BGP Techniques for Internet Service Providers

Philip Smith <pfs@cisco.com>

NANOG 31

San Francisco

23-25 May 2004

Presentation Slides

- **Slides are at:**

  ftp://ftp-eng.cisco.com /pfs/seminars/NANOG31-BGP-Techniques.pdf

  And on the NANOG meeting website

- **Feel free to ask questions any time**

BGP for Internet Service Providers

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

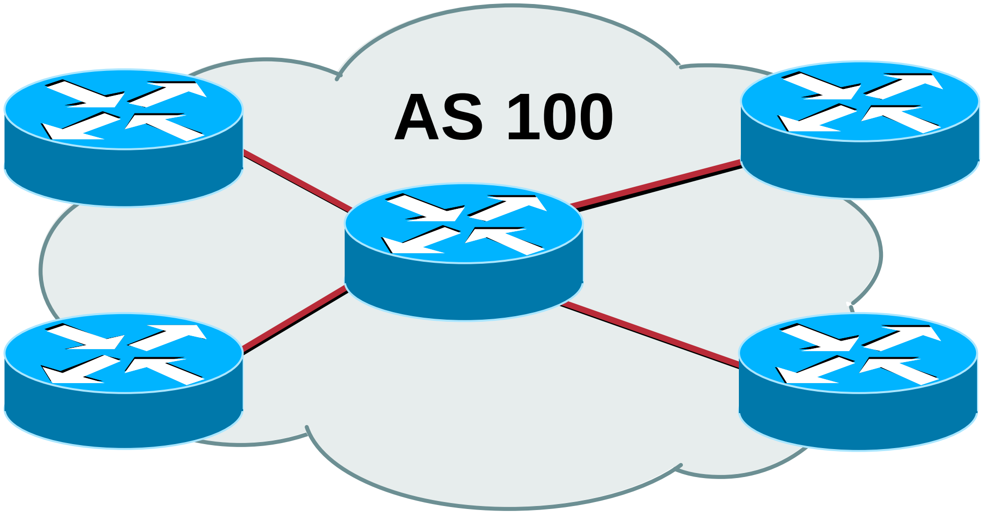
BGP Basics

What is this BGP thing?

Border Gateway Protocol

- **Routing Protocol used to exchange routing information between networks**
  - **exterior gateway protocol**
- **Described in RFC1771**
  - **work in progress to update**
  - **www.ietf.org/internet-drafts/draft-ietf-idr-bgp4-23.txt**
- **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 for public network use

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

  0 and 65535 are reserved

- **32 bit ASNs are coming soon**

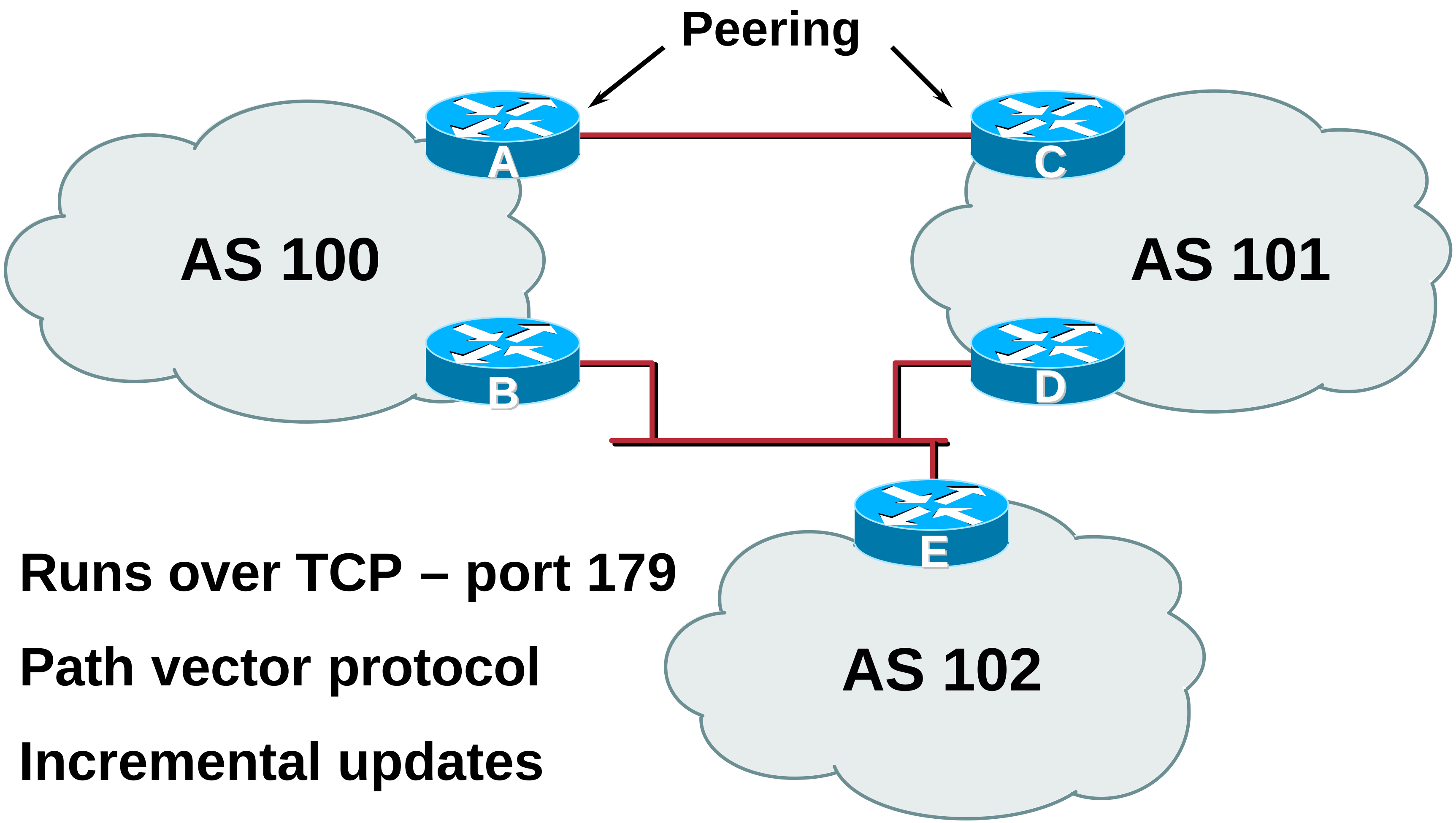
  www.ietf.org/internet-drafts/draft-ietf-idr-as4bytes-08.txt

  With ASN 23456 reserved for the transition

Autonomous System Number (ASN)

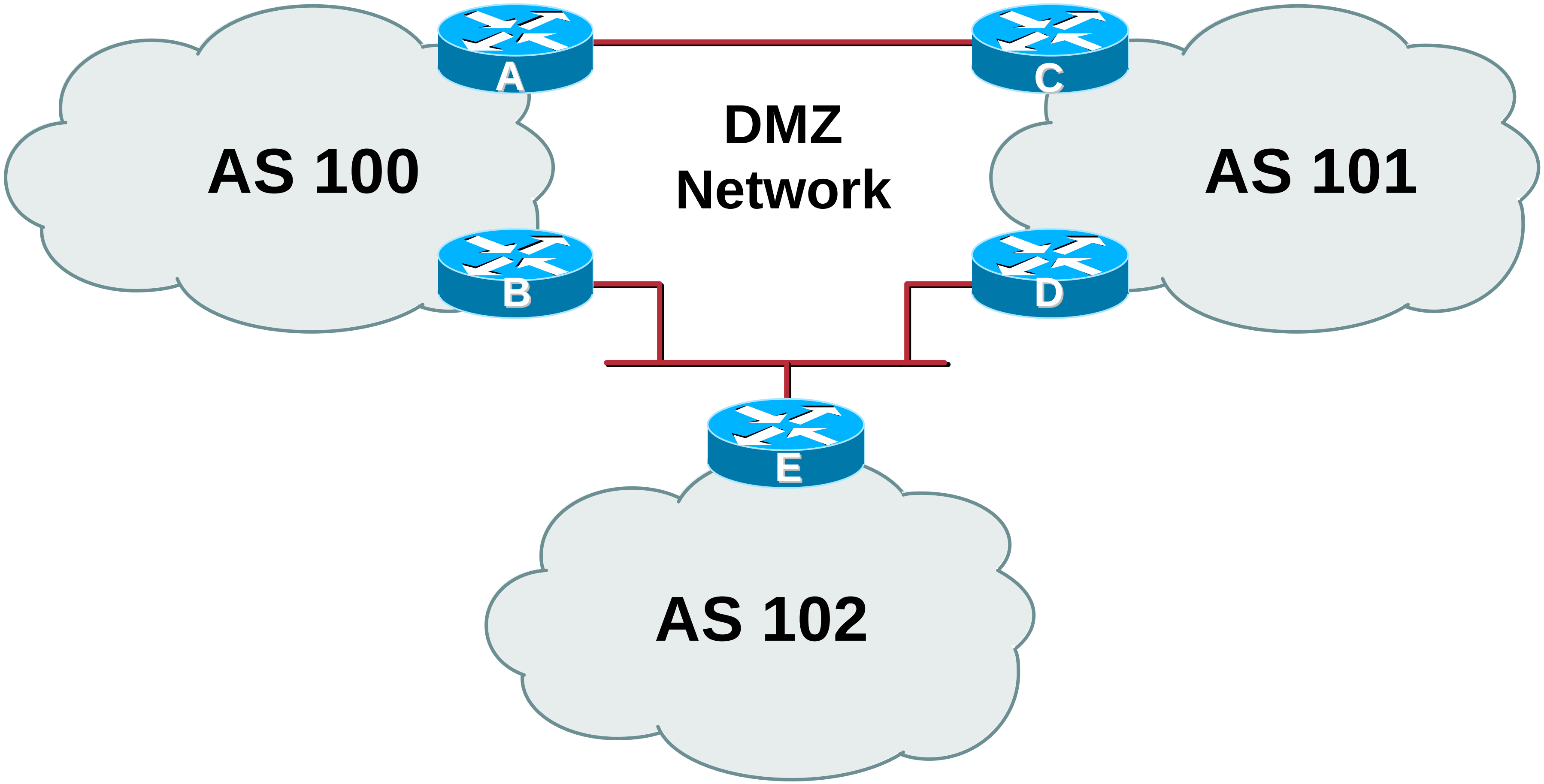
- **ASNs are distributed by the Regional Internet Registries**
  - **Also available from upstream ISPs who are members of one of the RIRs**
  - **Current ASN allocations up to 32767 have been made to the RIRs**
  - **Of these, around 17000 are visible on the Internet**

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

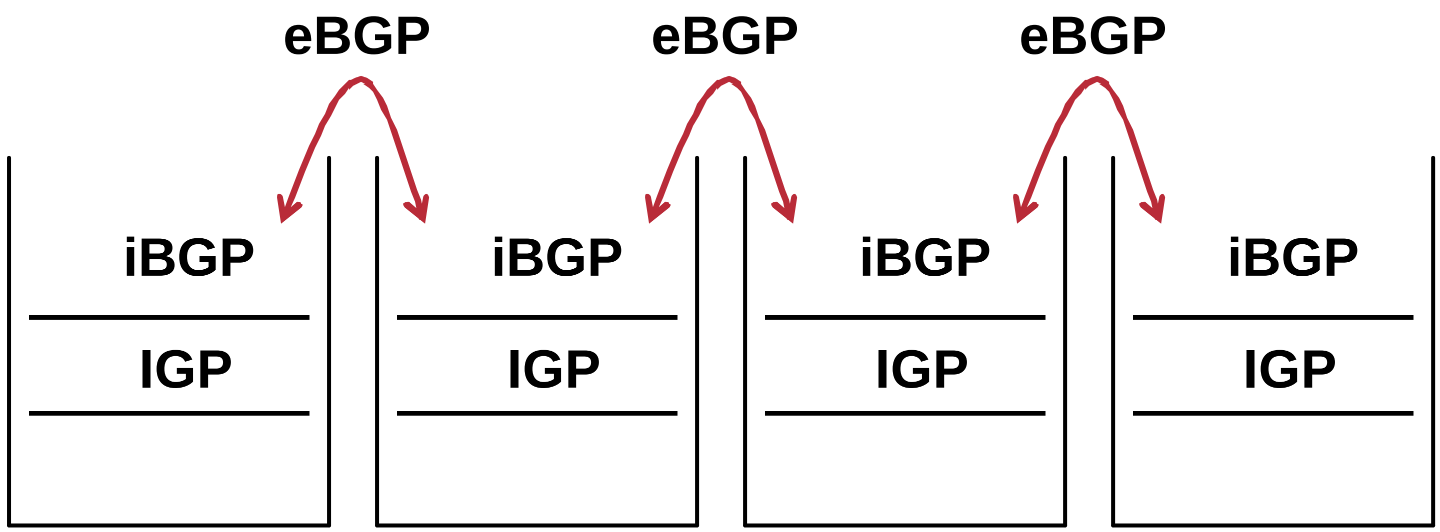
- Learns multiple paths via internal and external BGP speakers
- Picks the best path and installs in the forwarding table
- Best path is sent to external BGP neighbours
- Policies applied by influencing the best path selection

eBGP & iBGP

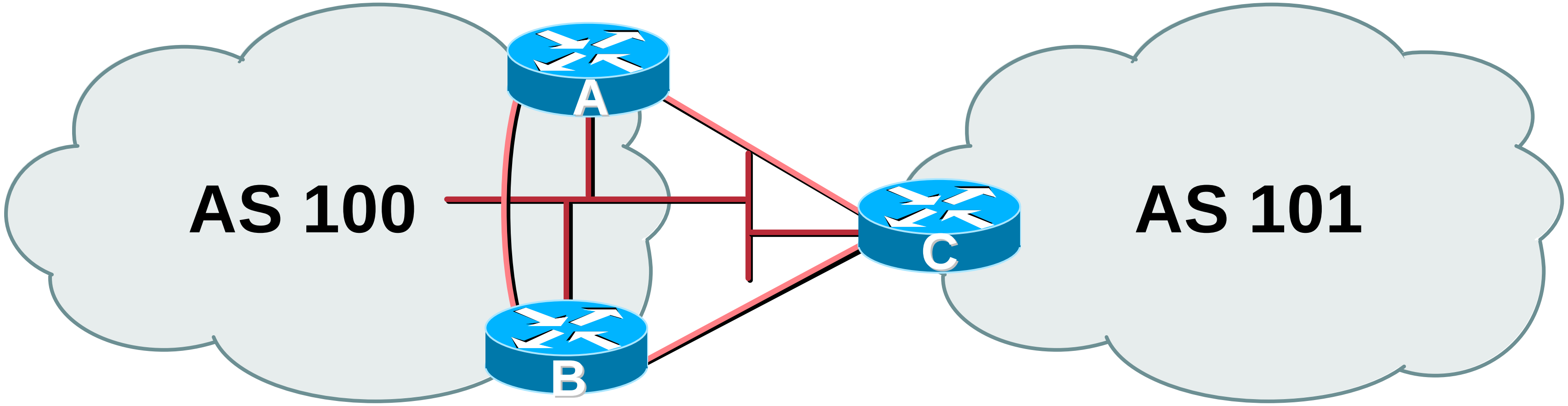
- BGP used internally (iBGP) and externally (eBGP)
- iBGP used to carry
    - some/all Internet prefixes across ISP backbone
    - ISP's customer prefixes
- eBGP used to
    - exchange prefixes with other ASes
    - implement routing policy

