp always-compared-med allows comparisons of MEDs from different ASes

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

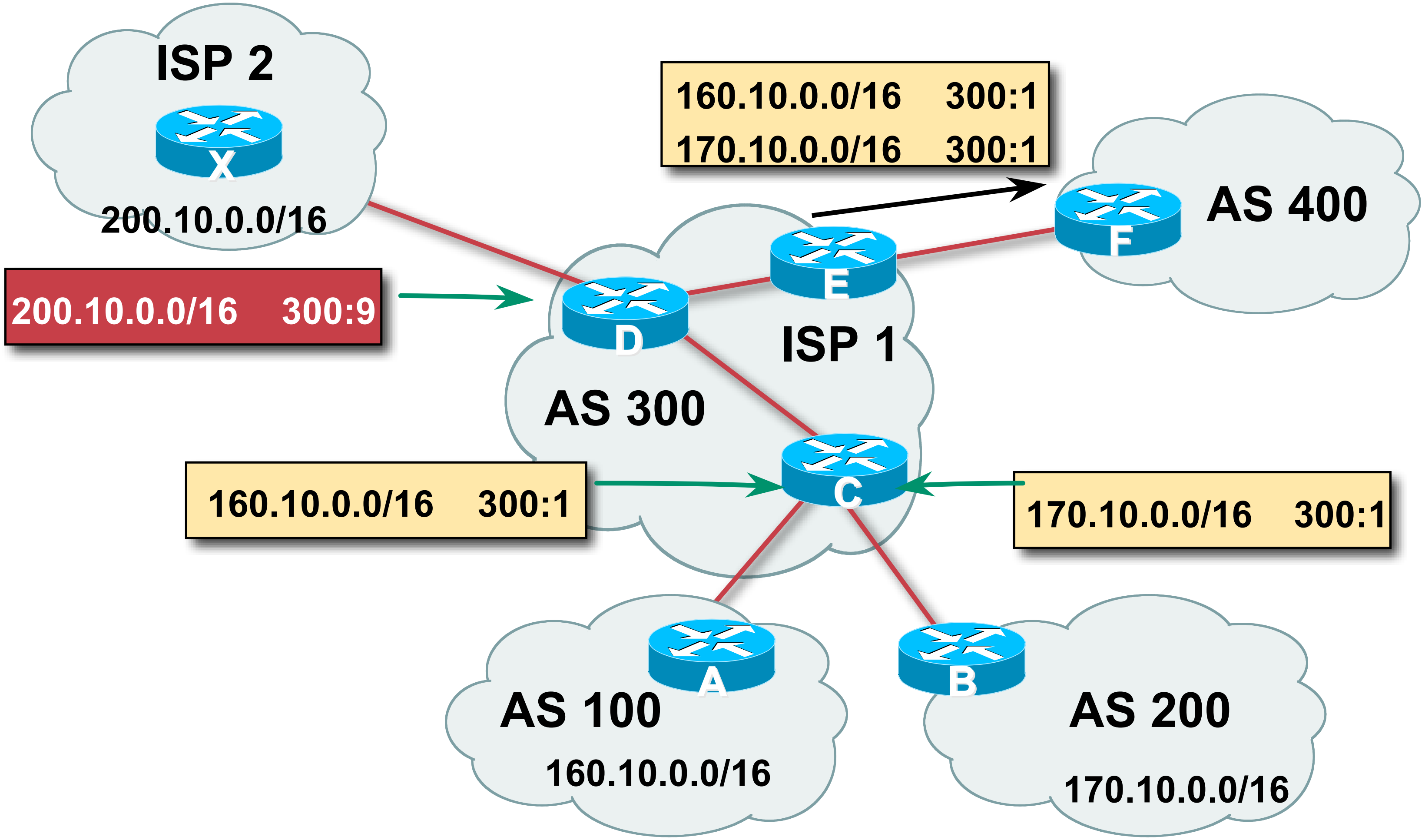
Multi-Exit Discriminator “metric confusion”

- 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 value varies according to vendor implementation
  - Original BGP spec made no recommendation
  - Some implementations said no metric was equivalent to 2^{32}-1 (the highest possible) or 2^{32}-2
  - Other implementations said no metric was equivalent to 0
- 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


ISP 2
X
200.10.0.0/16
200.10.0.0/16 300:9
160.10.0.0/16 300:1
170.10.0.0/16 300:1
AS 400
F
E
D
ISP 1
AS 300
C
160.10.0.0/16 300:1
170.10.0.0/16 300:1
A
B
AS 100
160.10.0.0/16
AS 200
170.10.0.0/16

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

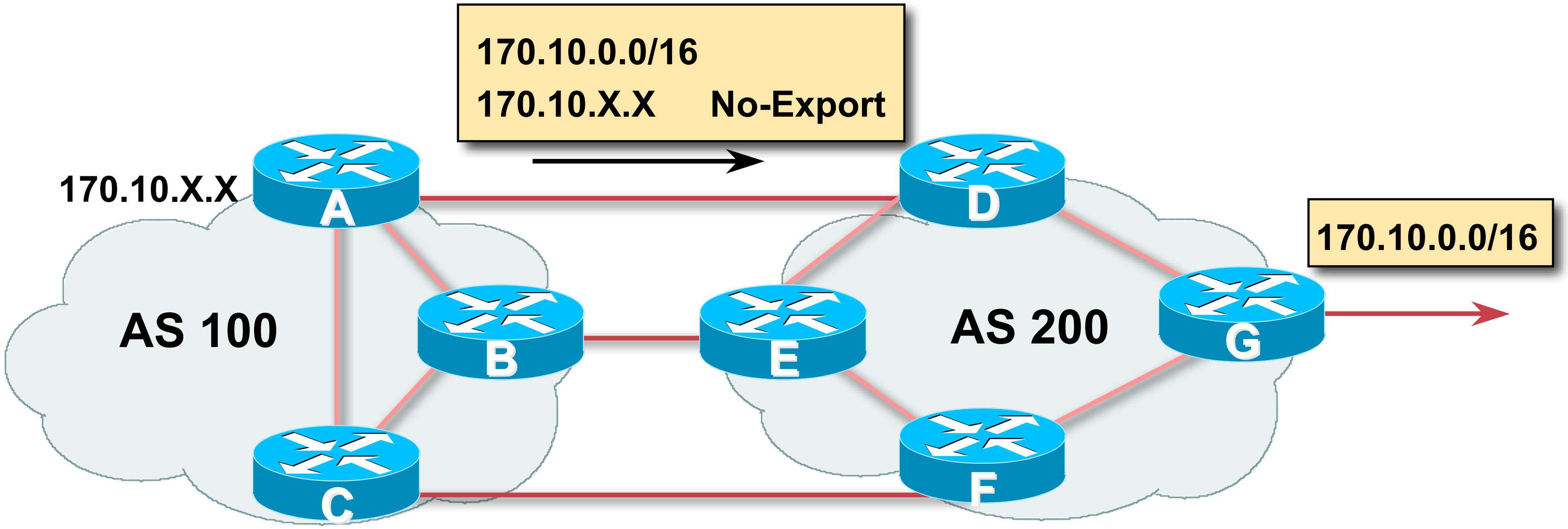
- **no-export-subconfed** **65535:65283**

  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

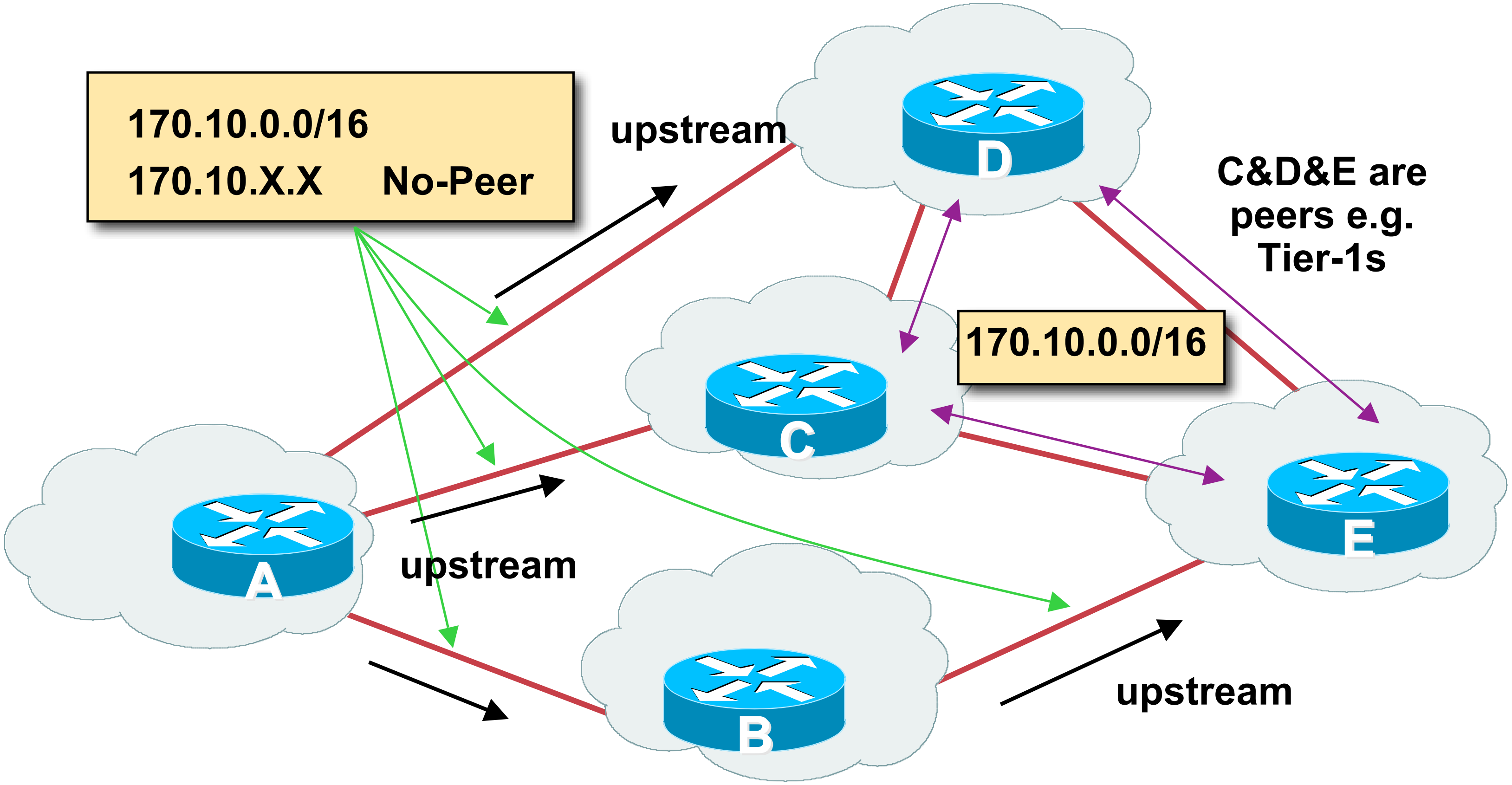


- AS100 announces aggregate and subprefixes

  aim 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)
- 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

  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

Control!

Applying Policy in BGP: Why?

- **Policies are applied to:**

  Influence BGP Path Selection by setting BGP attributes

  Determine which prefixes are announced or blocked

  Determine which AS-paths are preferred, permitted, or denied

  Determine route groupings and their effects

- **Decisions are generally based on prefix, AS-path and community**

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	[RFC2858]
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	[ID]
65	Support for 4 octet ASNs	[ID]
66	Deprecated 2003-03-06	
67	Support for Dynamic 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**
- **The other capabilities are still in development or not widely implemented or deployed yet**

BGP for Internet Service Providers

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

BGP Scaling Techniques

BGP Scaling Techniques

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

BGP Scaling Techniques

- Route Refresh
- Route flap damping
- 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 Flap Damping

Stabilising the Network

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**

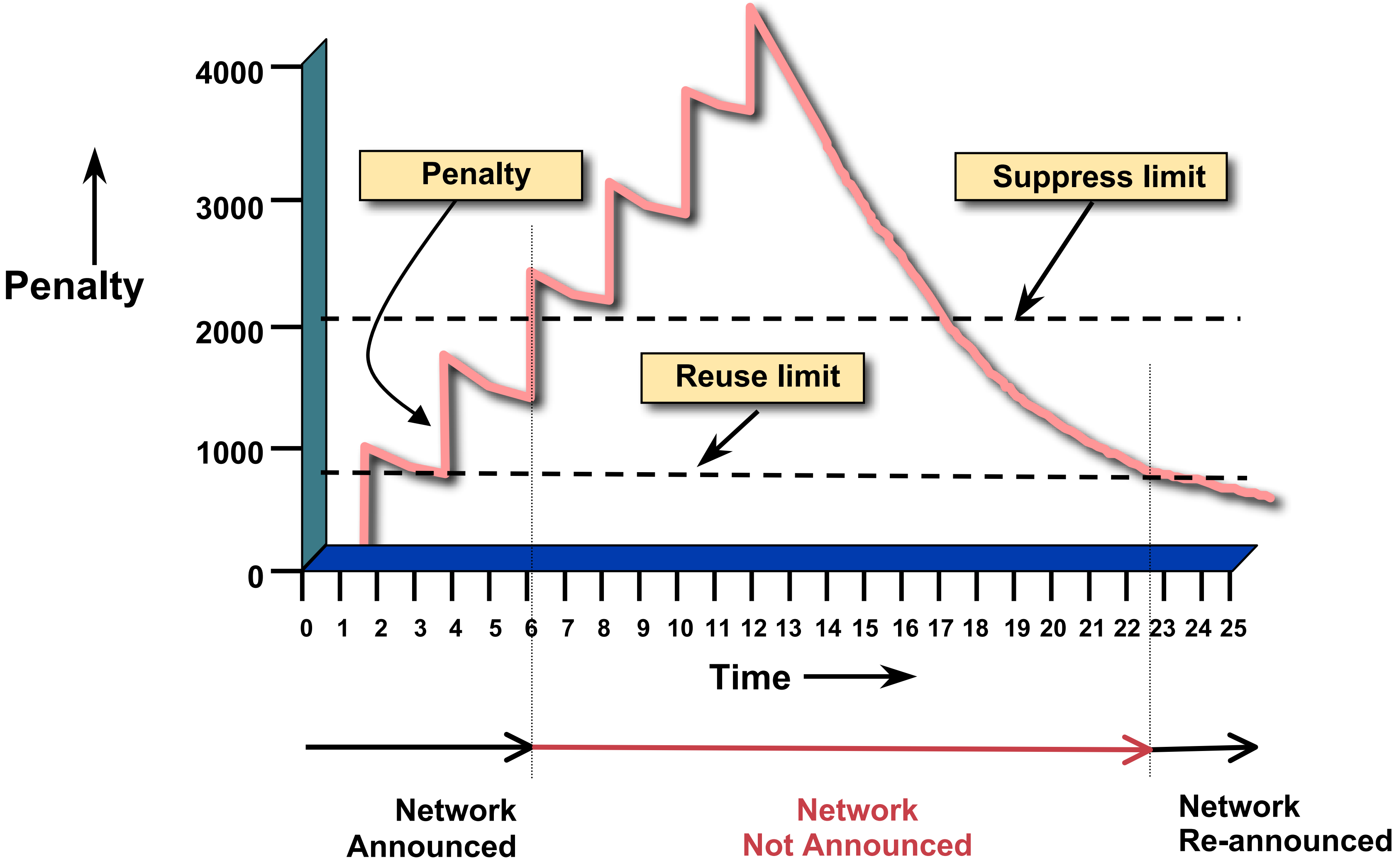
  do not advertise route to BGP peers

- **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

Implementing Flap Damping

- **Flap Damping should only be implemented to address a specific network stability problem**
- **Flap Damping can and does make stability worse**