BGP/IGP model used in ISP networks

- **Model representation**



External BGP Peering (eBGP)

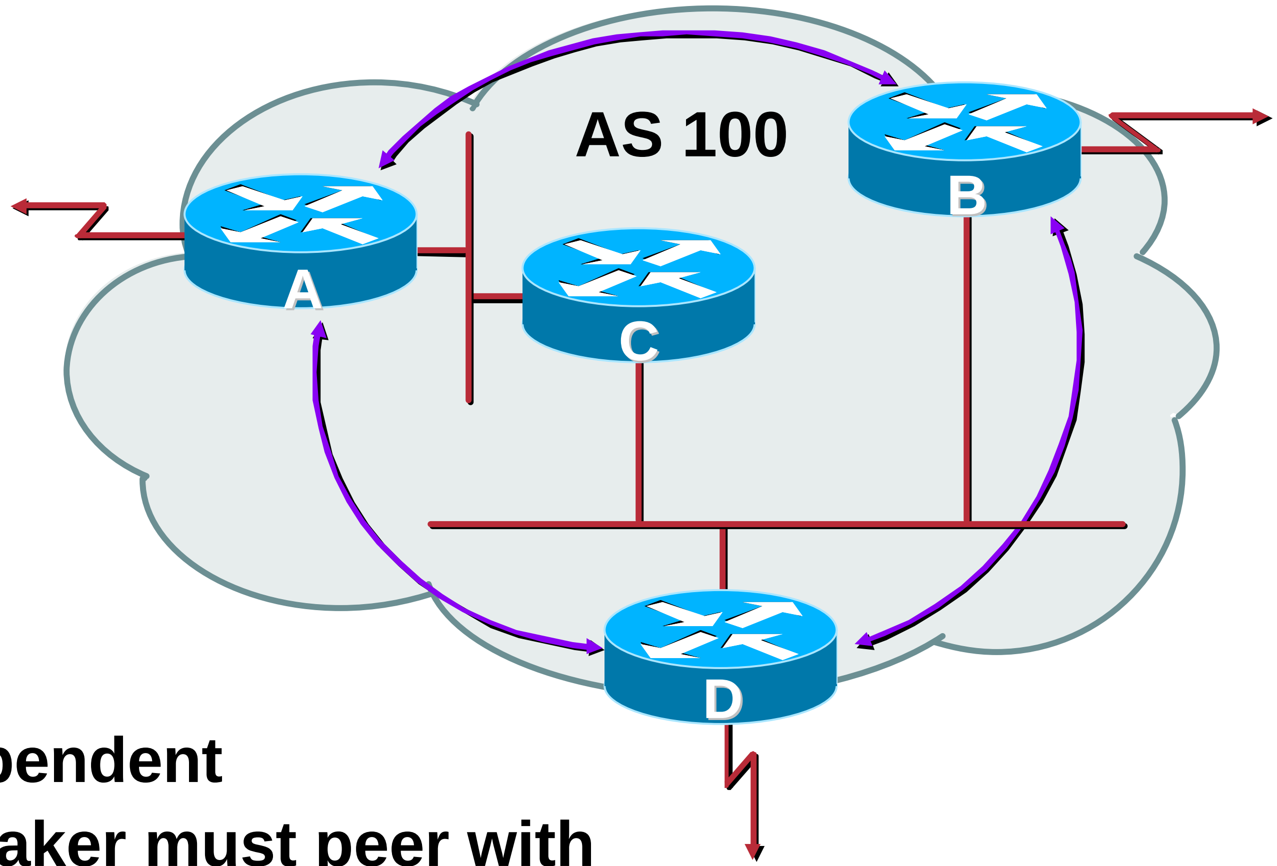


- **Between BGP speakers in different AS**
- **Should be directly connected**
- **Never run an IGP between eBGP peers**

Internal BGP (iBGP)

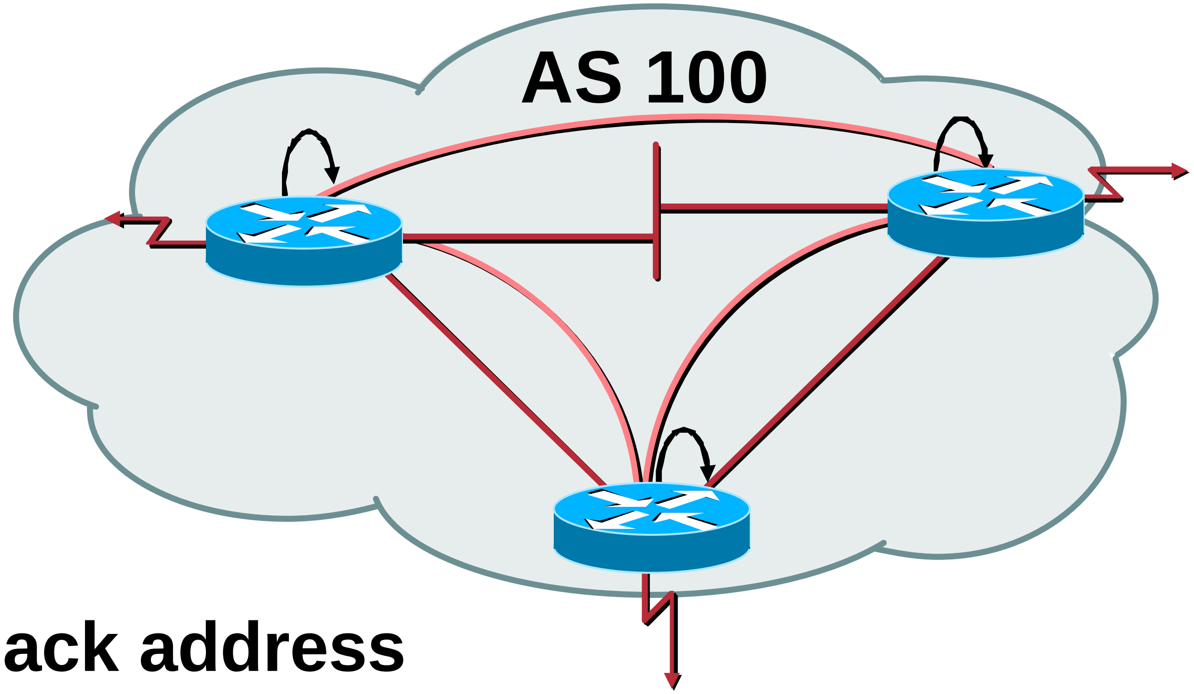
- BGP peer within the same AS
- Not required to be directly connected
  - IGP takes care of inter-BGP speaker connectivity
- iBGP speakers need to be fully meshed
  - they originate connected networks
  - they do not pass on prefixes learned from other iBGP speakers

Internal BGP Peering (iBGP)



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

Peering to loopback addresses



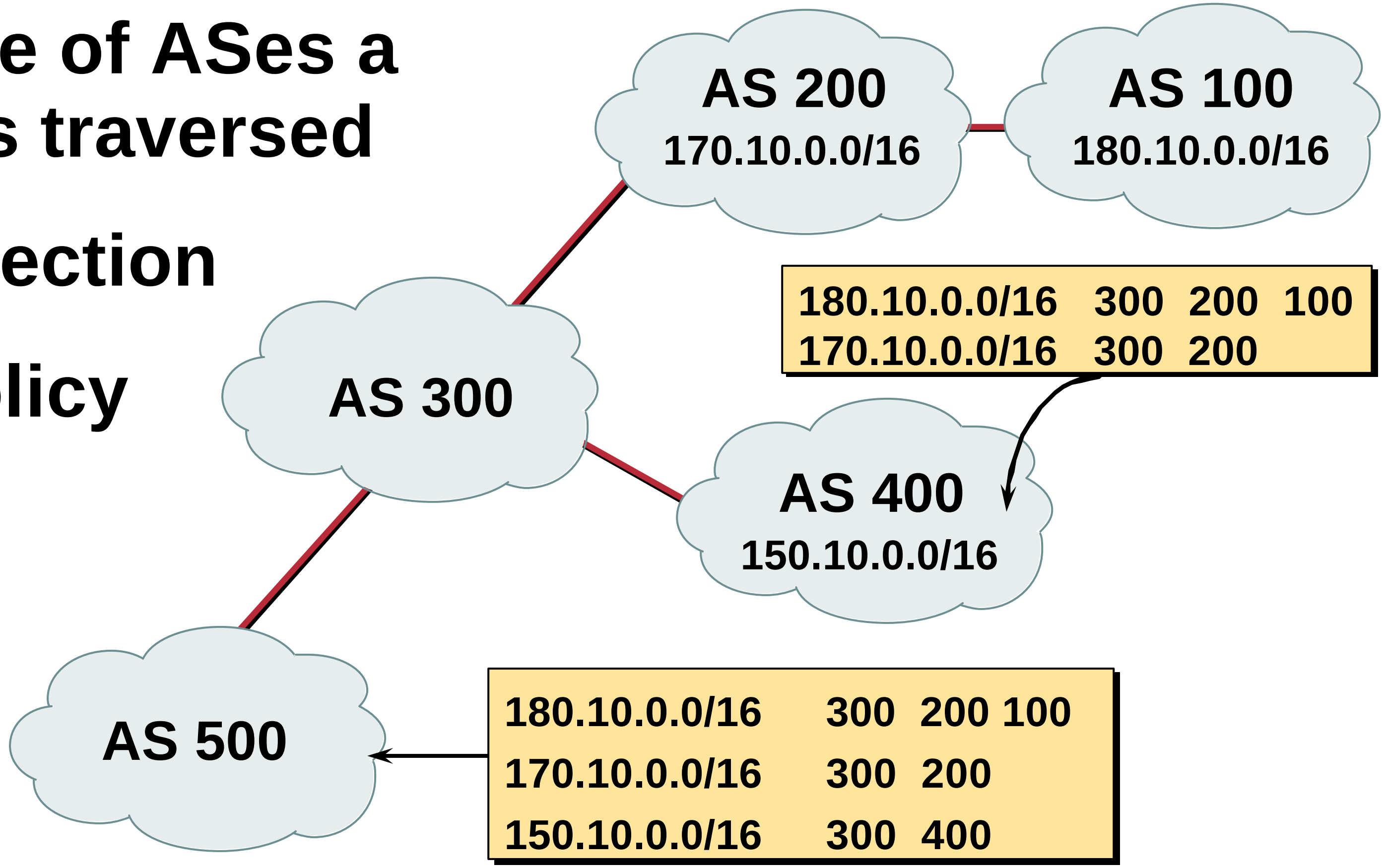
- **Peer with loop-back address**
  - **Loop-back interface does not go down – ever!**
- **iBGP session is not dependent on**
  - **State of a single interface**
  - **Physical topology**

BGP Attributes

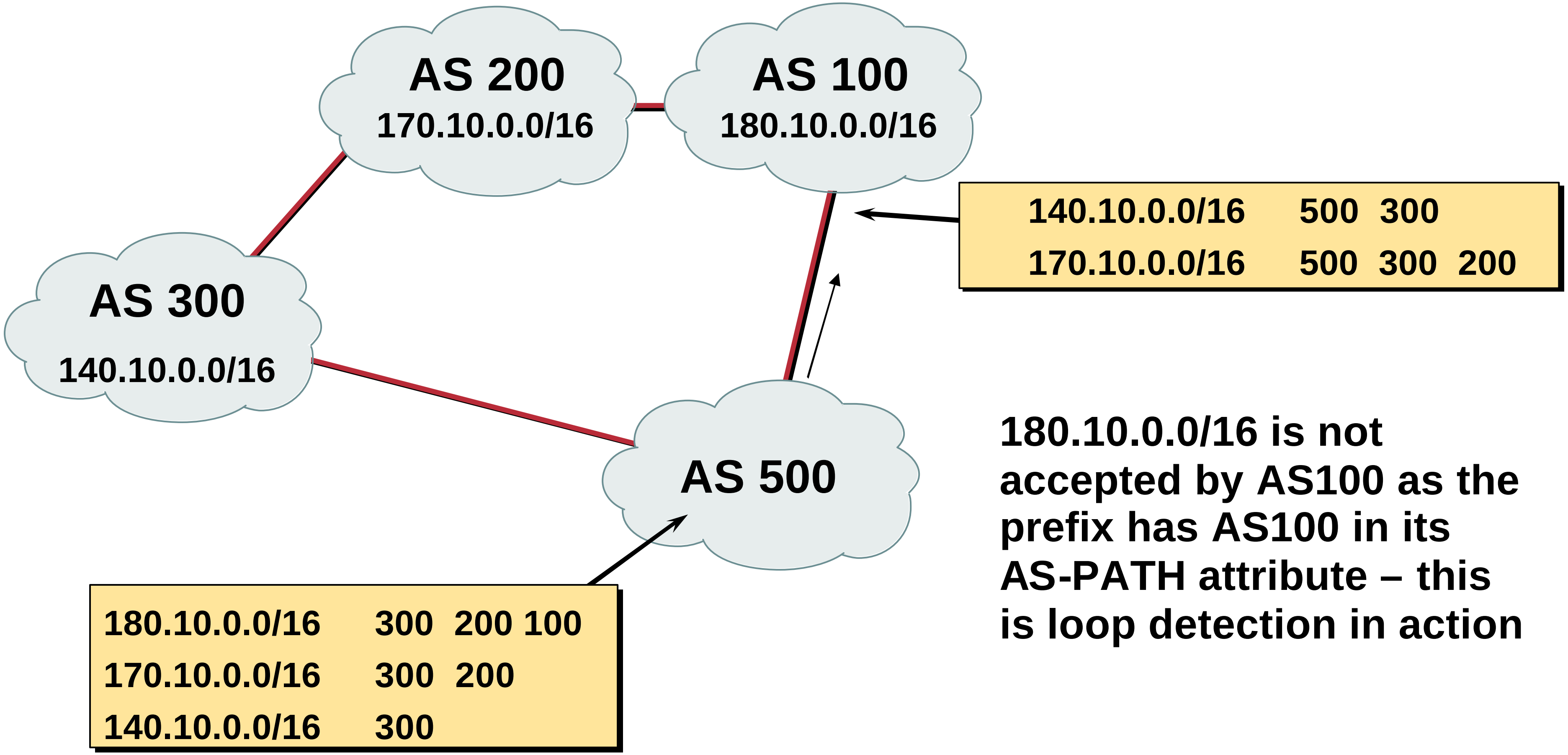
Information about BGP

AS-Path

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

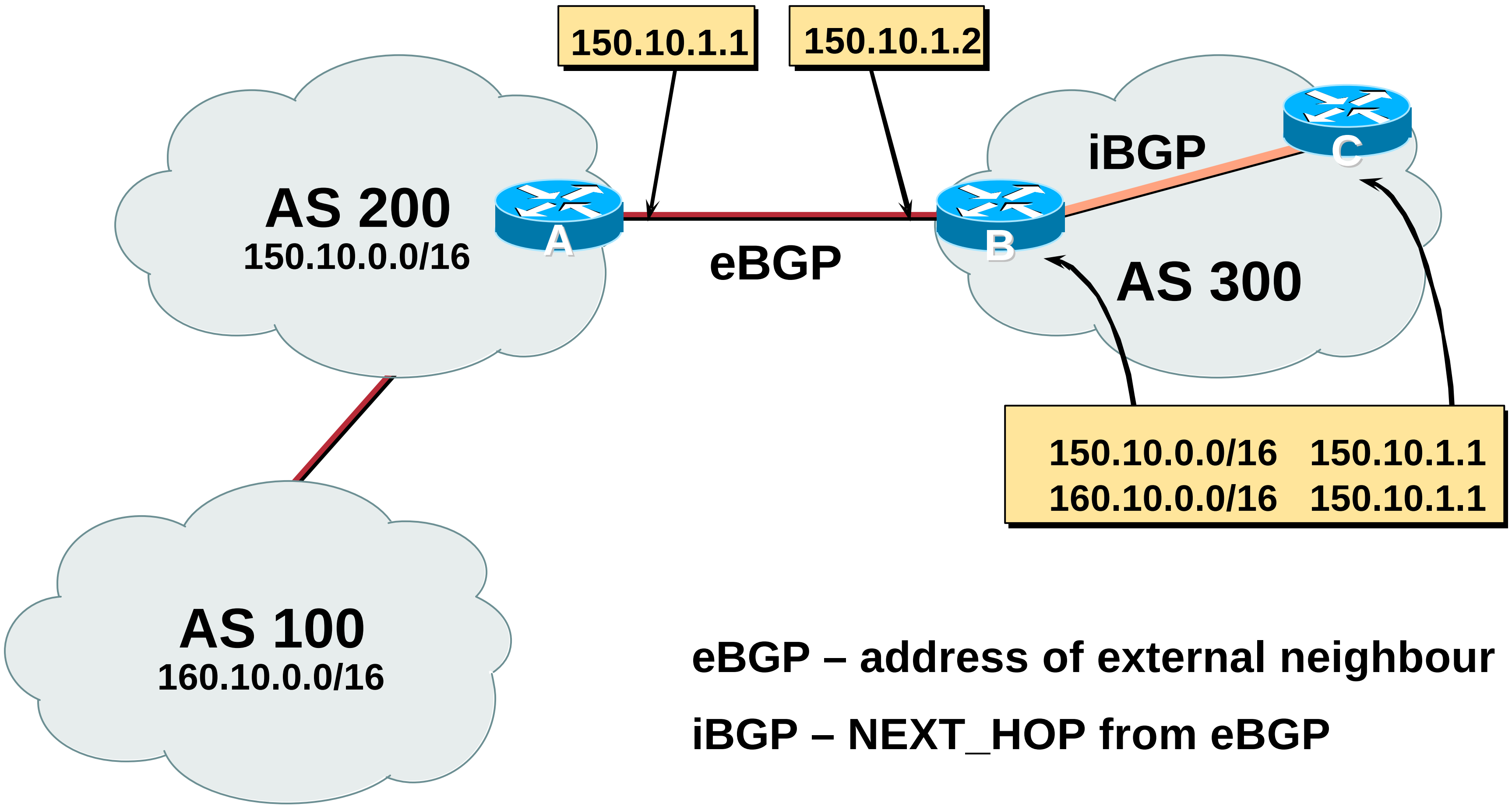


AS-Path loop detection



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

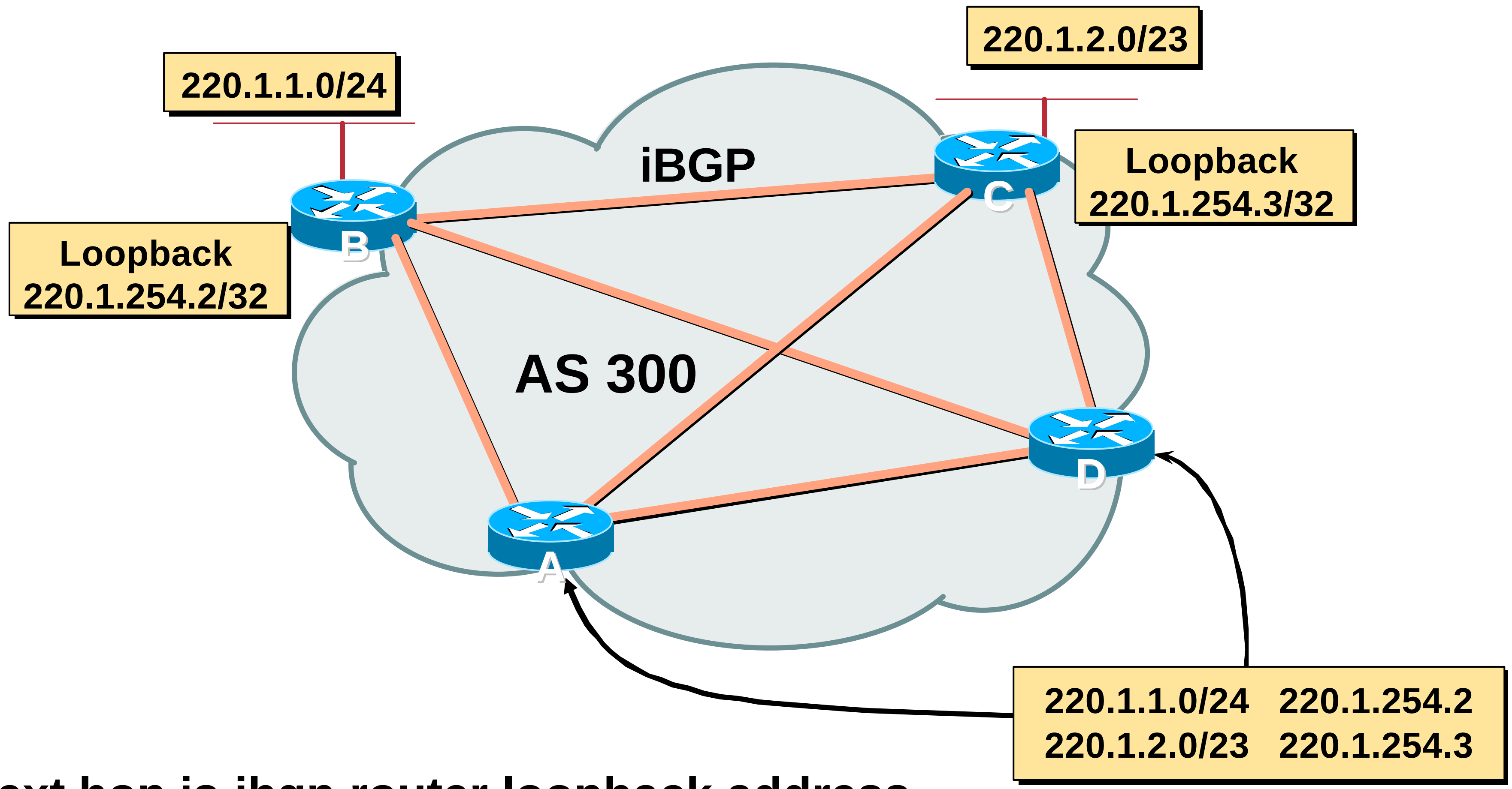
Next Hop



eBGP – address of external neighbour

iBGP – NEXT_HOP from eBGP

iBGP Next Hop



Next hop is ibgp router loopback address

Recursive route look-up

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

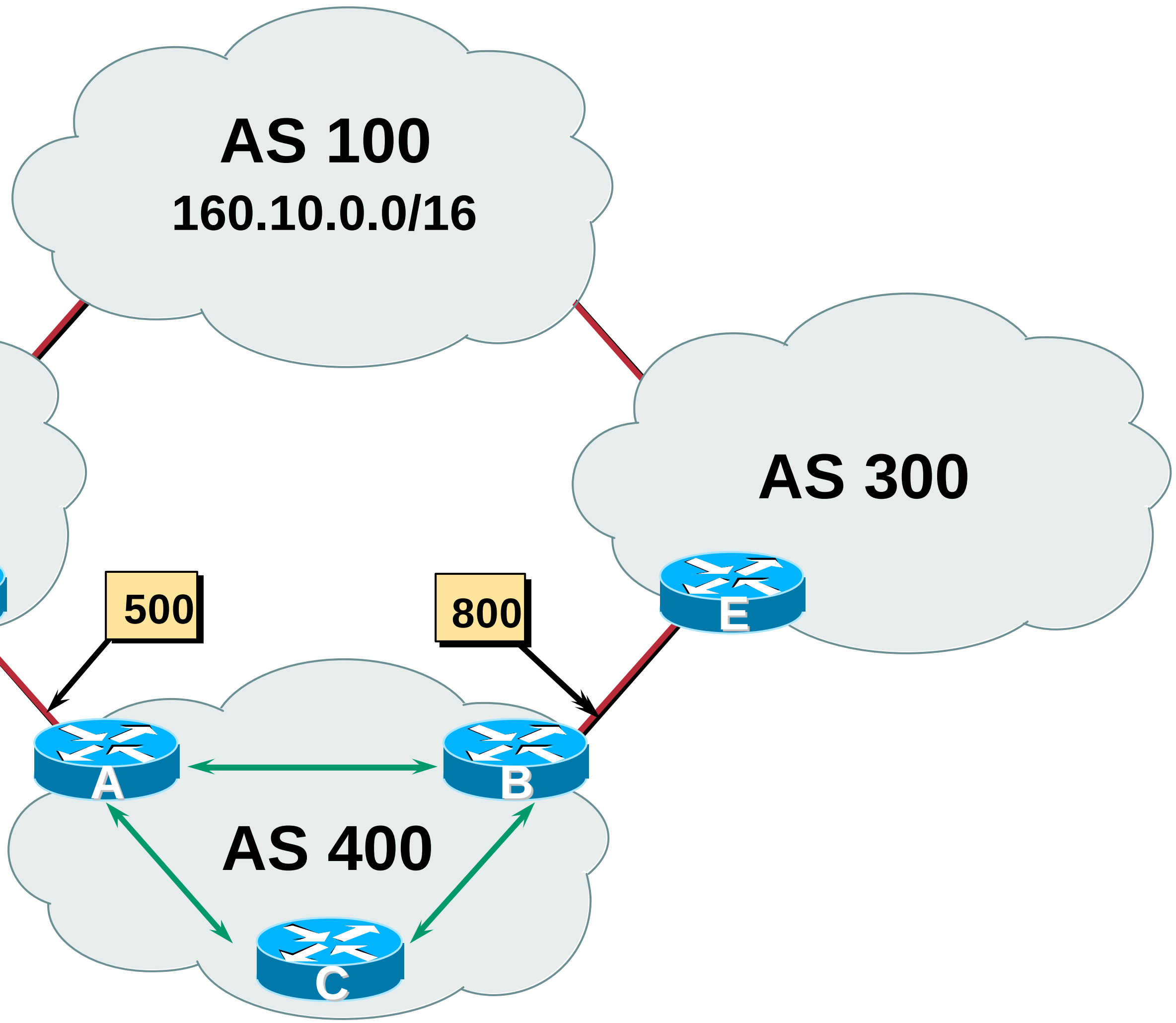
  EGP – generated by EGP

  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



	160.10.0.0/16	500
>	160.10.0.0/16	800

Local Preference

- **Local to an AS – non-transitive**

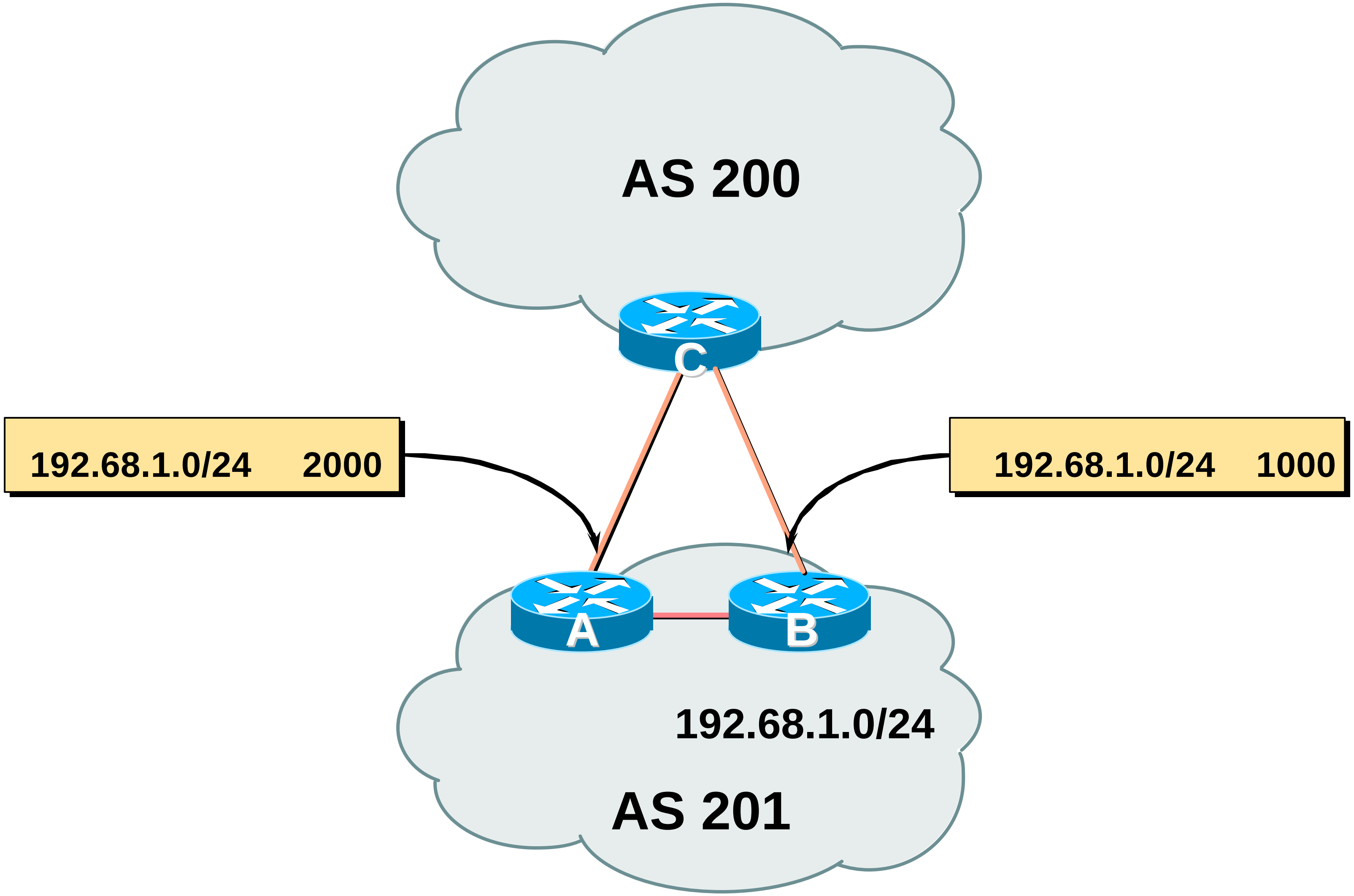
  Default local preference is 100 in most implementations

- **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

  Some implementations have option to relax this rule
- Path with lowest MED wins
- Absence of MED attribute implies MED value of zero (draft-ietf-idr-bgp4-23.txt)