  "Flap Amplification" from AS path attribute changes caused by BGP exploring alternate paths being unnecessarily penalised

  "Route Flap Damping Exacerbates Internet Routing Convergence"

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

Implementing Flap Damping

- If you have to implement flap damping, understand the impact on the network
  - Vendor defaults are very severe
  - Variable flap damping can bring benefits
  - Transit provider flap damping impacts peer ASes more harshly due to flap amplification
- Recommendations for ISPs
  - http://www.ripe.net/docs/ripe-229.html
  - (work by European and US ISPs a few years ago as vendor defaults were considered to be too aggressive)

Cisco Systems

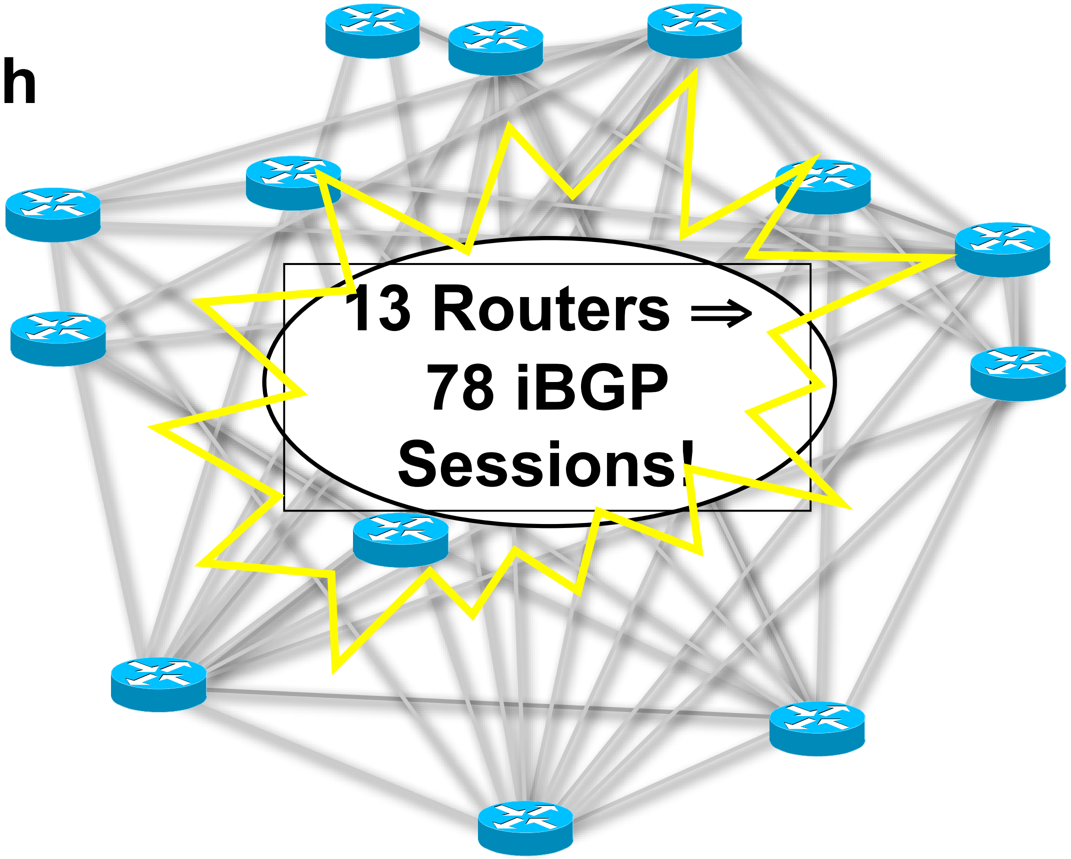
Route Reflectors

Scaling the iBGP mesh

Scaling iBGP mesh

Avoid ½n(n-1) iBGP mesh

n=1000 ⇒ nearly half a million ibgp sessions!

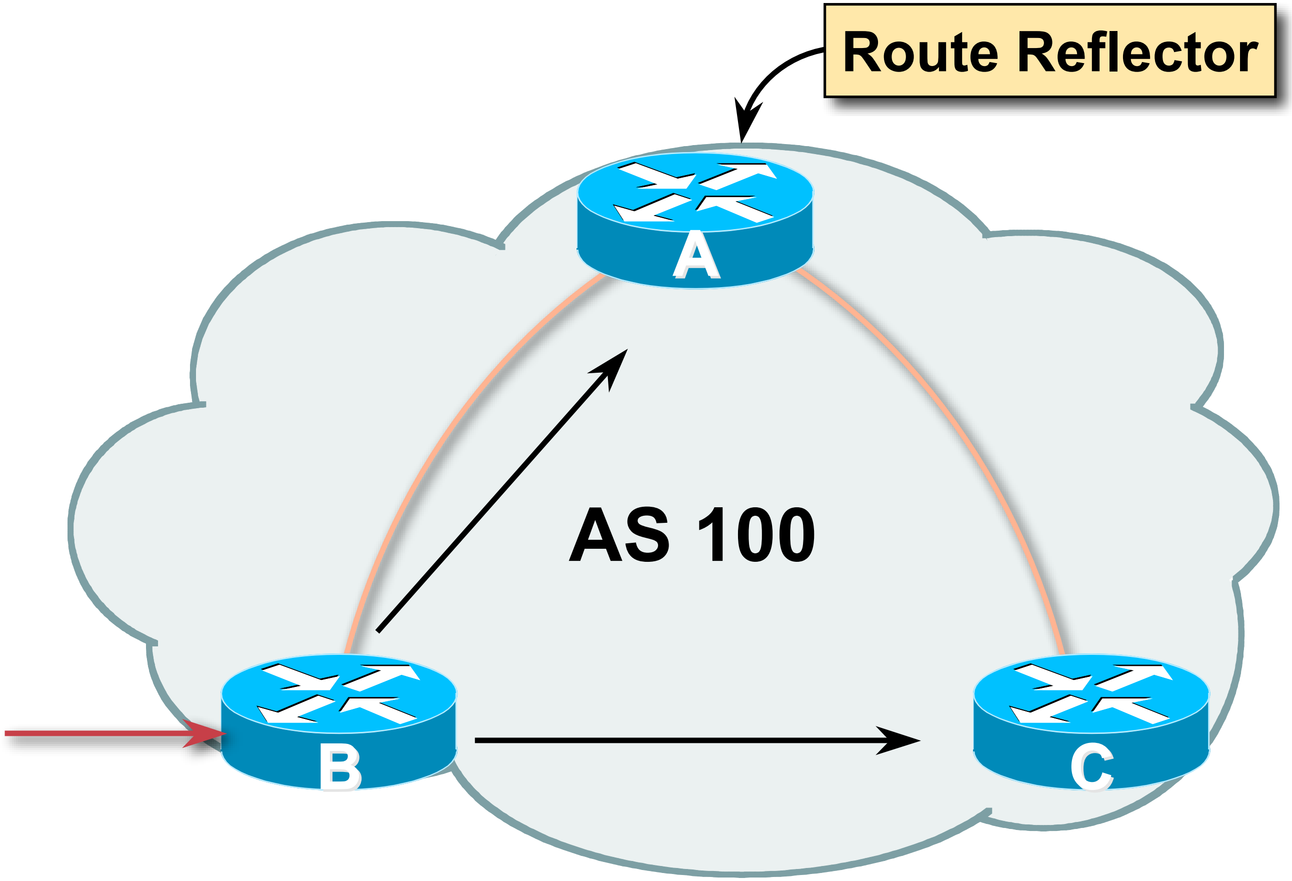


Two solutions

Route reflector – simpler to deploy and run

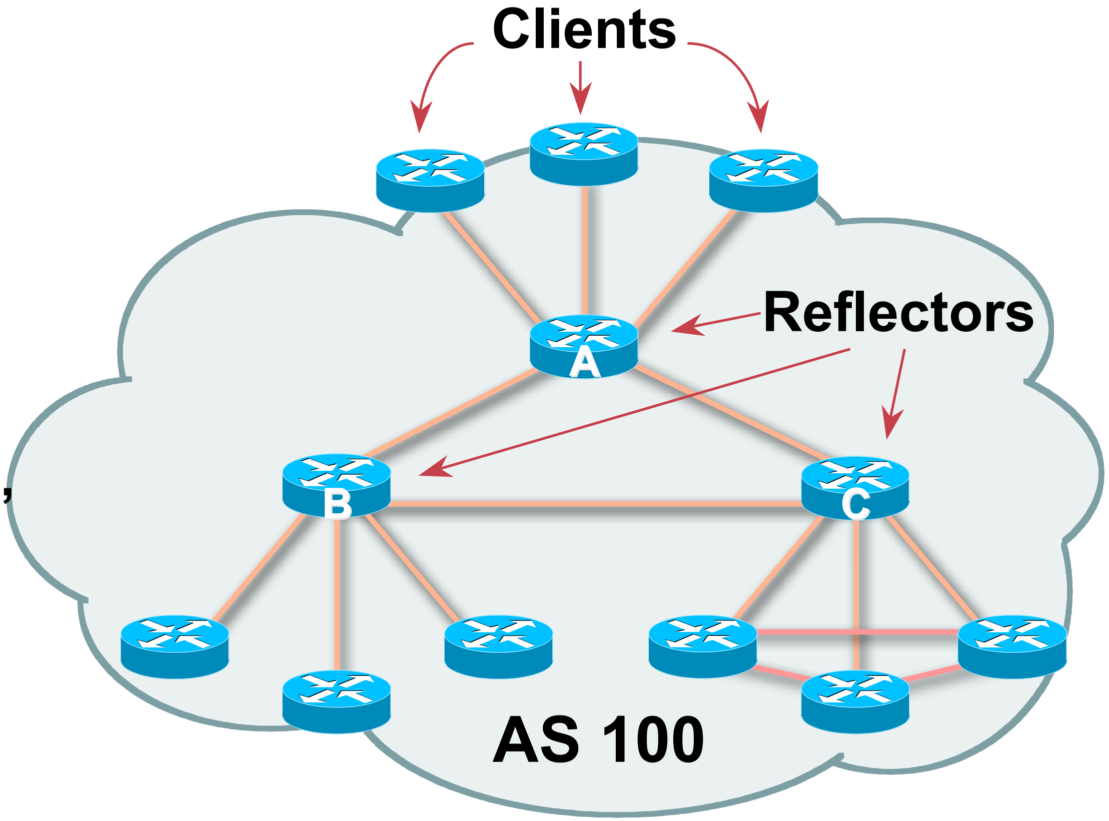
Confederation – more complex, has corner case advantages

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 RFC2796



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 Reflectors:
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 Reflectors: Redundancy

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

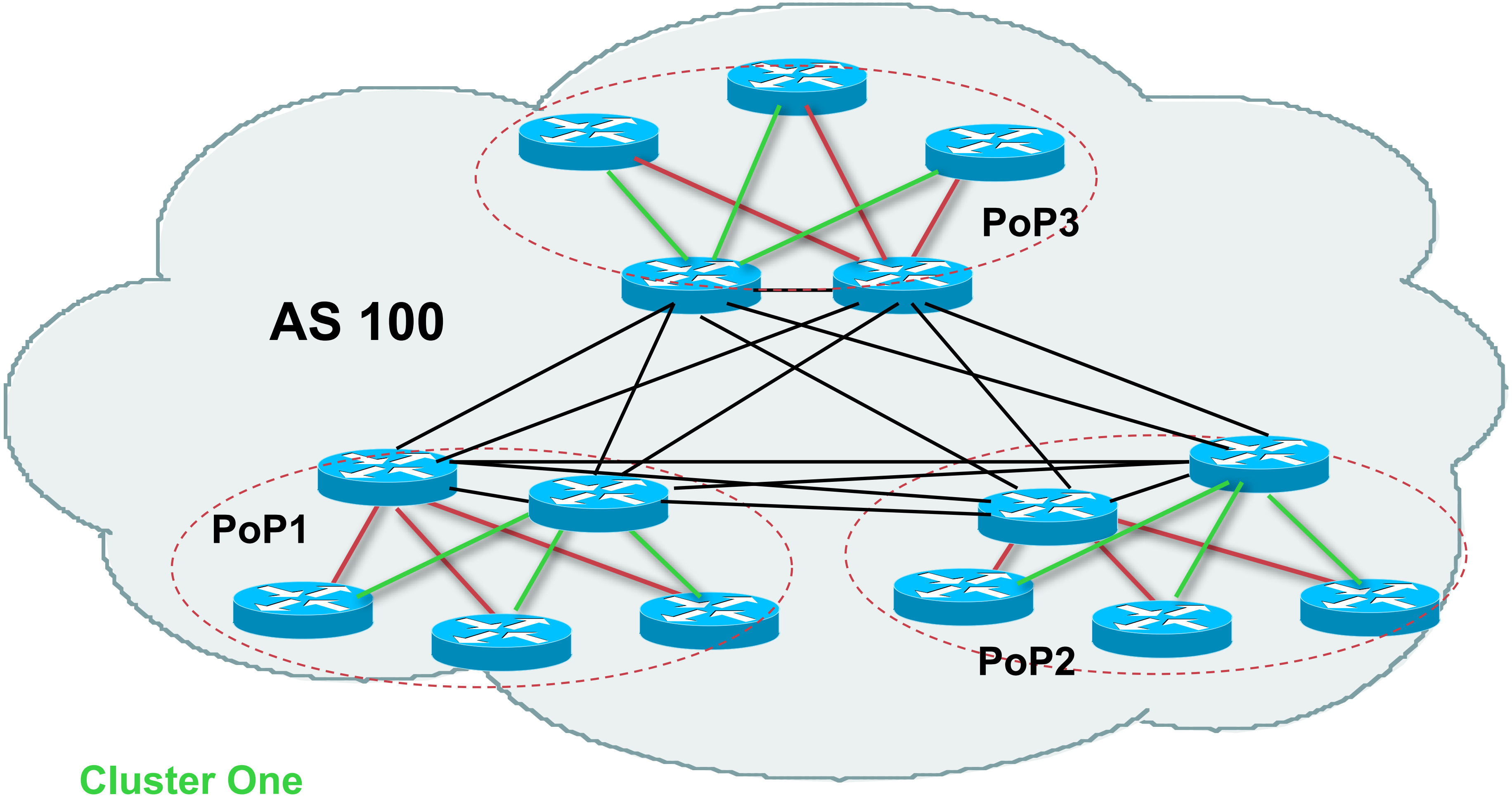
  All RRs in the cluster must have the same cluster-id (otherwise it is a different cluster)