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

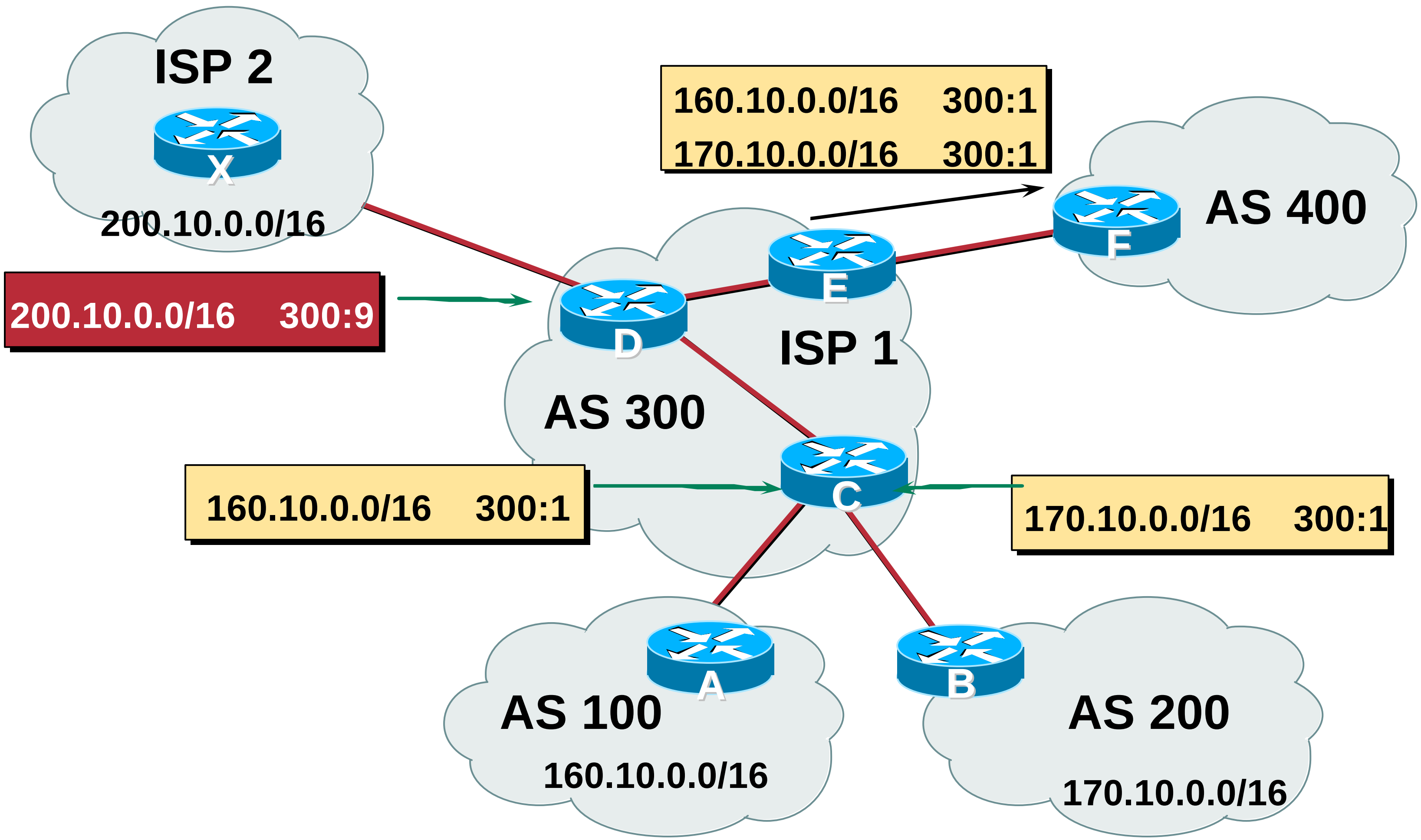
- Potential for "metric confusion"

  www.ietf.org/internet-drafts/draft-ietf-grow-bgp-med-considerations-01.txt

Community

- **Communities are described in RFC1997**
  - **Transitive & Optional attribute**
- **32 bit integer**
  - **Represented as two 16 bit integers (RFC1997/8)**
  - **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 for 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

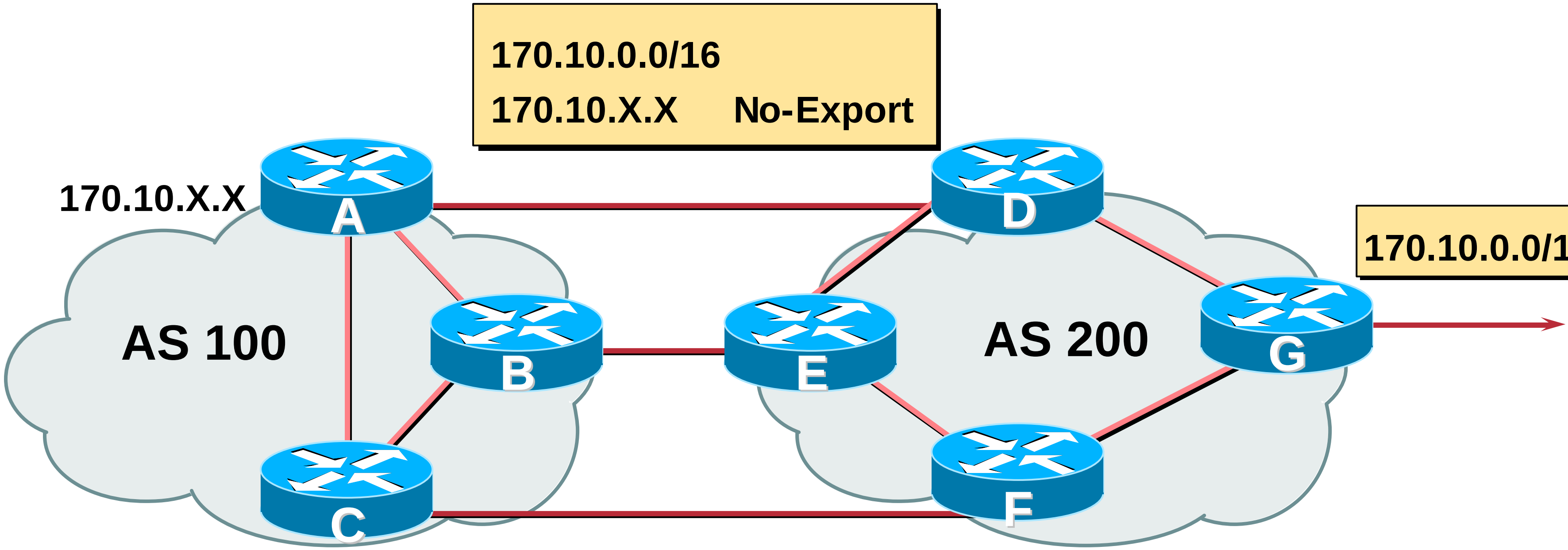
- **no-export-subconfed** 65535:65283

  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

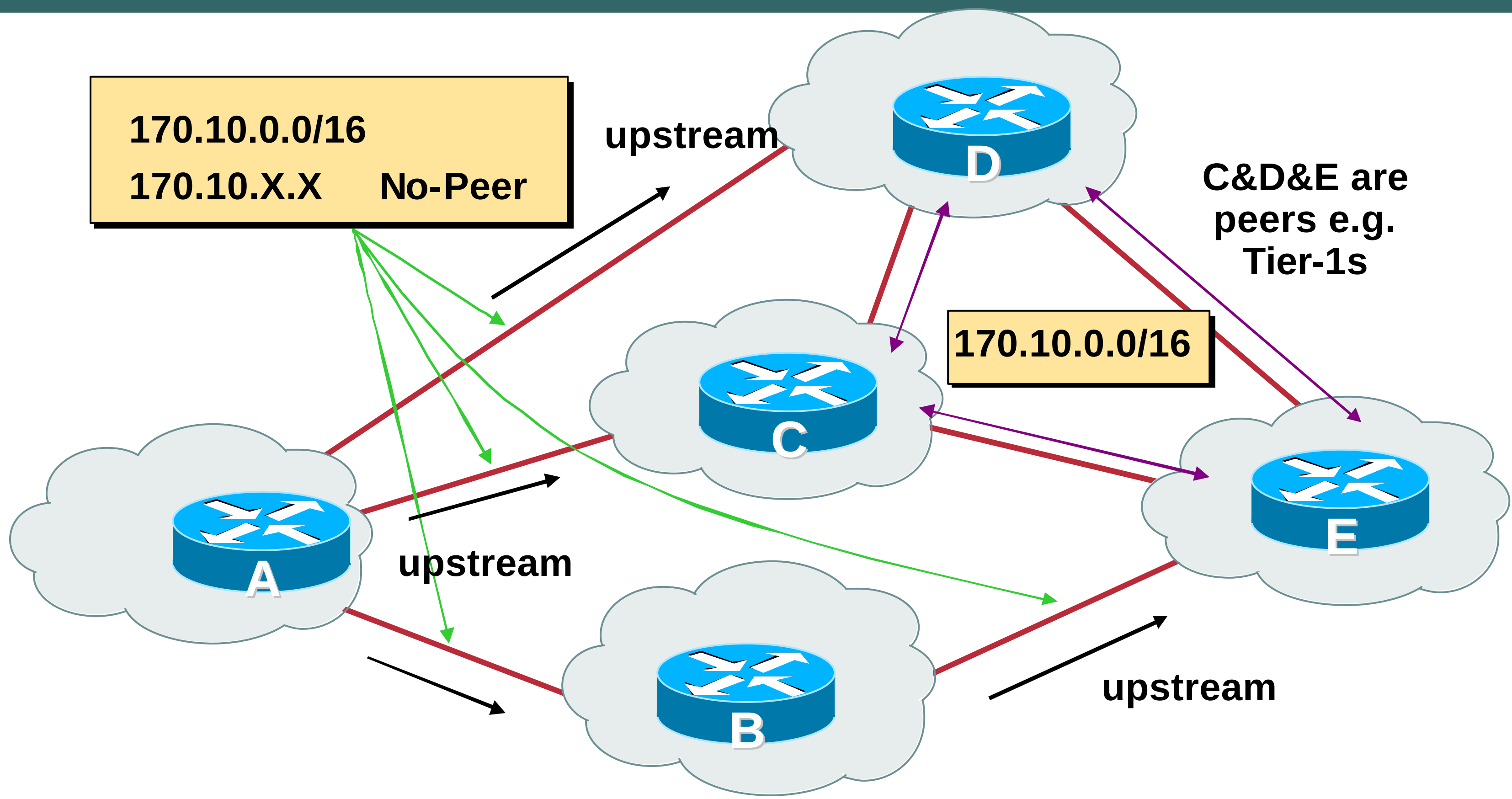


- **AS100 announces aggregate and subprefixes**

  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 Preparations

- **Before entering path selection algorithm, exclude:**
  - **Paths with no route to next hop**
  - **Paths where the AS_PATH attribute contains an AS loop**
- **Where there are multiple paths to the same destination with the same local preference, the tie-break is resolved with the "BGP Path Selection Algorithm"**