- **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 Reflectors: Migration

- **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

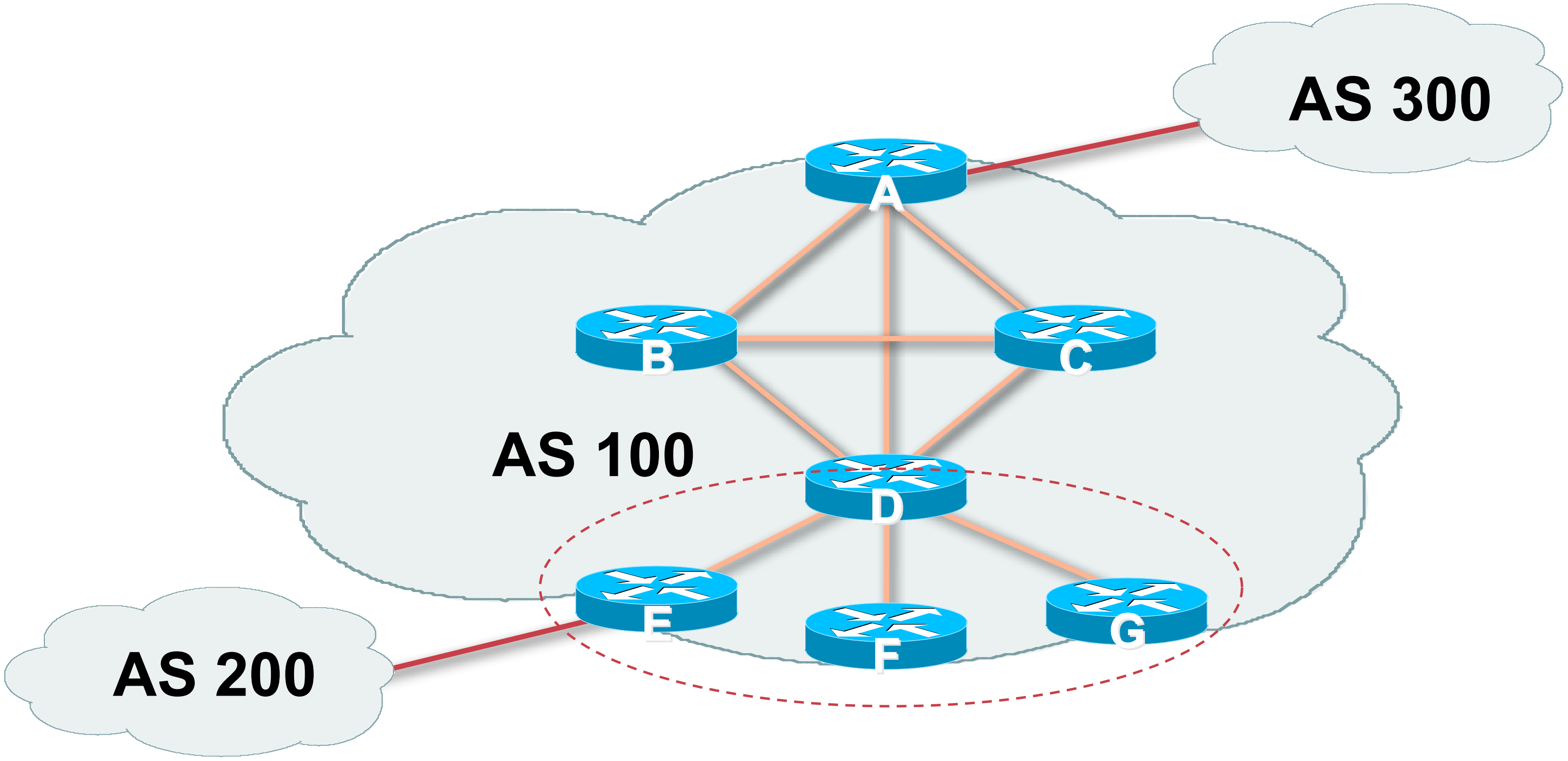
  Core routers are RR for the PoP

  Two overlaid clusters

Route Reflectors: 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 RFC3065

Confederations (Cont.)

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

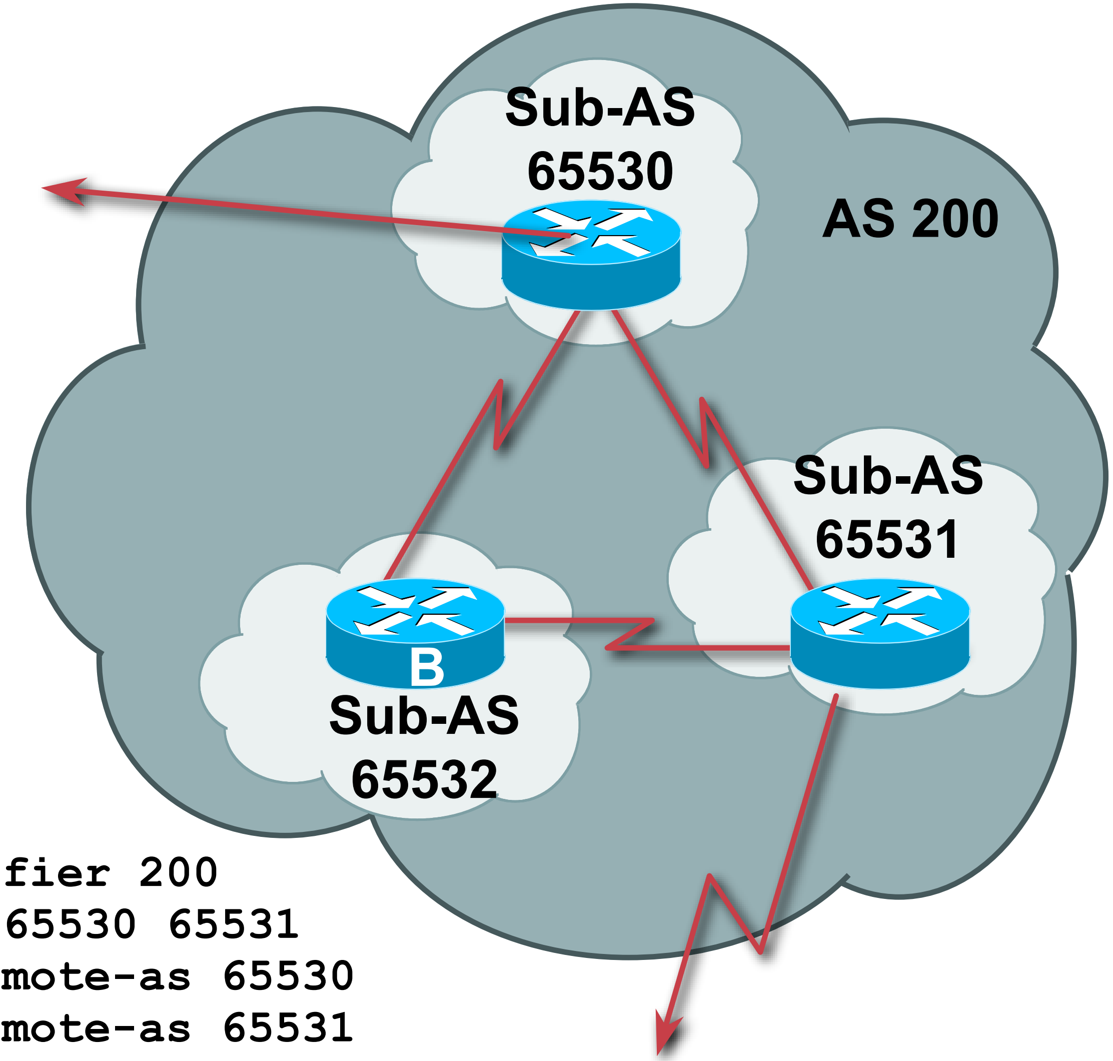
  Each sub-AS uses a number from the private AS range (64512-65534)

- **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

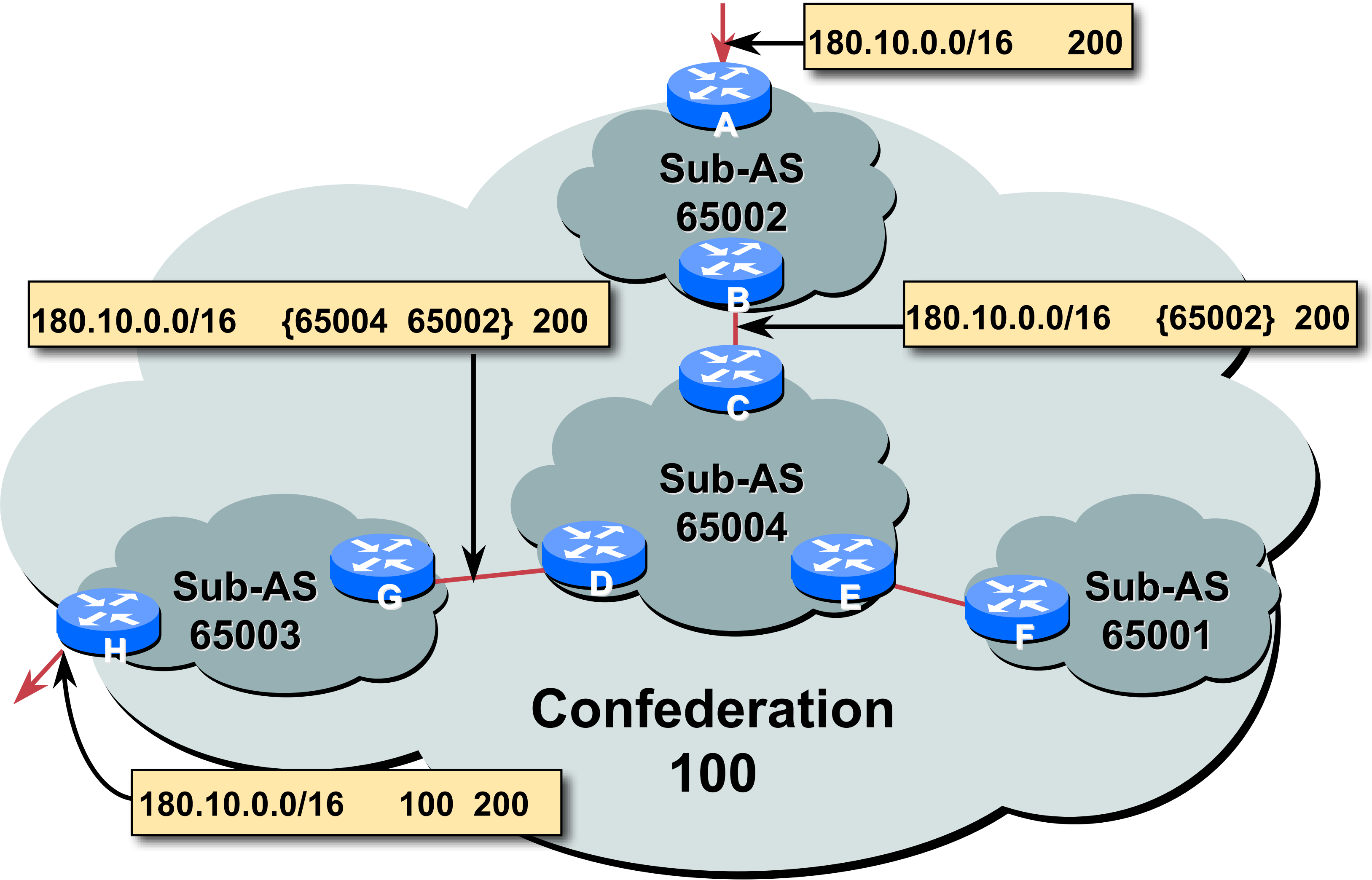
Confederations



- **Configuration (rtr B):**

```
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**

BGP Scaling Techniques

- Route Refresh
  - Use should be mandatory
- Route flap damping
  - Only use if you understand why
- Route Reflectors/Confederations
  - The only way to scale iBGP mesh

BGP for Internet Service Providers

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

Service Providers use of Communities

Some examples of how ISPs make life easier for themselves

BGP Communities

- Another ISP “scaling technique”
- Prefixes are grouped into different “classes” or communities within the ISP network
- Each community means a different thing, has a different result in the ISP network

BGP Communities

- Communities are generally set at the edge of the ISP network

    Customer edge: customer prefixes belong to different communities depending on the services they have purchased

    Internet edge: transit provider prefixes belong to difference communities, depending on the loadsharing or traffic engineering requirements of the local ISP, or what the demands from its BGP customers might be

- Two simple examples follow to explain the concept

Community Example – Customer Edge

- This demonstrates how communities might be used at the customer edge of an ISP network
- ISP has three connections to the Internet:

  IXP connection, for local peers

  Private peering with a competing ISP in the region

  Transit provider, who provides visibility to the entire Internet
- Customers have the option of purchasing combinations of the above connections

Community Example – Customer Edge

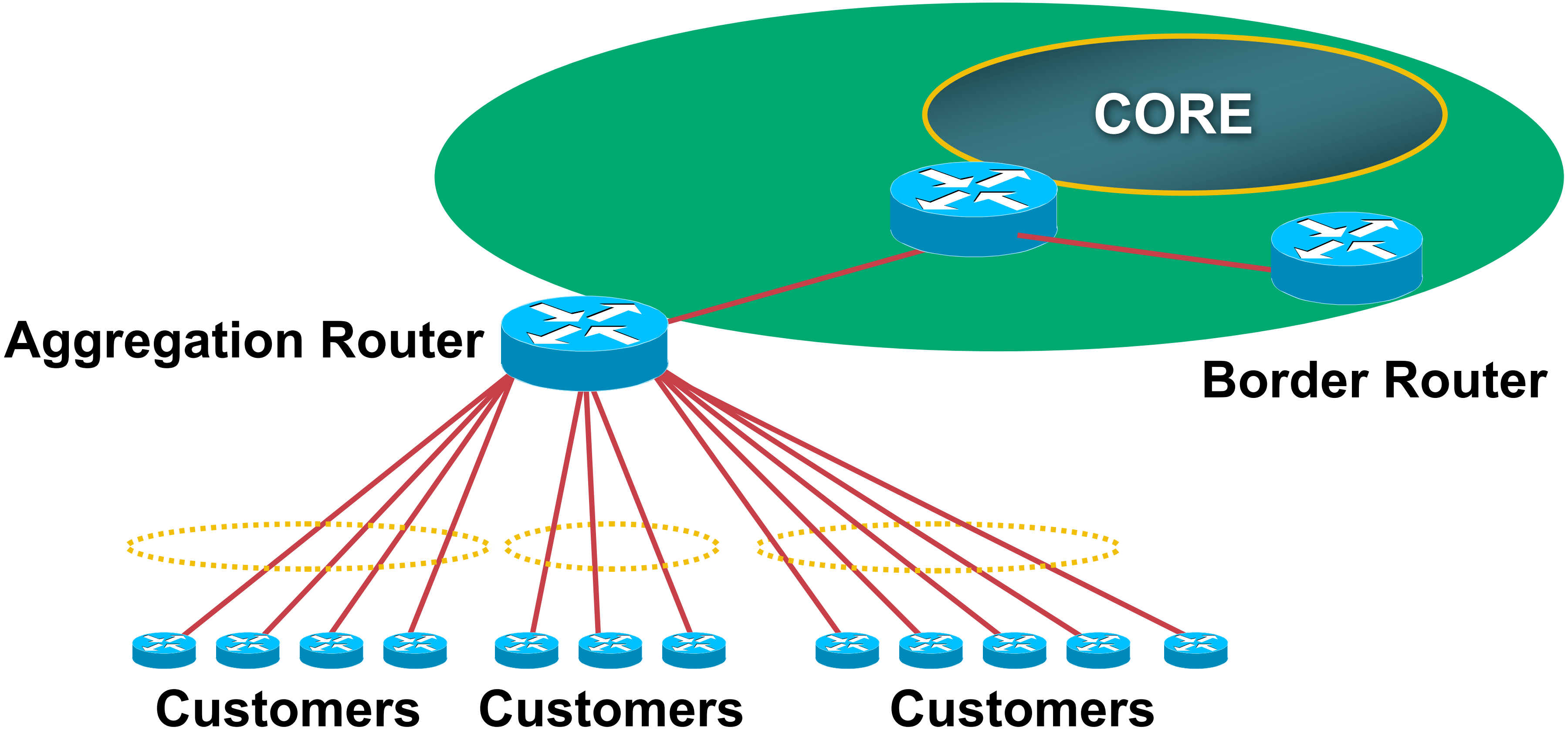
- **Community assignments:**

IXP connection:	**community 100:2100**
Private peer:	**community 100:2200**

- **Customer who buys local connectivity (via IXP) is put in community 100:2100**
- **Customer who buys peer connectivity is put in community 100:2200**
- **Customer who wants both IXP and peer connectivity is put in 100:2100 and 100:2200**
- **Customer who wants "the Internet" has no community set**

  We are going to announce his prefix everywhere

Community Example – Customer Edge



Communities set at the aggregation router where the prefix is injected into the ISP's iBGP

Community Example – Customer Edge

- **No need to alter filters at the network border when adding a new customer**
- **New customer simply is added to the appropriate community**

  Border filters already in place take care of announcements

  ⇒ Ease of operation!

Community Example – Internet Edge

- This demonstrates how communities might be used at the peering edge of an ISP network
- ISP has four types of BGP peers:
  - Customer
  - IXP peer
  - Private peer
  - Transit provider
- The prefixes received from each can be classified using communities
- Customers can opt to receive any or all of the above

Community Example – Internet Edge

- Community assignments:

Customer prefix:	community 100:3000
IXP prefix:	community 100:3100
Private peer prefix:	community 100:3200

- BGP customer who buys local connectivity gets 100:3000
- BGP customer who buys local and IXP connectivity receives community 100:3000 and 100:3100
- BGP customer who buys full peer connectivity receives community 100:3000, 100:3100, and 100:3200
- Customer who wants “the Internet” gets everything

  Gets default route originated by aggregation router

  Or pays money to get all 190k prefixes

Community Example – Internet Edge

- No need to create customised filters when adding customers

  Border router already sets communities

  Installation engineers pick the appropriate community set when establishing the customer BGP session

  ⇒ Ease of operation!

Community Example – Summary

- Two examples of customer edge and internet edge can be combined to form a simple community solution for ISP prefix policy control
- More experienced operators tend to have more sophisticated options available

  Advice is to start with the easy examples given, and then proceed onwards as experience is gained

BGP for Internet Service Providers

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

Deploying BGP in an ISP Network

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

Deploying BGP

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

The role of IGP and iBGP

Ships in the night?

Or

Good foundations?

BGP versus OSPF/ISIS

- **Internal Routing Protocols (IGPs)**

  examples are ISIS and OSPF

  used for carrying infrastructure addresses

  NOT used for carrying Internet prefixes or customer prefixes

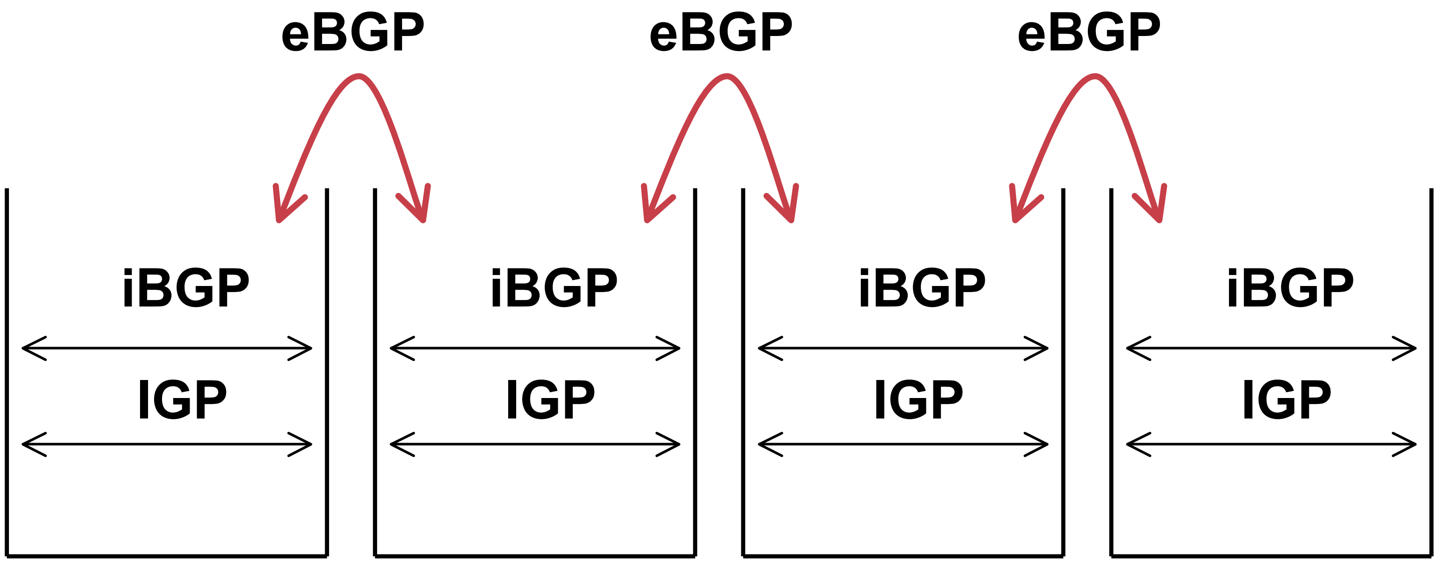
  design goal is to minimise number of prefixes in IGP to aid scalability and rapid convergence

BGP versus OSPF/ISIS

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

BGP/IGP model used in ISP networks

- **Model representation**



BGP versus OSPF/ISIS

- DO NOT:

  distribute BGP prefixes into an IGP

  distribute IGP routes into BGP

  use an IGP to carry customer prefixes

- YOUR NETWORK WILL NOT SCALE

Injecting prefixes into iBGP

- Use iBGP to carry customer prefixes

  don’t ever use IGP

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

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

  i.e. avoid iBGP flaps caused by interface flaps

Aggregation

Quality or Quantity?

Aggregation

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

Aggregation

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

  Not on the network borders!

Announcing an Aggregate

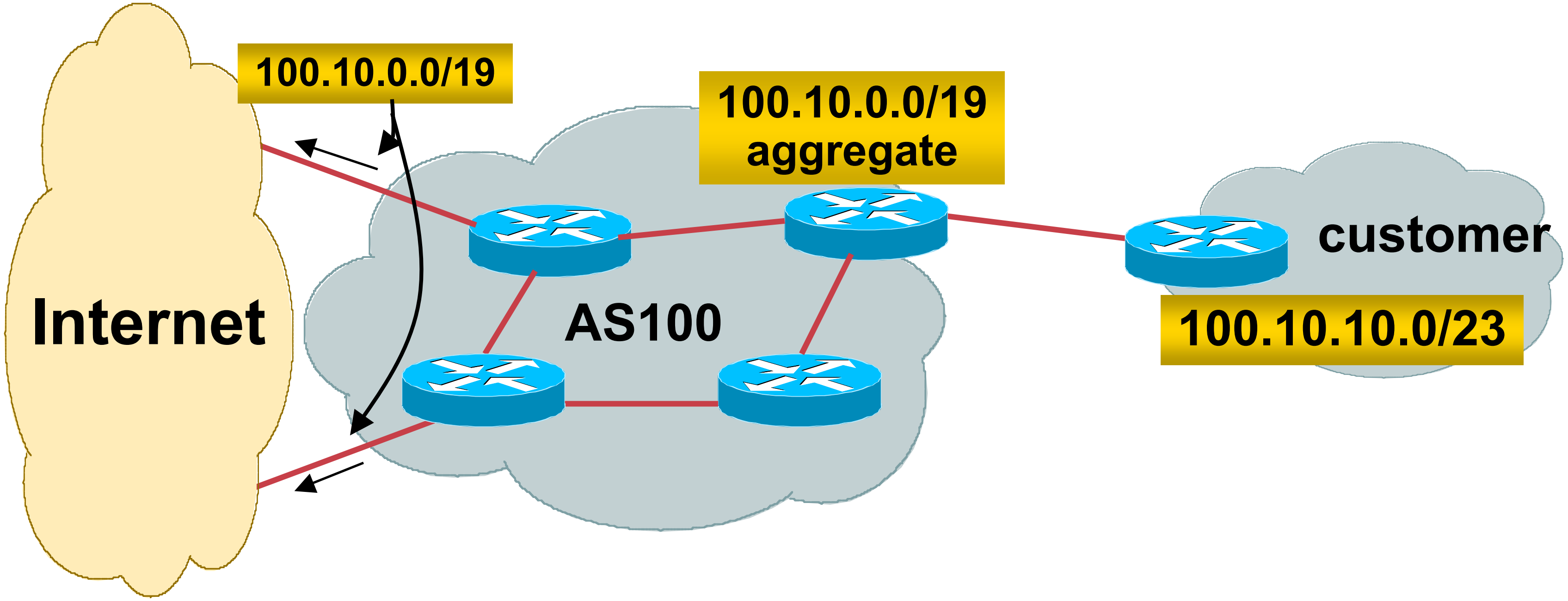
- ISPs who don’t and won’t aggregate are held in poor regard by community
- Registries publish their minimum allocation size

  Anything from a /20 to a /22 depending on RIR

- No real reason to see anything longer than a /22 prefix in the Internet

  BUT there are currently >99000 /24s!

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

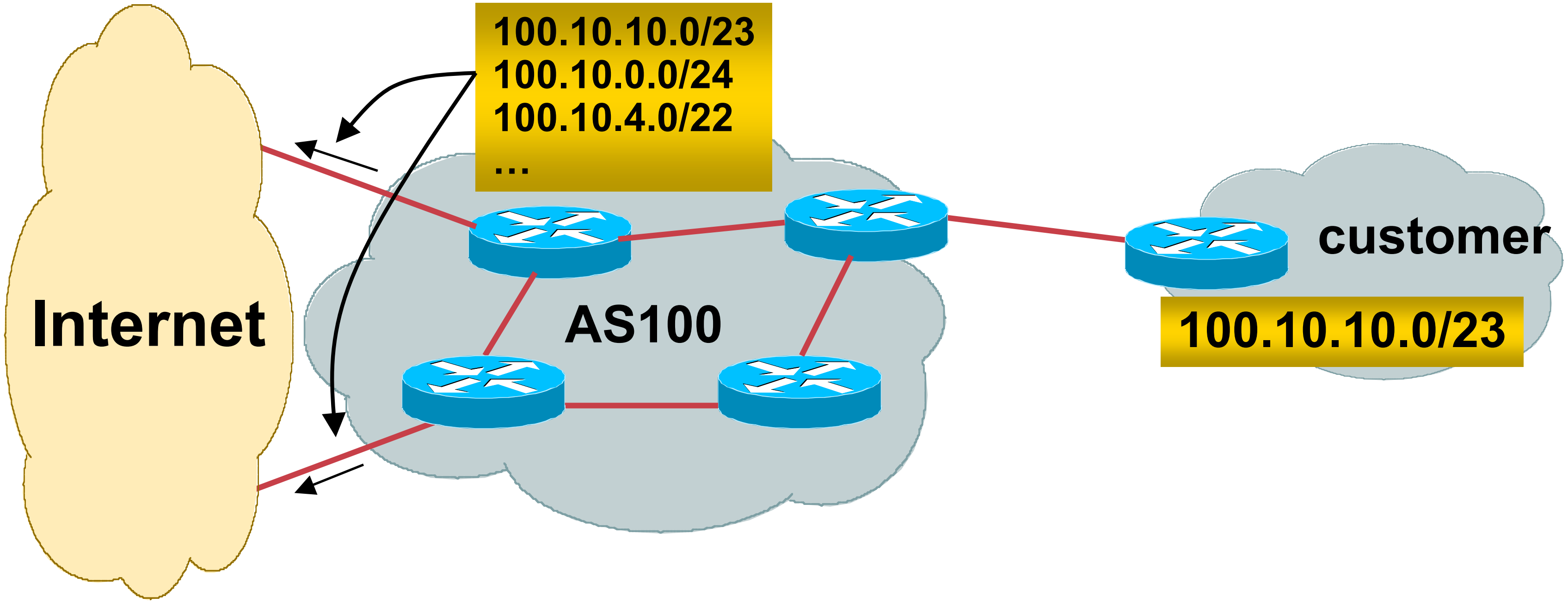
    no damping by other ISPs

- 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 – 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 – Summary

- **Good example is what everyone should do!**

  Adds to Internet stability

  Reduces size of routing table

  Reduces routing churn

  Improves Internet QoS for everyone

- **Bad example is what too many still do!**

  Why? Lack of knowledge?

  Laziness?

The Internet Today (February 2006)

- **Current Internet Routing Table Statistics**

BGP Routing Table Entries	182208
Prefixes after maximum aggregation	101495
Unique prefixes in Internet	88775
Prefixes smaller than registry alloc	88835
/24s announced	99131
only 5764 /24s are from 192.0.0.0/8	
ASes in use	21583

“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**	*74/8*	*0*	*124/8*	*0*	**205/8**	**2584**
41/8	*0*	*75/8*	*0*	*125/8*	*0*	**206/8**	**3127**
58/8	*0*	*76/8*	*0*	*126/8*	*0*	**207/8**	**2723**
59/8	*0*	*80/8*	*0*	*188/8*	*0*	**208/8**	**2817**
60/8	*0*	*81/8*	*0*	*189/8*	*0*	**209/8**	**2574**
61/8	**3**	*82/8*	*0*	*190/8*	*0*	**210/8**	**617**
62/8	**87**	*83/8*	*0*	**192/8**	**6275**	*211/8*	*0*
63/8	**20**	*84/8*	*0*	**193/8**	**2390**	**212/8**	**717**
64/8	*0*	*85/8*	*0*	**194/8**	**2932**	**213/8**	**1**
65/8	*0*	*86/8*	*0*	**195/8**	**1338**	**216/8**	**943**
66/8	*0*	*87/8*	*0*	**196/8**	**513**	*217/8*	*0*
67/8	*0*	*88/8*	*0*	**198/8**	**4034**	*218/8*	*0*
68/8	*0*	*89/8*	*0*	**199/8**	**3495**	*219/8*	*0*
69/8	*0*	*90/8*	*0*	**200/8**	**1348**	*220/8*	*0*
70/8	*0*	*91/8*	*0*	*201/8*	*0*	*221/8*	*0*
71/8	*0*	*121/8*	*0*	**202/8**	**2276**	*222/8*	*0*
72/8	*0*	*122/8*	*0*	**203/8**	**3622**		
73/8	*0*	*123/8*	*0*	**204/8**	**3792**		

“The New Swamp”
RIR Space – February 2006

RIR blocks contribute 161287 prefixes or 88% of the Internet Routing Table

Block	Networks	Block	Networks	Block	Networks	Block	Networks
24/8	3001	74/8	109	124/8	292	205/8	2934
41/8	41	75/8	2	125/8	682	206/8	3879
58/8	606	76/8	4	126/8	27	207/8	4385
59/8	628	80/8	1925	188/8	1	208/8	3239
60/8	468	81/8	1350	189/8	0	209/8	5611
61/8	2396	82/8	1158	190/8	39	210/8	3908