BGP Path Selection Algorithm
Part One

- **Highest local preference (global within AS)**
- **Shortest AS path**

  AS-set counts as 1 AS hop, regardless of the number of ASes in the AS-set

- **Lowest origin code**

  IGP < EGP < incomplete

BGP Path Selection Algorithm
Part Two

- **Lowest Multi-Exit Discriminator (MED)**

  MED is only compared if paths are from same AS

  (Routes with no MED are considered to have the lowest possible MED, i.e. 0)

- **Prefer eBGP path over iBGP path**
- **Prefer path with lowest IGP metric to next-hop**
- **Lowest BGP speaker router-id**
- **Lowest peer 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**

Route Refresh

Route Refresh

Problem:

- **Hard BGP peer reset required after every policy change because the router does not store prefixes that are rejected by policy**
- **Hard BGP peer reset:**

    Tears down BGP peering

    Consumes CPU

    Severely disrupts connectivity for all networks

Solution:

- **Route Refresh**

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 l[illegible] 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 peering reset 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
- Documented in RFC2439

Operation

- **Add penalty for each flap**

  NB: Change in attribute can also be penalized

- **Exponentially decay penalty**

  half life determines decay rate

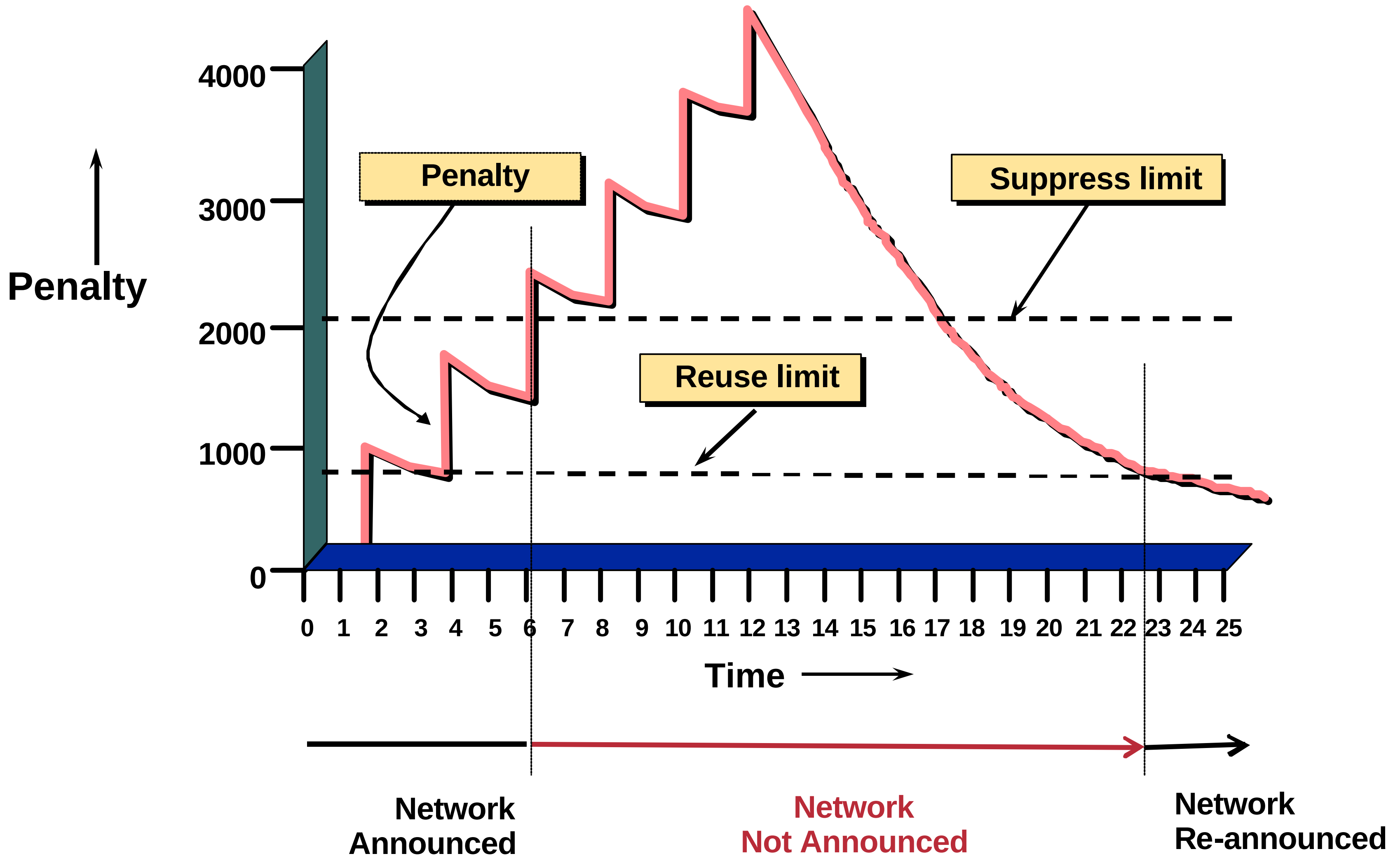
- **Penalty above suppress-limit**

  do not advertise route to BGP peers

- **Penalty decayed below reuse-limit**

  re-advertise route to BGP peers

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**

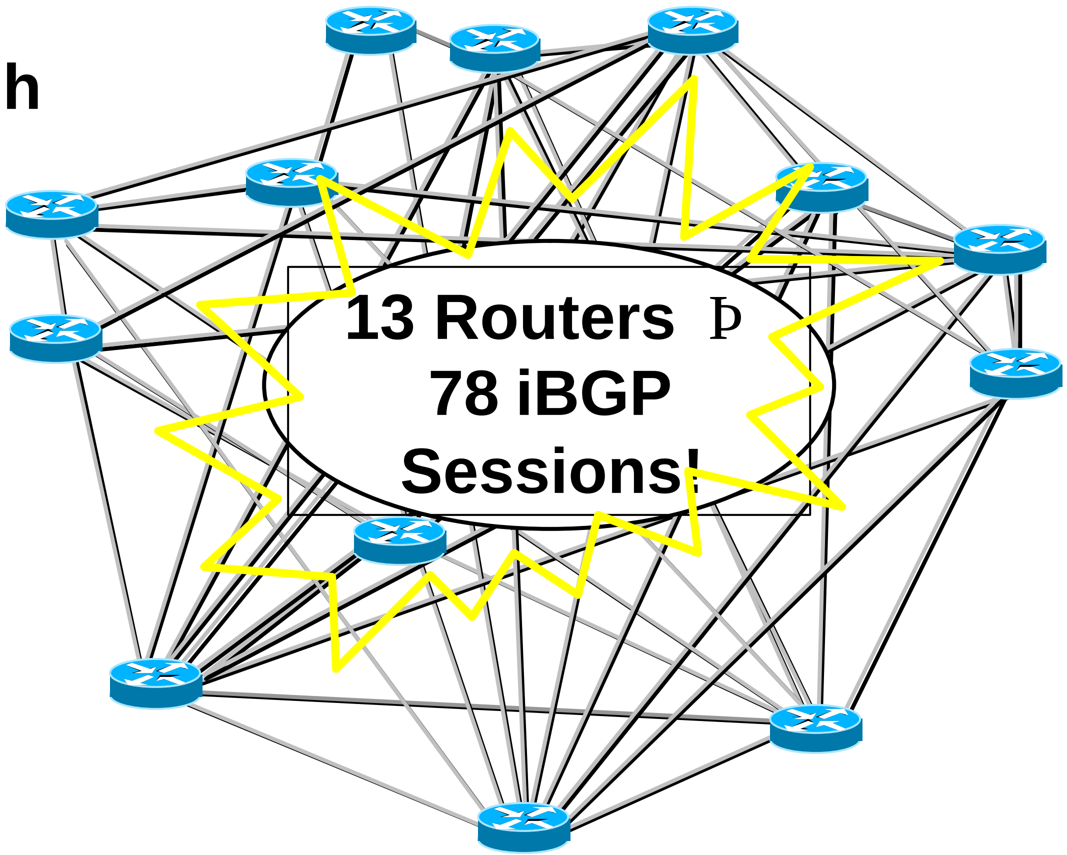
Configuration

- 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
- 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)

Route Reflectors and Confederations

Scaling iBGP mesh

Avoid ½n(n-1) iBGP mesh

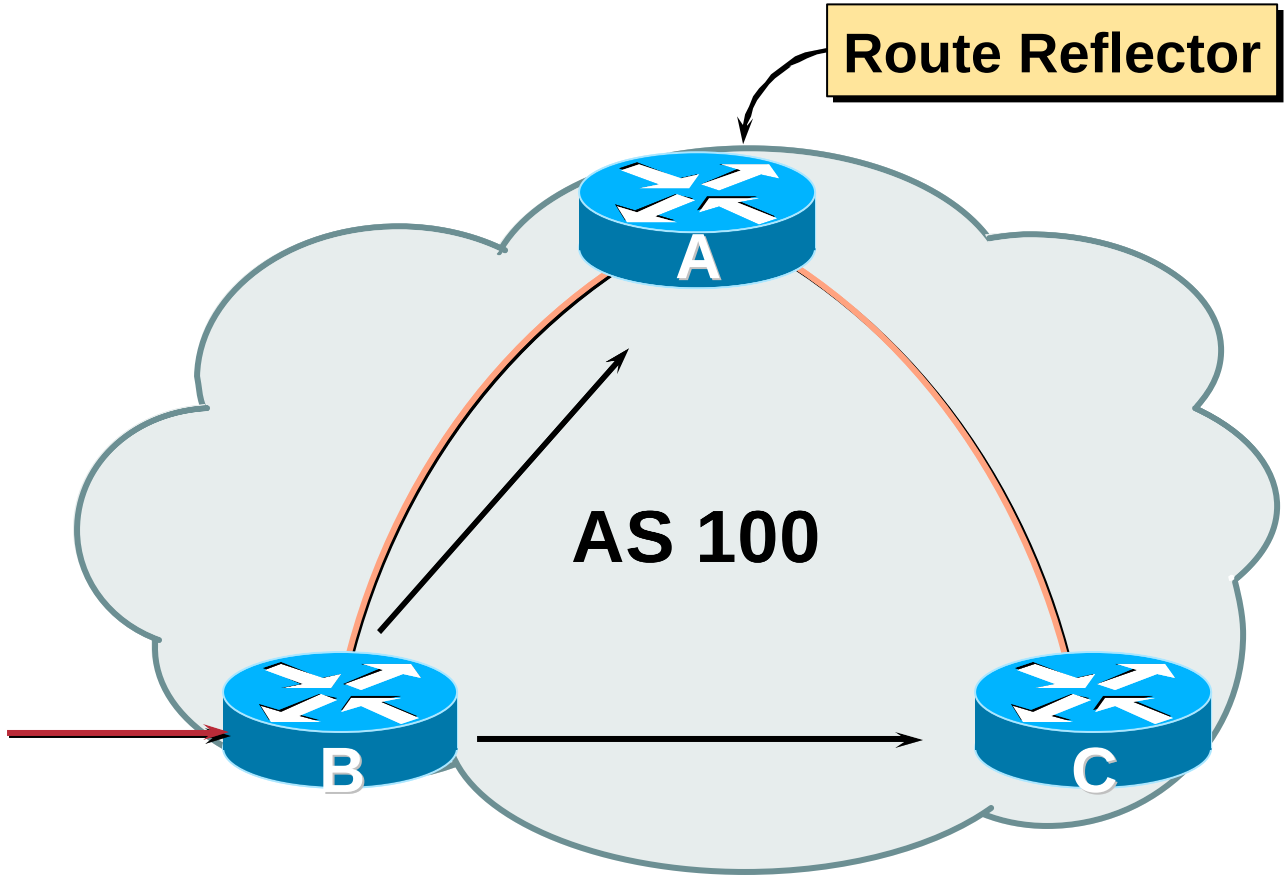


n=1000 Þ nearly half a million ibgp sessions!