62/8	1860	83/8	1130	192/8	6927	211/8	2291
63/8	2837	84/8	971	193/8	5203	212/8	2920
64/8	5374	85/8	1426	194/8	4061	213/8	3071
65/8	3785	86/8	650	195/8	3519	216/8	6893
66/8	6292	87/8	629	196/8	1264	217/8	2590
67/8	1832	88/8	328	198/8	4908	218/8	1220
68/8	3069	89/8	113	199/8	4156	219/8	1003
69/8	3315	90/8	2	200/8	6757	220/8	1657
70/8	1597	91/8	2	201/8	1614	221/8	765
71/8	888	121/8	0	202/8	9759	222/8	914
72/8	1772	122/8	0	203/8	9527		
73/8	274	123/8	0	204/8	5474		

“The New Swamp” Summary

- RIR space shows creeping deaggregation

  It seems that an RIR /8 block averages around 4000 prefixes once fully allocated

  So their existing 70 /8s will eventually cause 280000 prefix announcements

- Food for thought:

  Remaining 62 unallocated /8s and the 70 RIR /8s combined will cause:

  528000 prefixes with 4000 prefixes per /8 density

  792000 prefixes with 6000 prefixes per /8 density

  Plus 12% due to “non RIR space deaggregation”

  → Routing Table size of 887040 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)

- 179170 total announcements
- 116237 prefixes

  After aggregating including full AS PATH info

  i.e. including each ASN's traffic engineering

  35% saving possible

- 101167 prefixes

  After aggregating by Origin AS

  i.e. ignoring each ASN's traffic engineering

  8% saving possible

The excuses

- Traffic engineering causes 8% 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

Efforts to improve aggregation
The CIDR Report

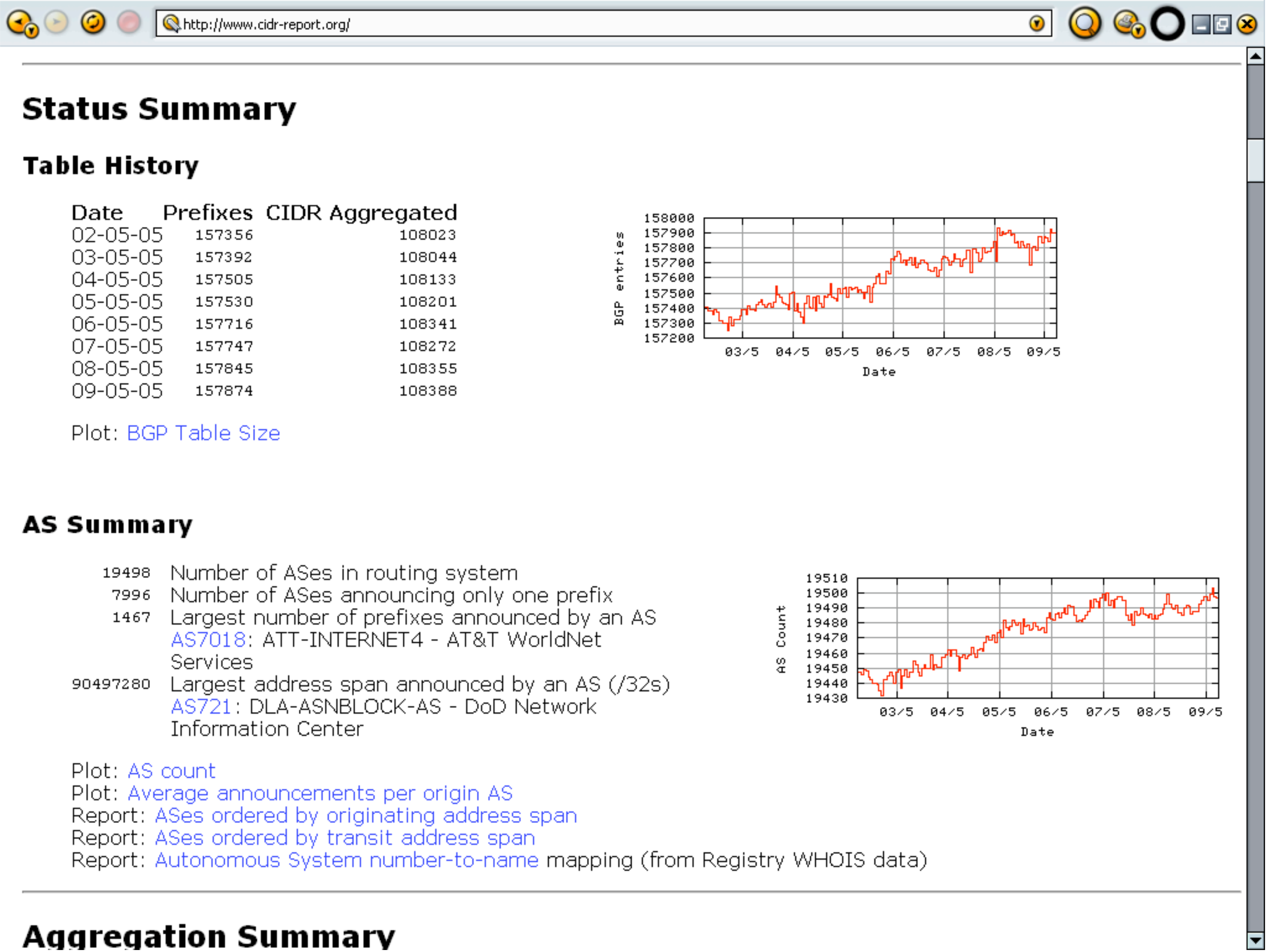
- Also computes the size of the routing table assuming ISPs performed optimal aggregation
- Website allows searches and computations of aggregation to be made on a per AS basis

  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
02-05-05	157356	108023
03-05-05	157392	108044
04-05-05	157505	108133
05-05-05	157530	108201
06-05-05	157716	108341
07-05-05	157747	108272
08-05-05	157845	108355
09-05-05	157874	108388

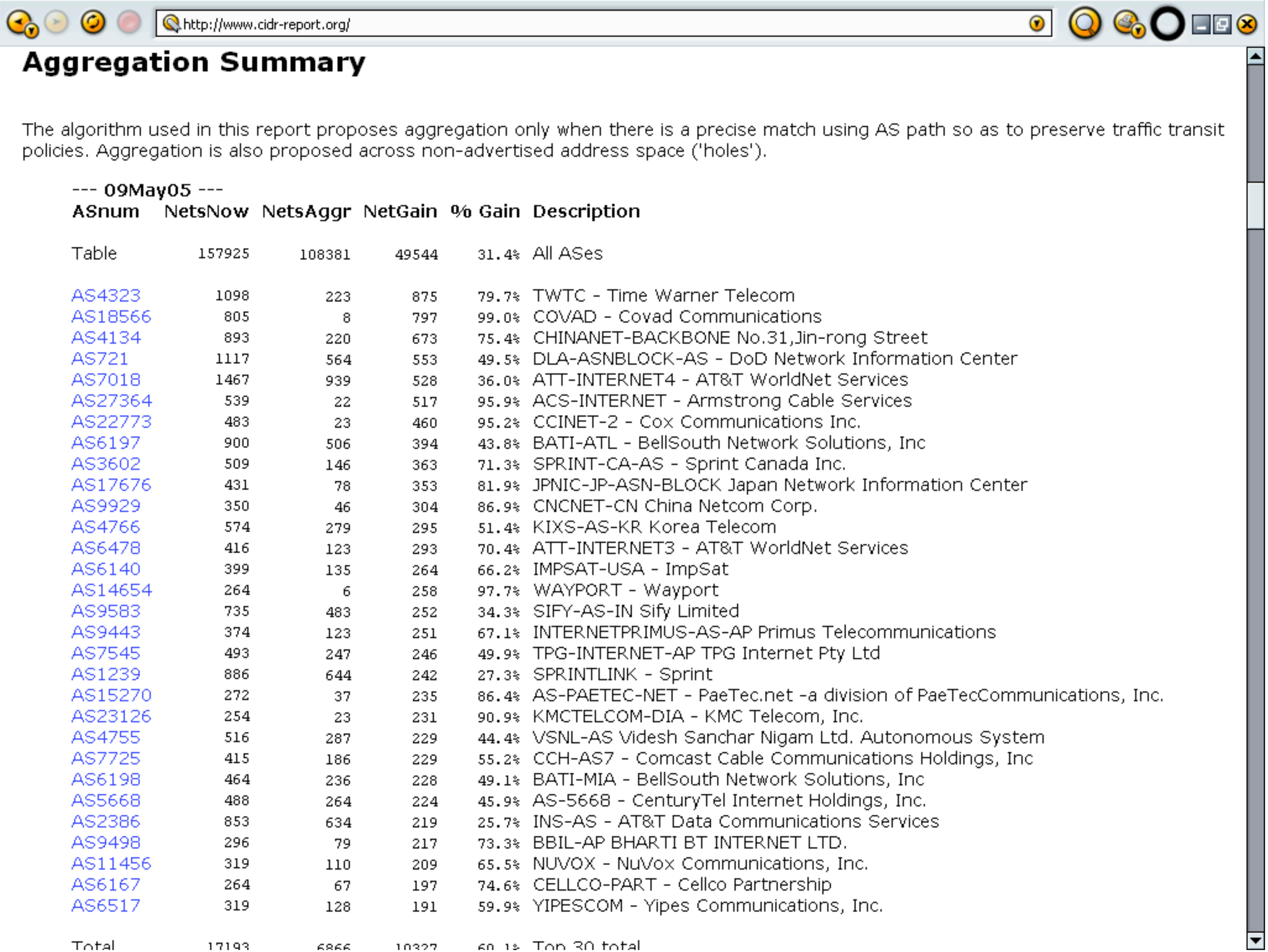
Plot: BGP Table Size

AS Summary

19498 Number of ASes in routing system
7996 Number of ASes announcing only one prefix
1467 Largest number of prefixes announced by an AS
AS7018: ATT-INTERNET4 - AT&T WorldNet Services
90497280 Largest address span announced by an AS (/32s)
AS721: DLA-ASNBLOCK-AS - 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

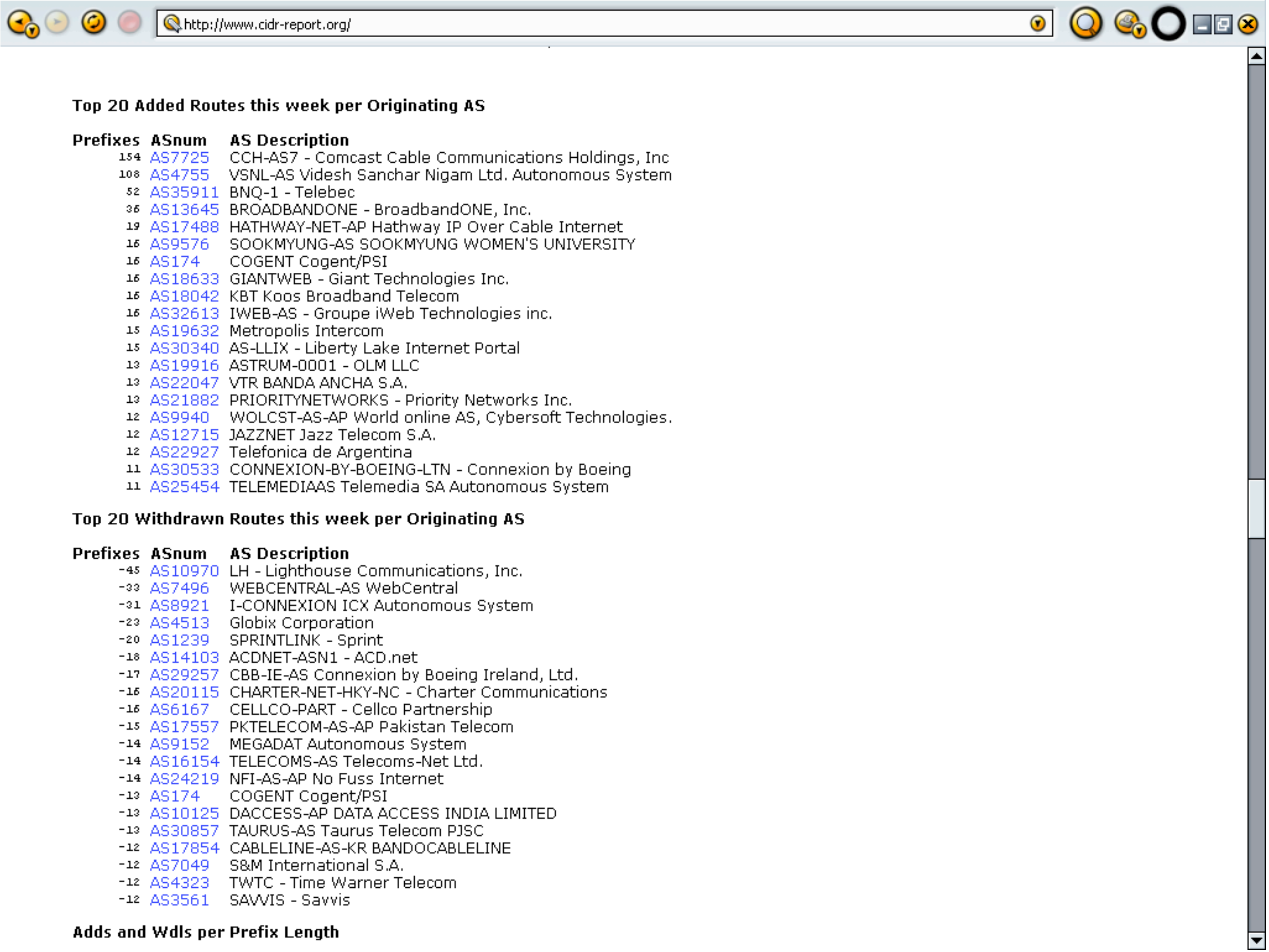


Aggregation Summary

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').

--- 09May05 ---

ASnum	NetsNow	NetsAggr	NetGain	% Gain	Description
Table	157925	108381	49544	31.4%	All ASes
AS4323	1098	223	875	79.7%	TWTC - Time Warner Telecom
AS18566	805	8	797	99.0%	COVAD - Covad Communications
AS4134	893	220	673	75.4%	CHINANET-BACKBONE No.31,Jin-rong Street
AS721	1117	564	553	49.5%	DLA-ASNBLOCK-AS - DoD Network Information Center
AS7018	1467	939	528	36.0%	ATT-INTERNET4 - AT&T WorldNet Services
AS27364	539	22	517	95.9%	ACS-INTERNET - Armstrong Cable Services
AS22773	483	23	460	95.2%	CCINET-2 - Cox Communications Inc.
AS6197	900	506	394	43.8%	BATI-ATL - BellSouth Network Solutions, Inc
AS3602	509	146	363	71.3%	SPRINT-CA-AS - Sprint Canada Inc.
AS17676	431	78	353	81.9%	JPNIC-JP-ASN-BLOCK Japan Network Information Center
AS9929	350	46	304	86.9%	CNCNET-CN China Netcom Corp.
AS4766	574	279	295	51.4%	KIXS-AS-KR Korea Telecom
AS6478	416	123	293	70.4%	ATT-INTERNET3 - AT&T WorldNet Services
AS6140	399	135	264	66.2%	IMPSAT-USA - ImpSat
AS14654	264	6	258	97.7%	WAYPORT - Wayport
AS9583	735	483	252	34.3%	SIFY-AS-IN Sify Limited
AS9443	374	123	251	67.1%	INTERNETPRIMUS-AS-AP Primus Telecommunications
AS7545	493	247	246	49.9%	TPG-INTERNET-AP TPG Internet Pty Ltd
AS1239	886	644	242	27.3%	SPRINTLINK - Sprint
AS15270	272	37	235	86.4%	AS-PAETEC-NET - PaeTec.net -a division of PaeTecCommunications, Inc.
AS23126	254	23	231	90.9%	KMCTELCOM-DIA - KMC Telecom, Inc.
AS4755	516	287	229	44.4%	VSNL-AS Videsh Sanchar Nigam Ltd. Autonomous System
AS7725	415	186	229	55.2%	CCH-AS7 - Comcast Cable Communications Holdings, Inc
AS6198	464	236	228	49.1%	BATI-MIA - BellSouth Network Solutions, Inc
AS5668	488	264	224	45.9%	AS-5668 - CenturyTel Internet Holdings, Inc.
AS2386	853	634	219	25.7%	INS-AS - AT&T Data Communications Services
AS9498	296	79	217	73.3%	BBIL-AP BHARTI BT INTERNET LTD.
AS11456	319	110	209	65.5%	NUVOX - NuVox Communications, Inc.
AS6167	264	67	197	74.6%	CELLCO-PART - Cellco Partnership
AS6517	319	128	191	59.9%	YIPESCOM - Yipes Communications, Inc.
Total	17193	6866	10327	60.1%	Top 30 total



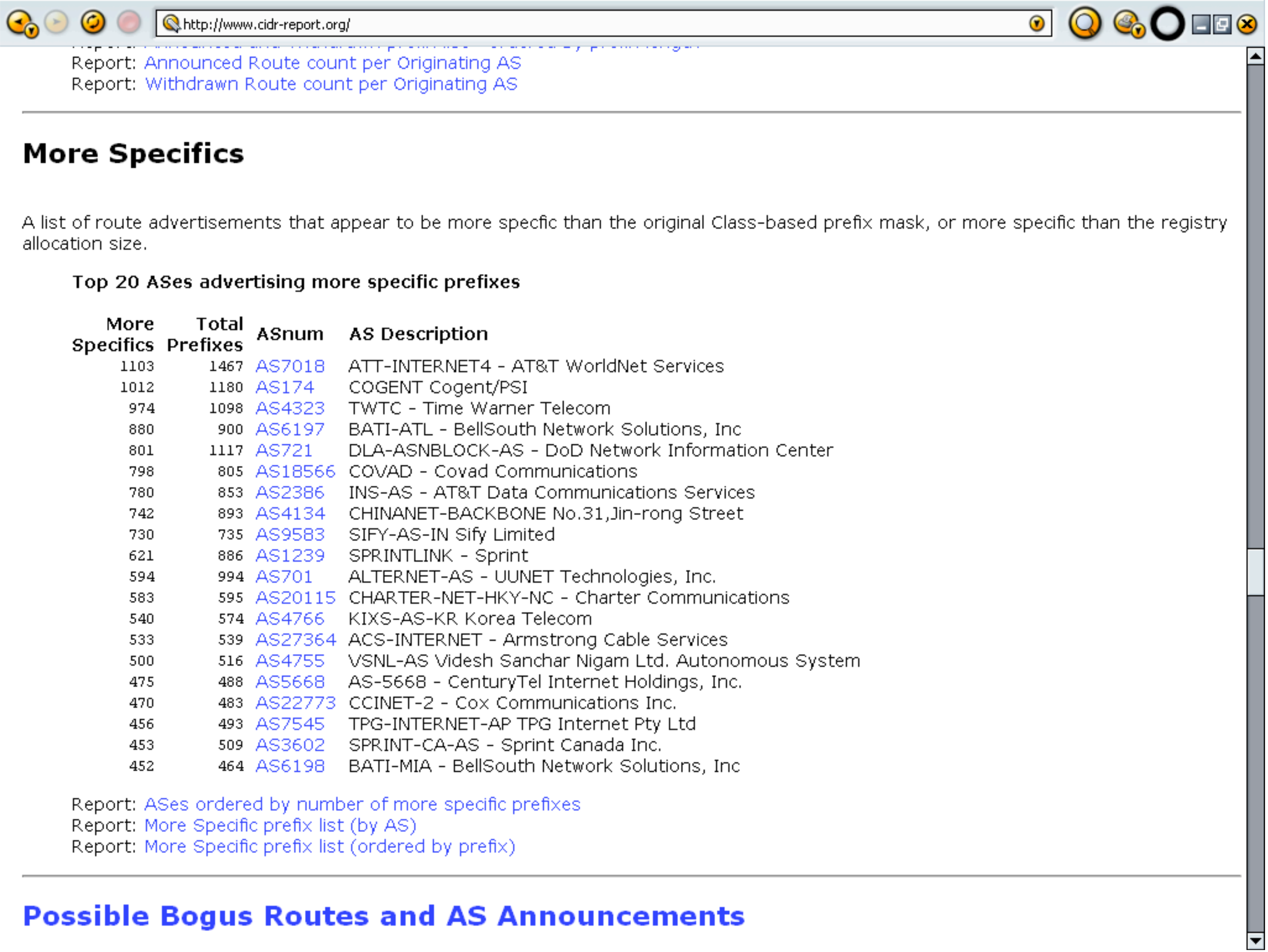
Top 20 Added Routes this week per Originating AS

Prefixes	ASnum	AS Description
154	AS7725	CCH-AS7 - Comcast Cable Communications Holdings, Inc
108	AS4755	VSNL-AS Videsh Sanchar Nigam Ltd. Autonomous System
52	AS35911	BNQ-1 - Telebec
36	AS13645	BROADBANDONE - BroadbandONE, Inc.
19	AS17488	HATHWAY-NET-AP Hathway IP Over Cable Internet
16	AS9576	SOOKMYUNG-AS SOOKMYUNG WOMEN'S UNIVERSITY
16	AS174	COGENT Cogent/PSI
16	AS18633	GIANTWEB - Giant Technologies Inc.
16	AS18042	KBT Koos Broadband Telecom
16	AS32613	IWEB-AS - Groupe iWeb Technologies inc.
15	AS19632	Metropolis Intercom
15	AS30340	AS-LLIX - Liberty Lake Internet Portal
13	AS19916	ASTRUM-0001 - OLM LLC
13	AS22047	VTR BANDA ANCHA S.A.
13	AS21882	PRIORITYNETWORKS - Priority Networks Inc.
12	AS9940	WOLCST-AS-AP World online AS, Cybersoft Technologies.
12	AS12715	JAZZNET Jazz Telecom S.A.
12	AS22927	Telefonica de Argentina
11	AS30533	CONNEXION-BY-BOEING-LTN - Connexion by Boeing
11	AS25454	TELEMEDIAAS Telemedia SA Autonomous System

Top 20 Withdrawn Routes this week per Originating AS

Prefixes	ASnum	AS Description
-45	AS10970	LH - Lighthouse Communications, Inc.
-33	AS7496	WEBCENTRAL-AS WebCentral
-31	AS8921	I-CONNEXION ICX Autonomous System
-23	AS4513	Globix Corporation
-20	AS1239	SPRINTLINK - Sprint
-18	AS14103	ACDNET-ASN1 - ACD.net
-17	AS29257	CBB-IE-AS Connexion by Boeing Ireland, Ltd.
-16	AS20115	CHARTER-NET-HKY-NC - Charter Communications
-16	AS6167	CELLCO-PART - Cellco Partnership
-15	AS17557	PKTELECOM-AS-AP Pakistan Telecom
-14	AS9152	MEGADAT Autonomous System
-14	AS16154	TELECOMS-AS Telecoms-Net Ltd.
-14	AS24219	NFI-AS-AP No Fuss Internet
-13	AS174	COGENT Cogent/PSI
-13	AS10125	DACCESS-AP DATA ACCESS INDIA LIMITED
-13	AS30857	TAURUS-AS Taurus Telecom PJSC
-12	AS17854	CABLELINE-AS-KR BANDOCABLELINE
-12	AS7049	S&M International S.A.
-12	AS4323	TWTC - Time Warner Telecom
-12	AS3561	SAVVIS - Savvis

Adds and Wdls per Prefix Length



Report: Announced Route count per Originating AS
Report: Withdrawn Route count per Originating AS

More Specifics

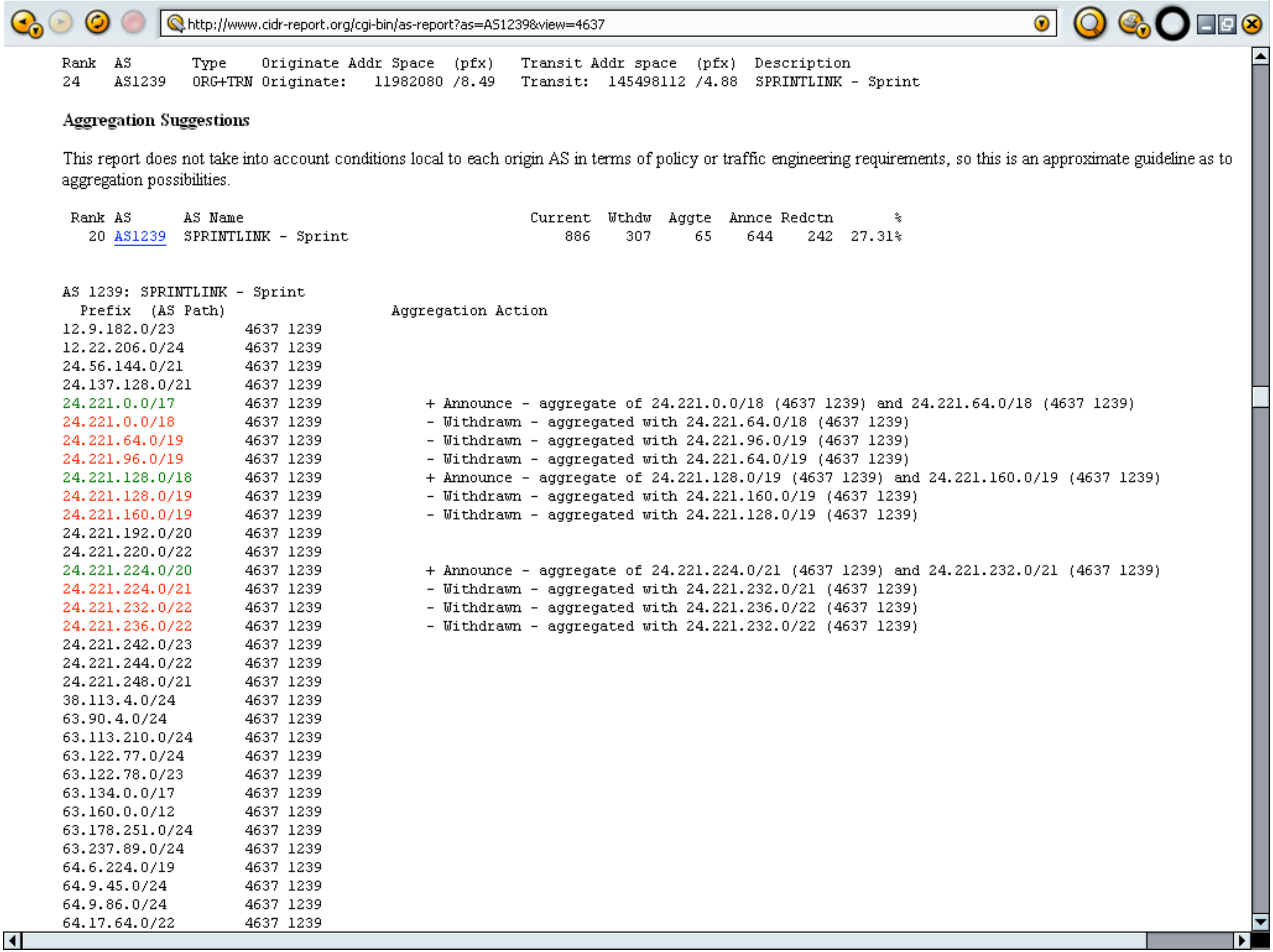
A list of route advertisements that appear to be more specfic than the original Class-based prefix mask, or more specific than the registry allocation size.

Top 20 ASes advertising more specific prefixes

More Specifics	Total Prefixes	ASnum	AS Description
1103	1467	AS7018	ATT-INTERNET4 - AT&T WorldNet Services
1012	1180	AS174	COGENT Cogent/PSI
974	1098	AS4323	TWTC - Time Warner Telecom
880	900	AS6197	BATI-ATL - BellSouth Network Solutions, Inc
801	1117	AS721	DLA-ASNBLOCK-AS - DoD Network Information Center
798	805	AS18566	COVAD - Covad Communications
780	853	AS2386	INS-AS - AT&T Data Communications Services
742	893	AS4134	CHINANET-BACKBONE No.31,Jin-rong Street
730	735	AS9583	SIFY-AS-IN Sify Limited
621	886	AS1239	SPRINTLINK - Sprint
594	994	AS701	ALTERNET-AS - UUNET Technologies, Inc.
583	595	AS20115	CHARTER-NET-HKY-NC - Charter Communications
540	574	AS4766	KIXS-AS-KR Korea Telecom
533	539	AS27364	ACS-INTERNET - Armstrong Cable Services
500	516	AS4755	VSNL-AS Videsh Sanchar Nigam Ltd. Autonomous System
475	488	AS5668	AS-5668 - CenturyTel Internet Holdings, Inc.
470	483	AS22773	CCINET-2 - Cox Communications Inc.
456	493	AS7545	TPG-INTERNET-AP TPG Internet Pty Ltd
453	509	AS3602	SPRINT-CA-AS - Sprint Canada Inc.
452	464	AS6198	BATI-MIA - BellSouth Network Solutions, Inc

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

Possible Bogus Routes and AS Announcements



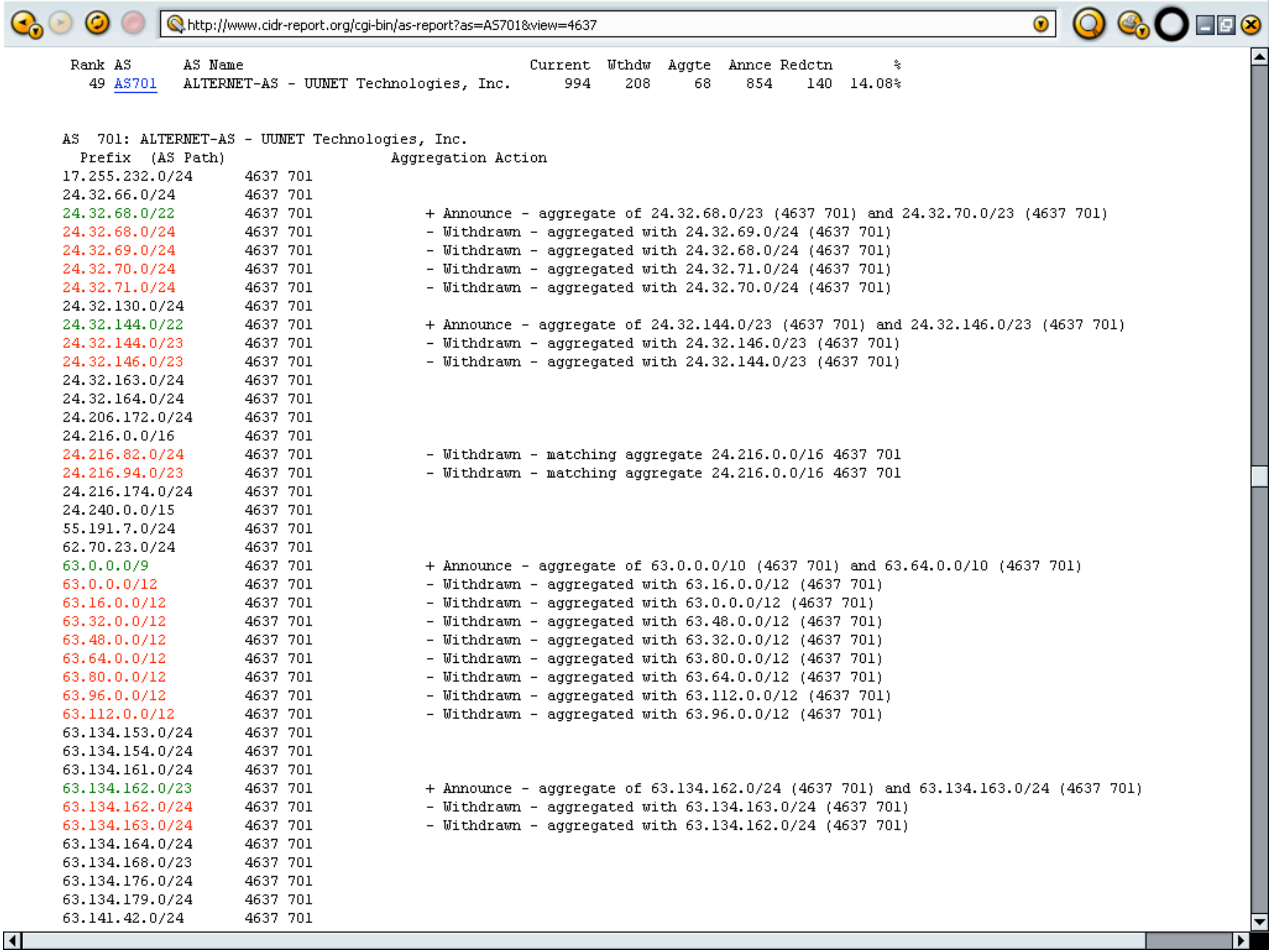
http://www.cidr-report.org/cgi-bin/as-report?as=AS1239&view=4637