Two solutions

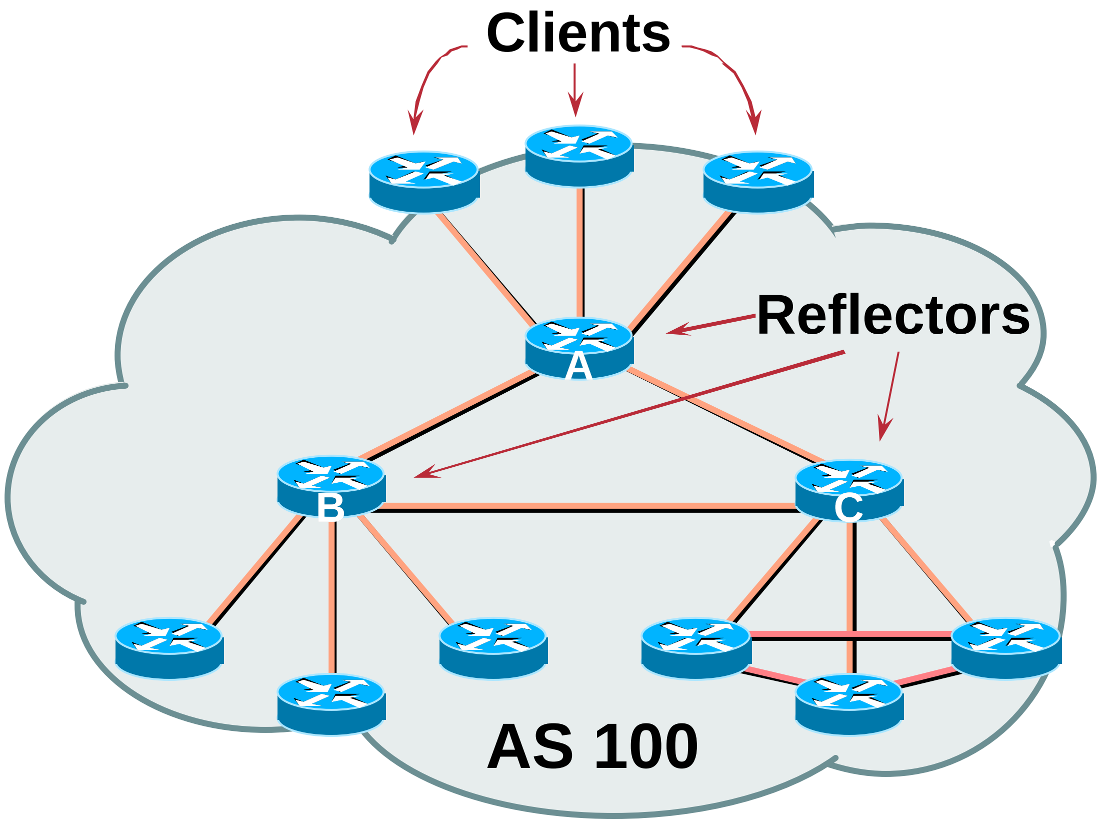
- **Route reflector – simpler to deploy and run**
- **Confederation – more complex, corner case benefits**

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

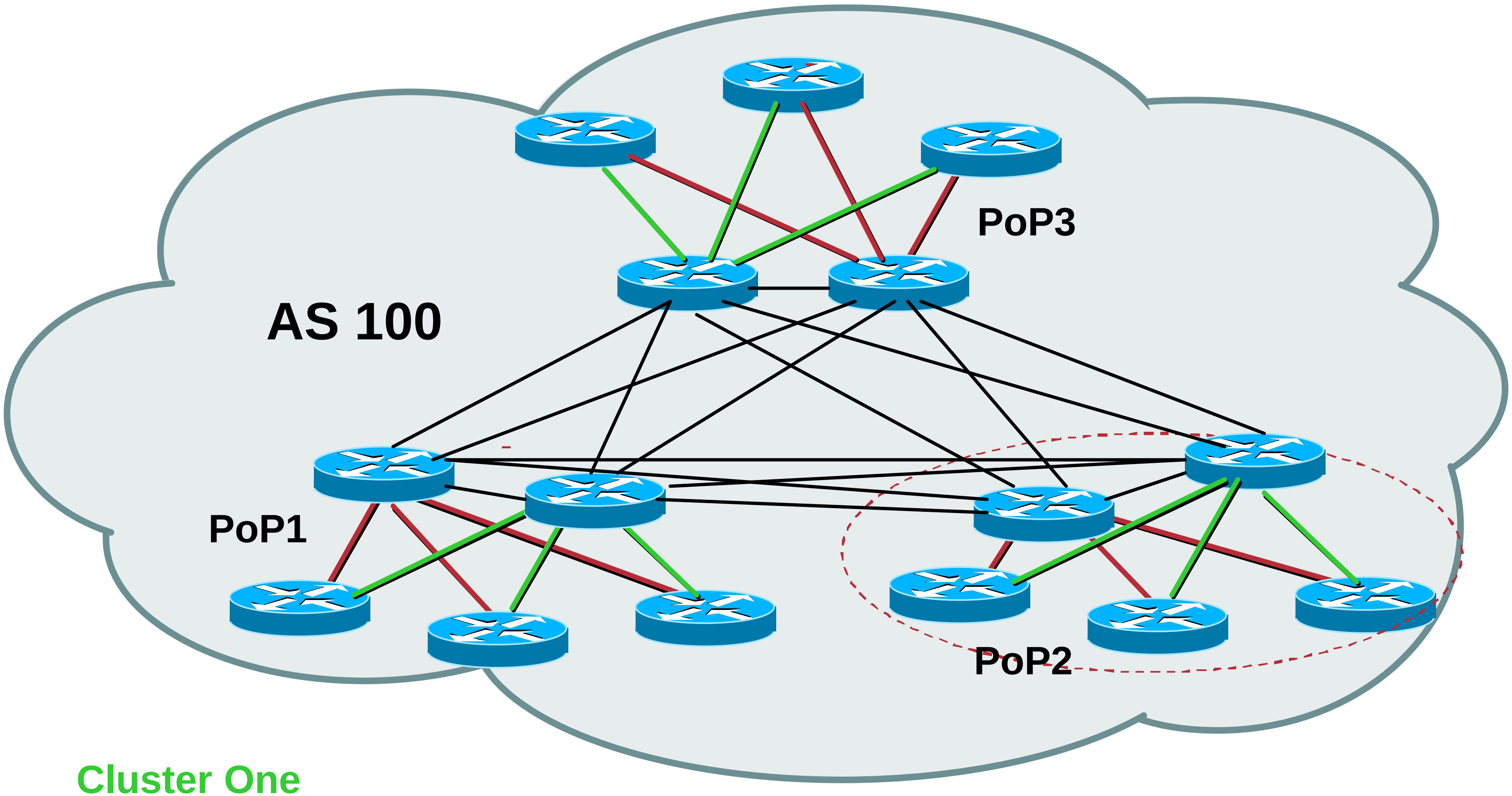
  Best to set cluster-id is from router-id (address of loopback)

  (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

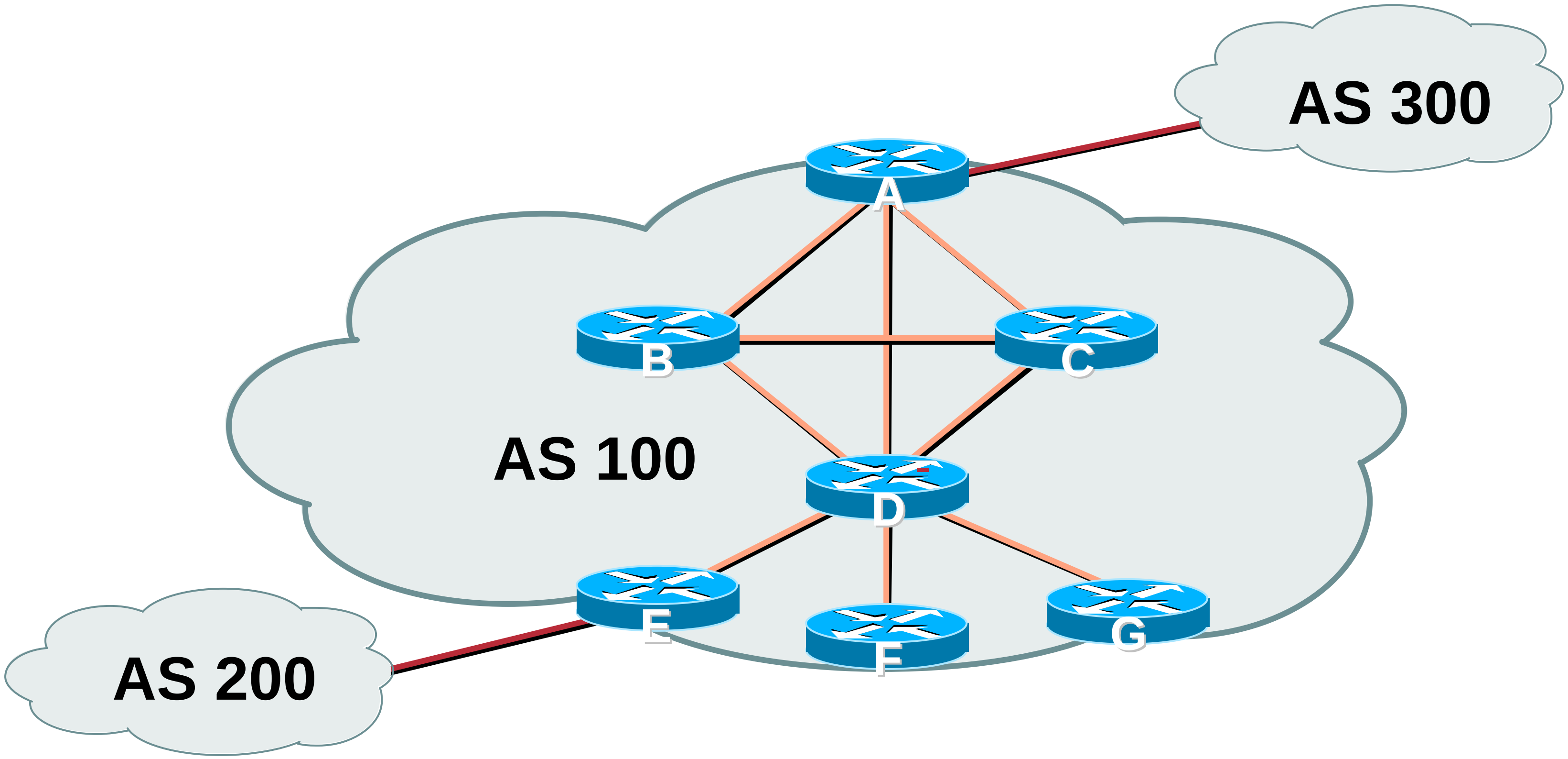
Route Reflectors: Migration

- **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-ASes
    - eBGP between sub-ASes, 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"**

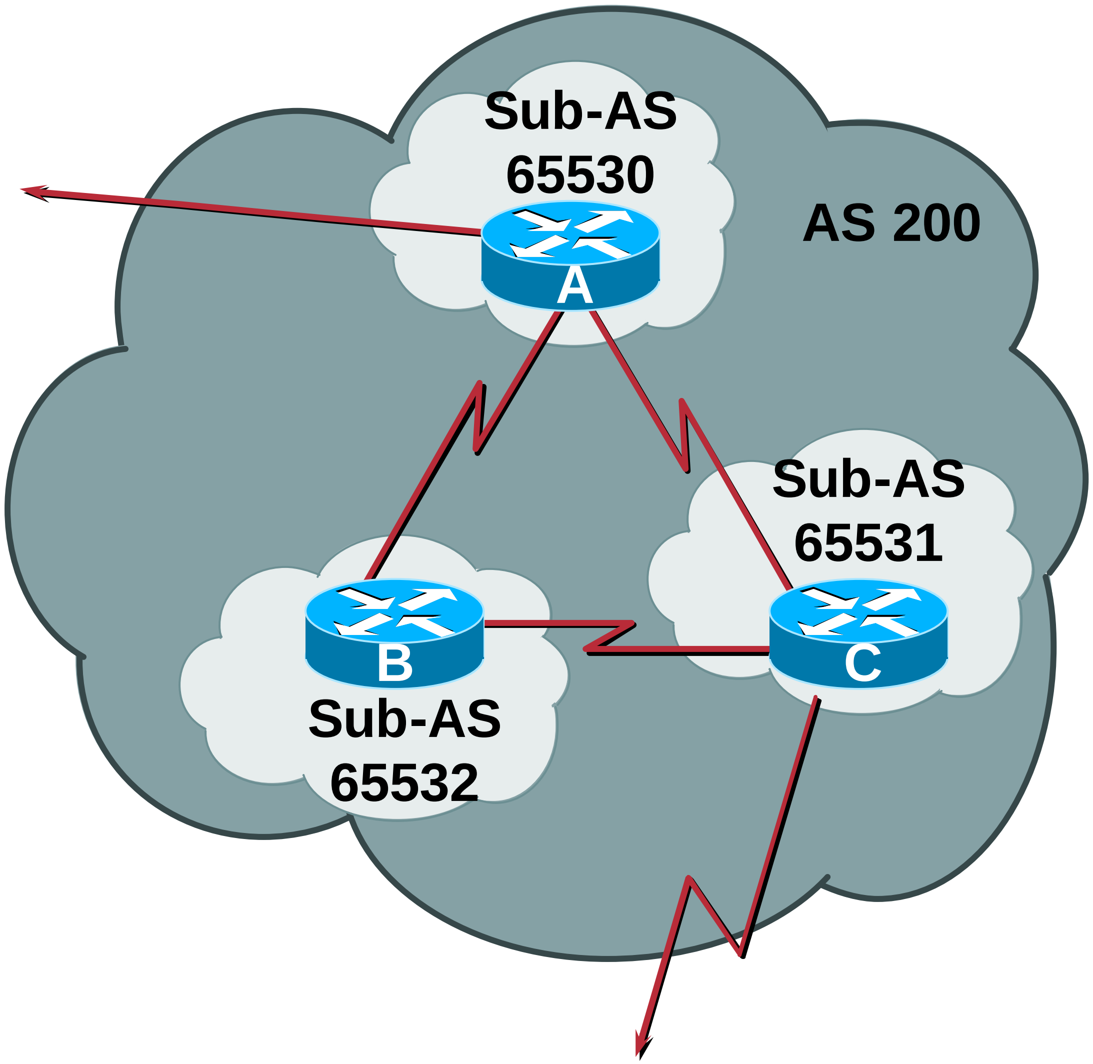
  Each sub-AS uses a number from the private AS range (64512-65534)