```
Rank  AS      Type    Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
24    AS1239  ORG+TRN Originate:   11982080 /8.49   Transit:  145498112 /4.88  SPRINTLINK - Sprint
```

Aggregation Suggestions

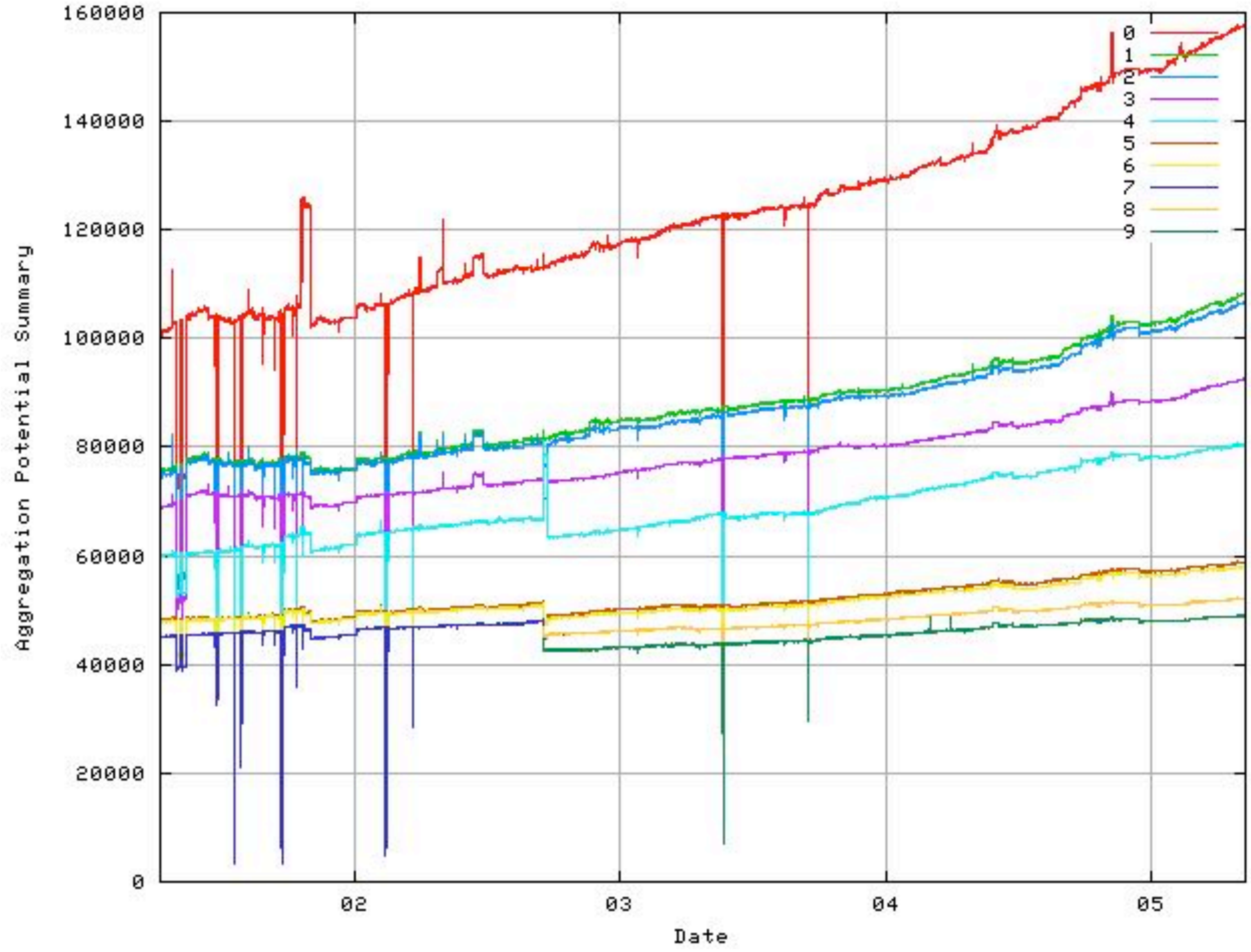
This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

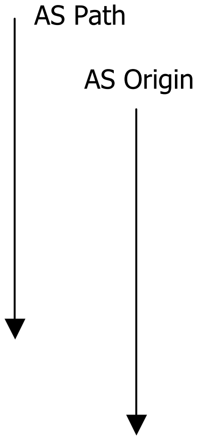
```
 Rank AS      AS Name                              Current  Wthdw  Aggte  Annce Redctn      %
   20 AS1239  SPRINTLINK - Sprint                      886    307     65    644    242  27.31%