- **iBGP speakers in each sub-AS are fully meshed**

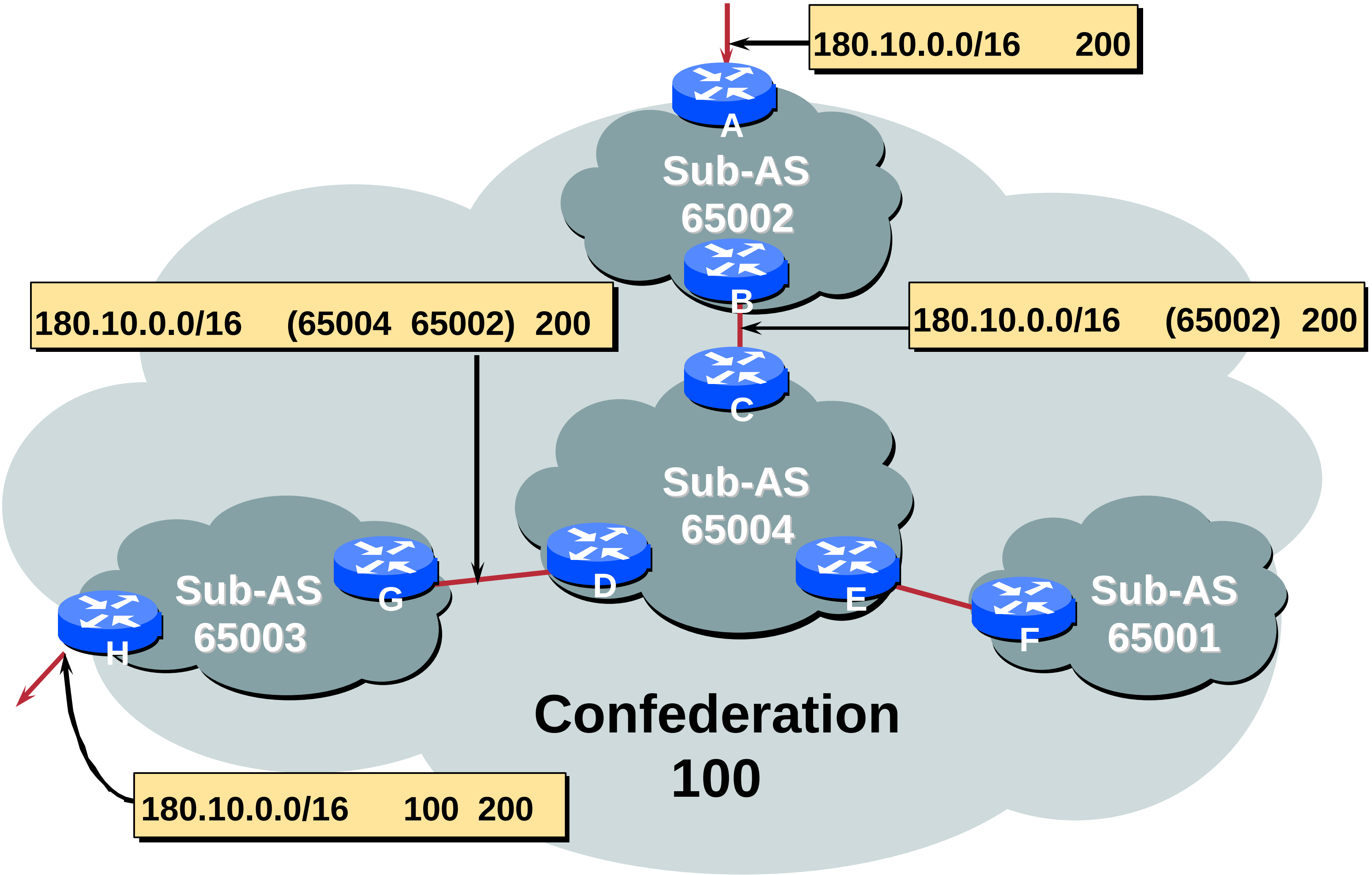
  The total number of neighbors is reduced by limiting the full mesh requirement to only the peers in the sub-AS

  Can also use Route-Reflector within sub-AS

Confederations (Cont.)



Confederations: AS-Sequence



Route Propagation Decisions

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

Route Reflectors 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	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

- **These 3 techniques should be core requirements in all ISP networks**
  - **Route Refresh**
  - **Route flap damping**
  - **Route reflectors/Confederations**

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:

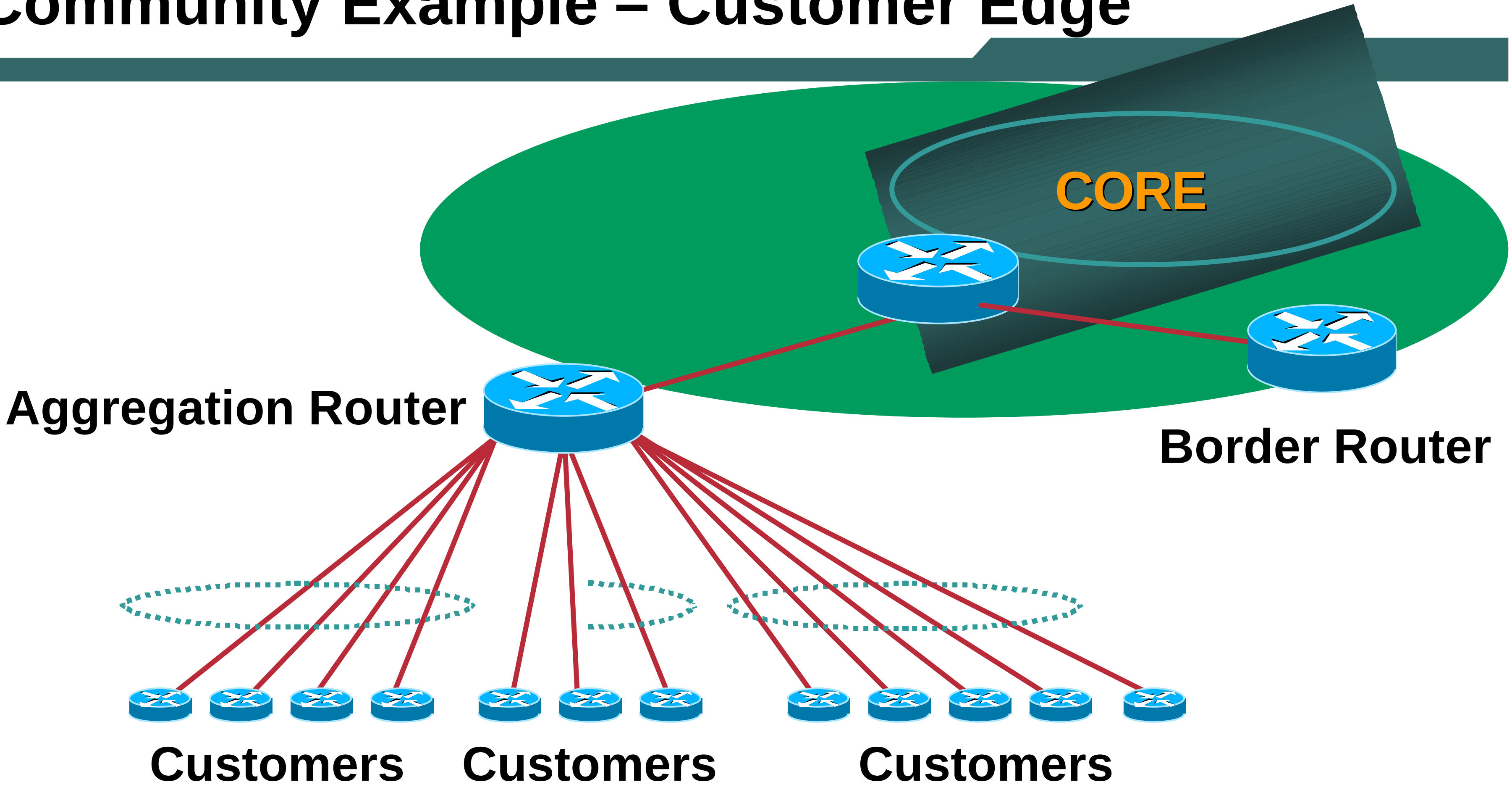
  IXP connection: community 100:2100

  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
  - Þ 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 135k 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**

  Þ **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

Some ISP Examples

- **ISPs also create communities to give customers bigger routing policy control**
- **Public policy is usually listed in the IRR**
  - **Following examples are all in the IRR**
  - **Examples build on the configuration concepts from the introductory example**
- **Consider creating communities to give policy control to customers**
  - **Reduces technical support burden**
  - **Reduces the amount of router reconfiguration, and the chance of mistakes**

Some ISP Examples
Connect.com.au

- **Australian ISP**
- **Run their own Routing Registry**

  Whois.connect.com.au

- **Permit customers to send up 8 types of communities to allow traffic engineering**

Some ISP Examples
Connect.com.au

```
aut-num:      AS2764
as-name:      ASN-CONNECT-NET
descr:        connect.com.au pty ltd
admin-c:      CC89
tech-c:       MP151
remarks:      Community  Definition
remarks:      ------------------------------------------------
remarks:      2764:1 Announce to "domestic" rate ASes only
remarks:      2764:2 Don't announce outside local POP
remarks:      2764:3 Lower local preference by 25
remarks:      2764:4 Lower local preference by 15
remarks:      2764:5 Lower local preference by 5
remarks:      2764:6 Announce to non customers with "no-export"
remarks:      2764:7 Only announce route to customers
remarks:      2764:8 Announce route over satellite link
notify:       routing@connect.com.au
mnt-by:       CONNECT-AU
changed:      mrp@connect.com.au 19990506
source:       CCAIR
```

Some ISP Examples
UUNET Europe

- **UUNET's European operation**
- **Permits customers to send communities which determine**
  - **local preferences within UUNET's network**
  - **Reachability of the prefix**
  - **How the prefix is announced outside of UUNET's network**

Some ISP Examples
UUNET Europe

```
aut-num: AS702
as-name: AS702
descr:   UUNET - Commercial IP service provider in Europe
remarks: -------------------------------------------------------------
         UUNET filters out inbound prefixes longer than /24.
         We also filter any networks within AS702:RS-INBOUND-FILTER.
         -------------------------------------------------------------
         UUNET uses the following communities with its customers:
         702:80   Set Local Pref 80 within AS702
         702:120  Set Local Pref 120 within AS702
         702:20   Announce only to UUNET AS'es and UUNET customers
         702:30   Keep within Europe, don't announce to other UUNET AS's
         702:1    Prepend AS702 once at edges of UUNET to Peers
         702:2    Prepend AS702 twice at edges of UUNET to Peers
         702:3    Prepend AS702 thrice at edges of UUNET to Peers
         -------------------------------------------------------------
         Advanced communities for customers
         702:7020 Do not announce to AS702 peers with a scope of
                  National but advertise to Global Peers, European
                  Peers and UUNET customers.

(more)
```

Some ISP Examples
UUNET Europe

```
(more)
          702:7001 Prepend AS702 once at edges of UUNET to AS702
                   peers with a scope of National.
          702:7002 Prepend AS702 twice at edges of UUNET to AS702
                   peers with a scope of  National.
          702:7003 Prepend AS702 thrice at edges of UUNET to AS702
                   peers with a scope  of National.
          702:8020 Do not announce to AS702 peers with a scope of
                   European but advertise to Global Peers, National
                   Peers and UUNET  customers.
          702:8001 Prepend AS702 once at edges of UUNET to AS702
                   peers with a scope of European.
          702:8002 Prepend AS702 twice at edges of UUNET to AS702
                   peers with a scope of  European.
          702:8003 Prepend AS702 thrice at edges of UUNET to AS702
                   peers with a scope  of European.
          ---------------------------------------------------------------
          Additional details of the UUNET communities are located at:
          http://global.mci.com/uk/customer/bgp/
          ---------------------------------------------------------------
mnt-by:   WCOM-EMEA-RICE-MNT
changed:  rice@lists.mci.com 20040523
source:   RIPE
```

Some ISP Examples
BT

- **Formerly Concert's European network**
- **One of the most comprehensive community lists around**

  Seems to be based on definitions originally used in Tiscali's network

  whois –h whois.ripe.net AS5400 reveals all

- **Extensive community definitions allow sophisticated traffic engineering by customers**

Some ISP Examples
BT

```
aut-num:        AS5400
as-name:        CIPCORE
descr:          BT European Backbone
remarks:        The following BGP communities can be set by BT
remarks:        BGP customers to affect announcements to major peers.
remarks:
remarks:        Community to                            Community to
remarks:        Not announce      To peer:              AS prepend 5400
remarks:
remarks:        5400:1000 All peers & Transits          5400:2000
remarks:
remarks:        TRANSITS:
remarks:
remarks:        5400:1500 All Transits                  5400:2500
remarks:        5400:1501 Sprint Transit (AS1239)       5400:2501
remarks:        5400:1502 C&W Transit (AS3561)          5400:2502
remarks:        5400:1503 Level 3 Transit (AS3356)      5400:2503
remarks:        5400:1504 AT&T Transit (AS7018)         5400:2504
remarks:        5400:1505 UUnet Transit (AS701)         5400:2505
remarks:

(more)
```

Some ISP Examples
BT

```
(more)

remarks:      Community to                                  Community to
remarks:      Not announce      To peer:                    AS prepend 5400

remarks:      PEERS:
remarks:
remarks:      5400:1001 Nexica (AS24592)                    5400:2001
remarks:      5400:1002 Fujitsu (AS3324)                    5400:2002
remarks:      5400:1003 Unisource (AS3300)                  5400:2003
remarks:      5400:1004 C&W EU (AS1273)                     5400:2004
remarks:      5400:1005 UUnet (AS702)                       5400:2005
remarks:      5400:1006 Eltec (AS30892)                     5400:2006
remarks:      5400:1007 SupportNet (8582)                   5400:2007
remarks:      5400:1008 AT&T (AS2686)                       5400:2008
remarks:      5400:1009 ACENS (AS16371)                     5400:2009
remarks:      5400:1010 RIPE (AS3333)                       5400:2010
remarks:      5400:1011 Altecom (AS24983)                   5400:2011
remarks:      5400:1012 Globix (AS4513)                     5400:2012
<snip>
notify:       notify@eu.bt.net
mnt-by:       CIP-MNT
source:       RIPE
```

And many many more!