AS 1239: SPRINTLINK - Sprint
  Prefix  (AS Path)                 Aggregation Action
12.9.182.0/23        4637 1239
12.22.206.0/24       4637 1239
24.56.144.0/21       4637 1239
24.137.128.0/21      4637 1239
24.221.0.0/17        4637 1239          + Announce - aggregate of 24.221.0.0/18 (4637 1239) and 24.221.64.0/18 (4637 1239)
24.221.0.0/18        4637 1239          - Withdrawn - aggregated with 24.221.64.0/18 (4637 1239)
24.221.64.0/19       4637 1239          - Withdrawn - aggregated with 24.221.96.0/19 (4637 1239)
24.221.96.0/19       4637 1239          - Withdrawn - aggregated with 24.221.64.0/19 (4637 1239)
24.221.128.0/18      4637 1239          + Announce - aggregate of 24.221.128.0/19 (4637 1239) and 24.221.160.0/19 (4637 1239)
24.221.128.0/19      4637 1239          - Withdrawn - aggregated with 24.221.160.0/19 (4637 1239)
24.221.160.0/19      4637 1239          - Withdrawn - aggregated with 24.221.128.0/19 (4637 1239)
24.221.192.0/20      4637 1239
24.221.220.0/22      4637 1239
24.221.224.0/20      4637 1239          + Announce - aggregate of 24.221.224.0/21 (4637 1239) and 24.221.232.0/21 (4637 1239)
24.221.224.0/21      4637 1239          - Withdrawn - aggregated with 24.221.232.0/21 (4637 1239)
24.221.232.0/22      4637 1239          - Withdrawn - aggregated with 24.221.236.0/22 (4637 1239)
24.221.236.0/22      4637 1239          - Withdrawn - aggregated with 24.221.232.0/22 (4637 1239)
24.221.242.0/23      4637 1239
24.221.244.0/22      4637 1239
24.221.248.0/21      4637 1239
38.113.4.0/24        4637 1239
63.90.4.0/24         4637 1239
63.113.210.0/24      4637 1239
63.122.77.0/24       4637 1239
63.122.78.0/23       4637 1239
63.134.0.0/17        4637 1239
63.160.0.0/12        4637 1239
63.178.251.0/24      4637 1239
63.237.89.0/24       4637 1239
64.6.224.0/19        4637 1239
64.9.45.0/24         4637 1239
64.9.86.0/24         4637 1239
64.17.64.0/22        4637 1239
```


http://www.cidr-report.org/cgi-bin/as-report?as=AS701&view=4637

```
 Rank AS      AS Name                                Current  Wthdw  Aggte  Annce Redctn     %
   49 AS701   ALTERNET-AS - UUNET Technologies, Inc.     994    208     68    854    140  14.08%


AS  701: ALTERNET-AS - UUNET Technologies, Inc.
  Prefix  (AS Path)                  Aggregation Action
17.255.232.0/24      4637 701
24.32.66.0/24        4637 701
24.32.68.0/22        4637 701           + Announce - aggregate of 24.32.68.0/23 (4637 701) and 24.32.70.0/23 (4637 701)
24.32.68.0/24        4637 701           - Withdrawn - aggregated with 24.32.69.0/24 (4637 701)
24.32.69.0/24        4637 701           - Withdrawn - aggregated with 24.32.68.0/24 (4637 701)
24.32.70.0/24        4637 701           - Withdrawn - aggregated with 24.32.71.0/24 (4637 701)
24.32.71.0/24        4637 701           - Withdrawn - aggregated with 24.32.70.0/24 (4637 701)
24.32.130.0/24       4637 701
24.32.144.0/22       4637 701           + Announce - aggregate of 24.32.144.0/23 (4637 701) and 24.32.146.0/23 (4637 701)
24.32.144.0/23       4637 701           - Withdrawn - aggregated with 24.32.146.0/23 (4637 701)
24.32.146.0/23       4637 701           - Withdrawn - aggregated with 24.32.144.0/23 (4637 701)
24.32.163.0/24       4637 701
24.32.164.0/24       4637 701
24.206.172.0/24      4637 701
24.216.0.0/16        4637 701
24.216.82.0/24       4637 701           - Withdrawn - matching aggregate 24.216.0.0/16 4637 701
24.216.94.0/23       4637 701           - Withdrawn - matching aggregate 24.216.0.0/16 4637 701
24.216.174.0/24      4637 701
24.240.0.0/15        4637 701
55.191.7.0/24        4637 701
62.70.23.0/24        4637 701
63.0.0.0/9           4637 701           + Announce - aggregate of 63.0.0.0/10 (4637 701) and 63.64.0.0/10 (4637 701)
63.0.0.0/12          4637 701           - Withdrawn - aggregated with 63.16.0.0/12 (4637 701)
63.16.0.0/12         4637 701           - Withdrawn - aggregated with 63.0.0.0/12 (4637 701)
63.32.0.0/12         4637 701           - Withdrawn - aggregated with 63.48.0.0/12 (4637 701)
63.48.0.0/12         4637 701           - Withdrawn - aggregated with 63.32.0.0/12 (4637 701)
63.64.0.0/12         4637 701           - Withdrawn - aggregated with 63.80.0.0/12 (4637 701)
63.80.0.0/12         4637 701           - Withdrawn - aggregated with 63.64.0.0/12 (4637 701)
63.96.0.0/12         4637 701           - Withdrawn - aggregated with 63.112.0.0/12 (4637 701)
63.112.0.0/12        4637 701           - Withdrawn - aggregated with 63.96.0.0/12 (4637 701)
63.134.153.0/24      4637 701
63.134.154.0/24      4637 701
63.134.161.0/24      4637 701
63.134.162.0/23      4637 701           + Announce - aggregate of 63.134.162.0/24 (4637 701) and 63.134.163.0/24 (4637 701)
63.134.162.0/24      4637 701           - Withdrawn - aggregated with 63.134.163.0/24 (4637 701)
63.134.163.0/24      4637 701           - Withdrawn - aggregated with 63.134.162.0/24 (4637 701)
63.134.164.0/24      4637 701
63.134.168.0/23      4637 701
63.134.176.0/24      4637 701
63.134.179.0/24      4637 701
63.141.42.0/24       4637 701
```

Aggregation Potential (source: bgp.potaroo.net/as4637/)





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 in the four RIR databases to see if this address space really has been assigned to the customer

  The tool: whois –h whois.apnic.net x.x.x.0/24

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 Rob Thomas’ list of “bogons”

  http://www.cymru.org/Documents/bogon-list.html

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
Example One

- 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 IGP: OSPF or ISIS ☺
- Assign loopback interfaces and /32 addresses 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

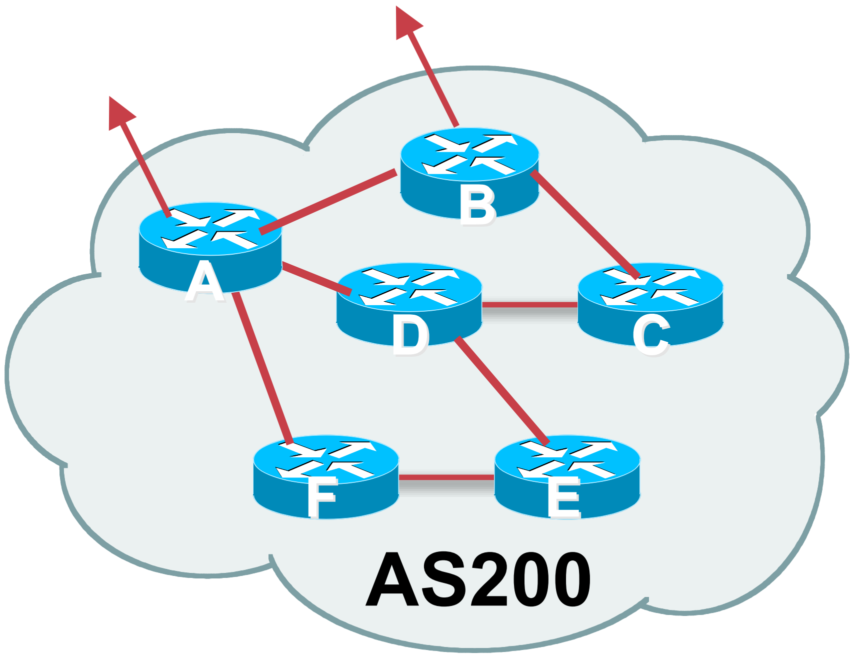
- **Second step is to configure the local network to use iBGP**
- **iBGP can run on**

  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***

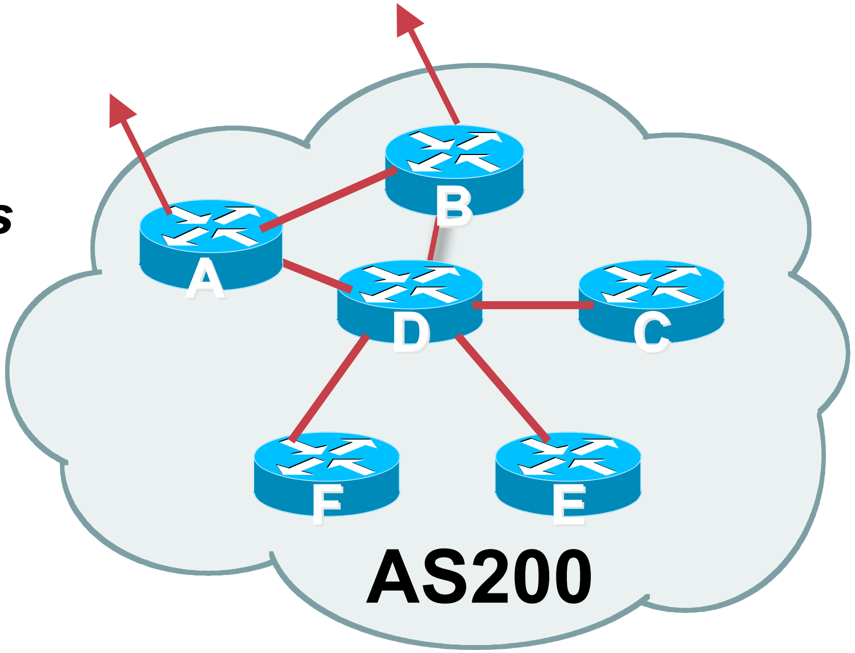


Preparing the Network
Second Step: iBGP (Transit Path)

- ***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
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

CISCO SYSTEMS

Configuration Tips

Of templates, passwords, tricks, and more 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!

Next-hop-self

- 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

  BGP speaker announces external network using local address (loopback) as next-hop

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!
- Use passwords on iBGP session

  Not being paranoid, VERY necessary

eBGP Template
Example

- BGP damping
  - Do NOT use it unless you understand why
  - Use RIPE-229 parameters, or something even weaker
  - 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

- 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!!

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   3FFE:C00:8023:5::2   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

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

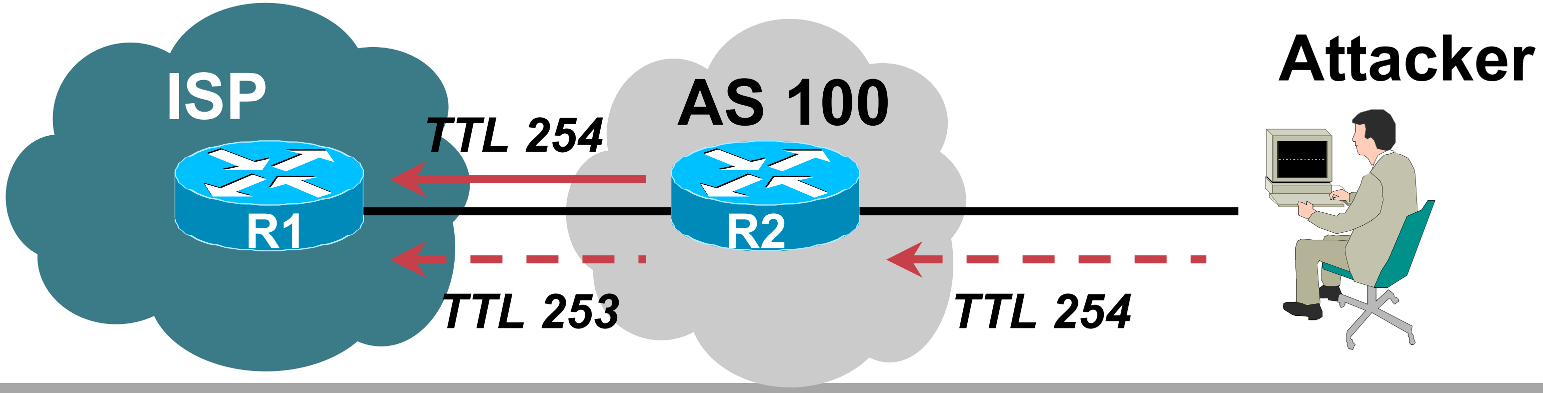
BGP TTL “hack”

- **Implement RFC3682 on BGP peerings**

  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

Passwords on BGP sessions

- *Yes, I am mentioning passwords again*
- Put password on the BGP session
  - 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”

Using Communities

- Use communities to:

    Scale iBGP management

    Ease iBGP management

- Come up with a strategy for different classes of customers

    Which prefixes stay inside network

    Which prefixes are announced by eBGP

    …etc…

Using Communities: Strategy

- **BGP customers**
  - **Offer max 3 types of feeds (easier than custom configuration per peer)**
  - **Use communities**
- **Static customers**
  - **Use communities**
- **Differentiate between different types of prefixes**
  - **Makes eBGP filtering easy**

Using Communities: BGP Customer Aggregation Guidelines

- Define at least three groups of peers:

  cust-default—send default route only

  cust-cust—send customer routes only

  cust-full —send full Internet routes

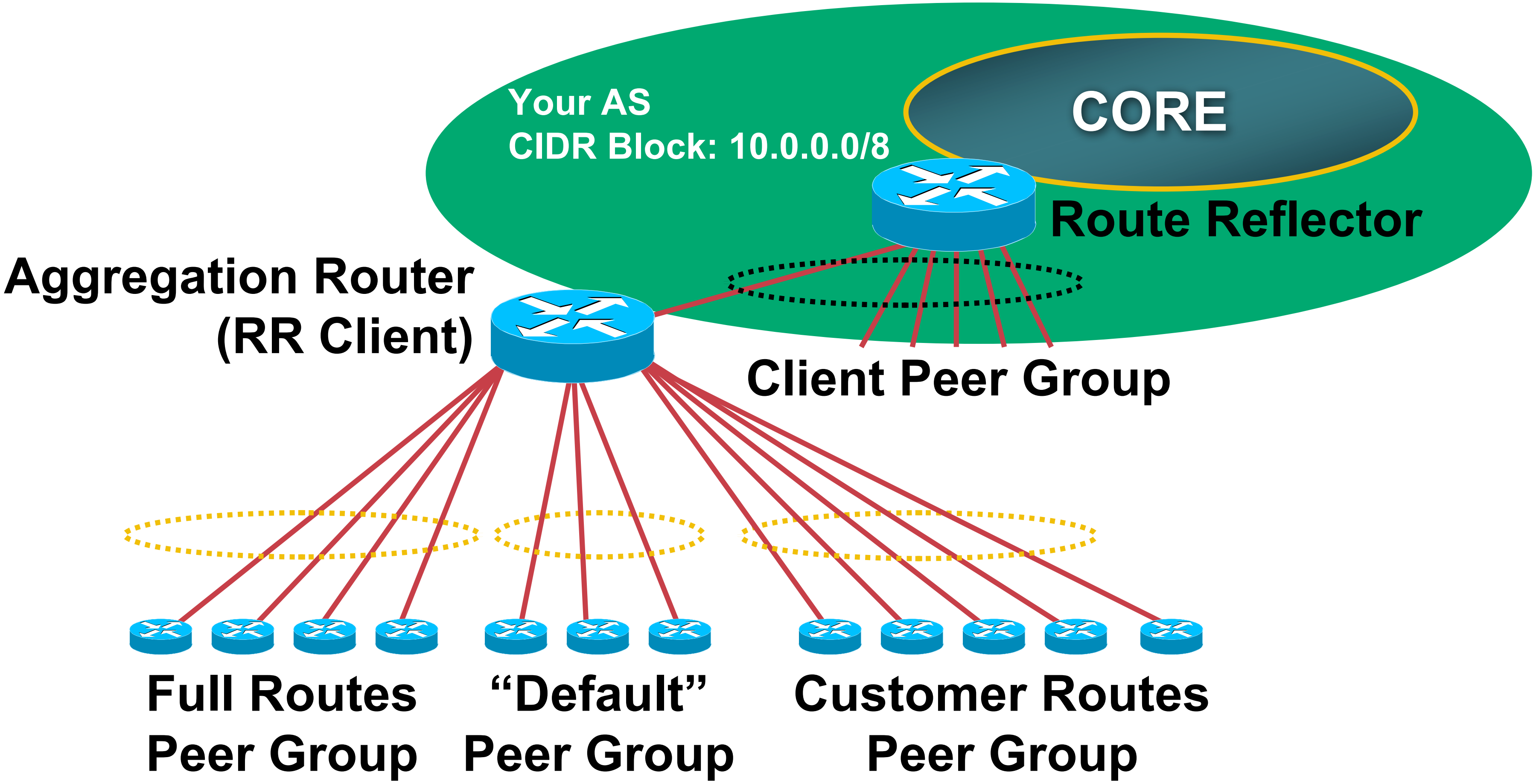
- Identify routes via communities e.g.

  100:4100=customers; 100:4500=peers

- Apply passwords per neighbour

- Apply inbound & outbound prefix filters per neighbour

BGP Customer Aggregation



Apply passwords and in/outbound prefix-list directly to each neighbour

Using Communities: Static Customer Aggregation Guidelines

- Identify routes via communities, e.g.

  100:4000 = my address blocks

  100:4100 = “specials” from my blocks

  100:4200 = customers from my blocks

  100:4300 = customers outside my blocks

  Helps with aggregation, iBGP, filtering

- Set correct community as networks are installed in BGP on aggregation routers

Using Communities: Sample core configuration

- **eBGP peers and upstreams**

  Send communities 100:4000, 100:4100 and 100:4300, receive everything

- **iBGP full routes**

  Send everything (only to network core)

- **iBGP partial routes**

  Send communities 100:4000, 100:4100, 100:4200, 100:4300 and 100:4500 (to edge routers, peering routers, IXP routers)

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

Deploying BGP
Next: Multihoming

Philip Smith <pfs@cisco.com>

NZNOG 2006

22-24 Mar 2006

Wellington