Some ISP Examples
Carrier1

- **European ISP**
- **Another very comprehensive list of community definitions**

  whois –h whois.ripe.net AS8918 reveals all

Some ISP Examples
Carrier1

```
aut-num:  AS8918
descr:    Carrier1 Autonomous System
<snip>
remarks:  The Following communities can be used by Carrier1 Customers
remarks:  to control outbound routing announcements
remarks:  *
remarks:  Community  Definition
remarks:  ------------------------------------------------------
remarks:  *
remarks:  8918:2000  Do not announce to C1 customers
remarks:  8918:2010  Do not announce to C1 peers, peers+ and transit
remarks:  8918:2015  Do not announce to C1 transit providers
remarks:  *
remarks:  8918:2020  Do not announce to Global Crossing (AS 3549)
remarks:  8918:2035  Do not announce to UUNet           (AS 702)
remarks:  8918:2040  Do not announce to Lambdanet       (AS 13237)
remarks:  8918:2060  Do not announce to SPRINT          (AS 1239)
remarks:  *
remarks:  8918:2070  Do not announce to AMS-IX peers
remarks:  8918:2080  Do not announce to NL-IX peers
remarks:  8918:2090  Do not announce to Packet Exchange Peers
remarks:  ------------------------------------------------------

(more)
```

Some ISP Examples
Carrier1

```
(more)

remarks: Communities to prepend AS8918 to outbound routing announcements:
remarks: *
remarks: Community Definition
remarks: -------------------------------------------------------------
remarks: *
remarks: 8918:3001 Lambdanet (AS 13237) prepend 8918
remarks: 8918:3002 Lambdanet (AS 13237) prepend 8918 8918
remarks: 8918:3003 Lambdanet (AS 13237) prepend 8918 8918 8918
remarks: 8918:3004 Lambdanet (AS 13237) prepend 8918 8918 8918 8918
remarks: 8918:3005 Lambdanet (AS 13237) prepend 8918 8918 8918 8918 8918
remarks: *
remarks: Global-Crossing (GBLX)
remarks: 8918:3011 GBLX (AS 3549) prepend 8918
remarks: 8918:3012 GBLX (AS 3549) prepend 8918 8918
remarks: 8918:3013 GBLX (AS 3549) prepend 8918 8918 8918
remarks: 8918:3014 GBLX (AS 3549) prepend 8918 8918 8918 8918
remarks: 8918:3015 GBLX (AS 3549) prepend 8918 8918 8918 8918 8918
<snip>
notify:  inoc@carrier1.net
mnt-by:  CARRIER1-MNT
source:  RIPE
```

Some ISP Examples
Level 3

- **Highly detailed AS object held on the RIPE Routing Registry**
- **Also a very comprehensive list of community definitions**

  whois –h whois.ripe.net AS3356 reveals all

Some ISP Examples
Level 3

```
aut-num:        AS3356
descr:          Level 3 Communications
<snip>
remarks:        -----------------------------------------------------------
remarks:        customer traffic engineering communities - Suppression
remarks:        -----------------------------------------------------------
remarks:        64960:XXX - announce to AS XXX if 65000:0
remarks:        65000:0   - announce to customers but not to peers
remarks:        65000:XXX - do not announce at peerings to AS XXX
remarks:        -----------------------------------------------------------
remarks:        customer traffic engineering communities - Prepending
remarks:        -----------------------------------------------------------
remarks:        65001:0   - prepend once  to all peers
remarks:        65001:XXX - prepend once  at peerings to AS XXX
remarks:        65002:0   - prepend twice to all peers
remarks:        65002:XXX - prepend twice at peerings to AS XXX
remarks:        65003:0   - prepend 3x to all peers
remarks:        65003:XXX - prepend 3x    at peerings to AS XXX
remarks:        65004:0   - prepend 4x to all peers
remarks:        65004:XXX - prepend 4x    at peerings to AS XXX
<snip>
mnt-by:         LEVEL3-MNT
source:         RIPE
```

And many many more!

BGP for Internet Service Providers

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

Deploying BGP in an ISP Network

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

Deploying BGP

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

The role of IGP and iBGP

Ships in the night?

Or

Good foundations?

BGP versus OSPF/ISIS

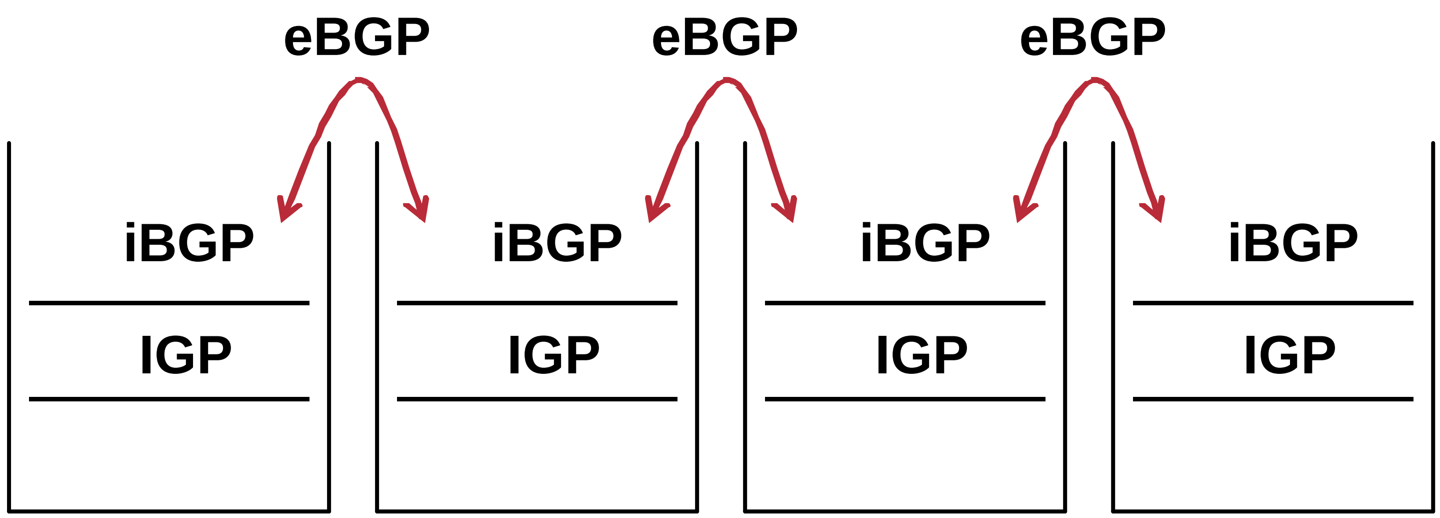
- Internal Routing Protocols (IGPs)
  - examples are ISIS and OSPF
  - used for carrying infrastructure addresses
  - NOT used for carrying Internet prefixes or customer prefixes
  - design goal is to minimise number of prefixes in IGP to aid scalability and rapid convergence

BGP versus OSPF/ISIS

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

BGP/IGP model used in ISP networks

- **Model representation**



BGP versus OSPF/ISIS

- **DO NOT:**
  - **distribute BGP prefixes into an IGP**
  - **distribute IGP routes into BGP**
  - **use an IGP to carry customer prefixes**
- **YOUR NETWORK WILL NOT SCALE**

Injecting prefixes into iBGP

- **Use iBGP to carry customer prefixes**

  don't ever use IGP

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

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

  i.e. avoid iBGP flaps caused by interface flaps

Aggregation

Quality or Quantity?

Aggregation

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

Aggregation

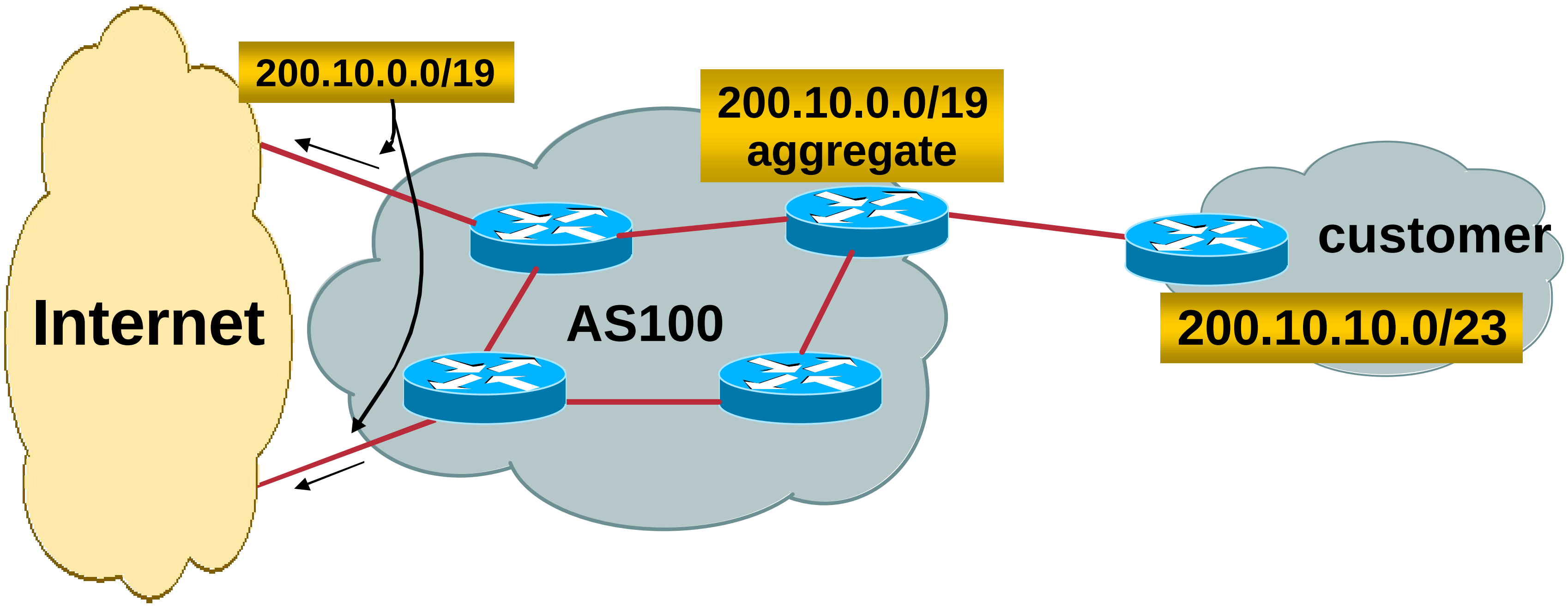
- **Address block should be announced to the Internet as an aggregate**
- **Subprefixes of address block should NOT be announced to Internet unless special circumstances (more later)**
- **Aggregate should be generated internally**

Not on the network borders!

Announcing an Aggregate

- ISPs who don't and won't aggregate are held in poor regard by community
- Registries publish their minimum allocation size
  - Anything from a /20 to a /22 depending on RIR
  - Different sizes for different address blocks
- No real reason to see anything longer than a /22 prefix in the Internet
  - BUT there are currently >71000 /24s!

Aggregation – Example



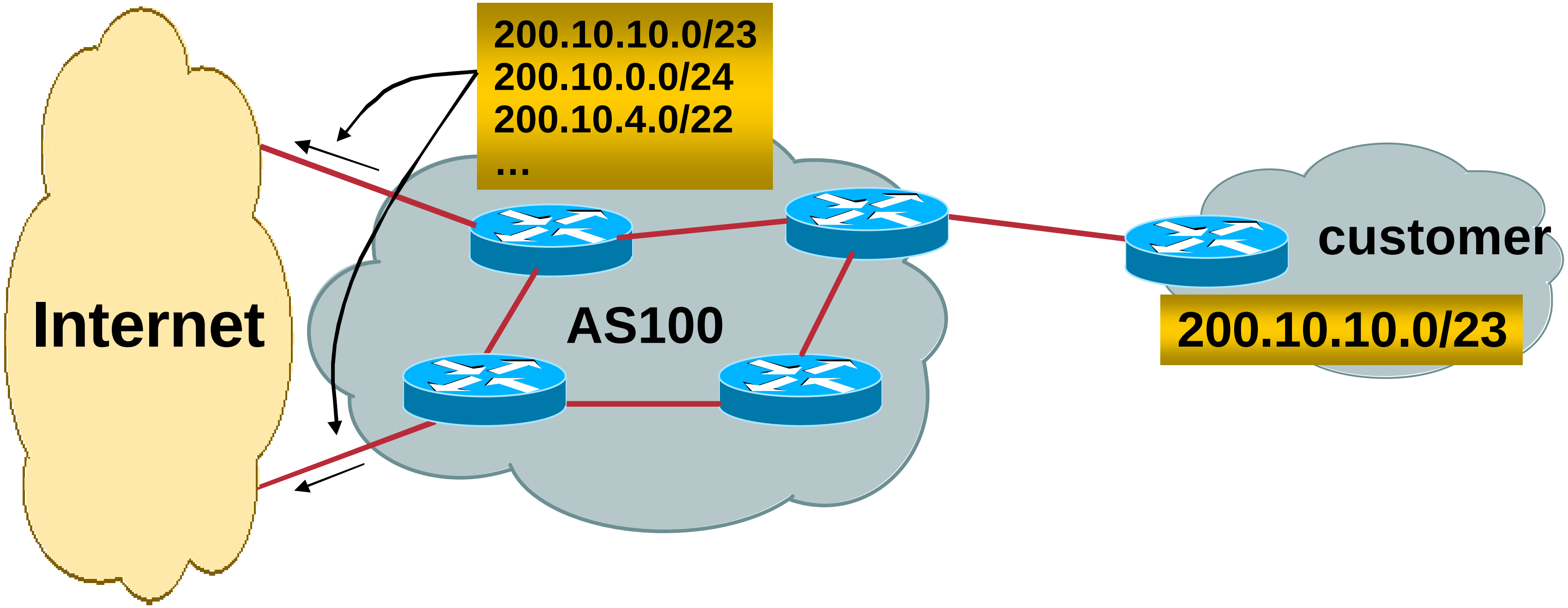
- **Customer has /23 network assigned from AS100's /19 address block**
- **AS100 announced /19 aggregate to the Internet**

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 (May 2004)

- **Current Internet Routing Table Statistics**

BGP Routing Table Entries	138240
Prefixes after maximum aggregation	84036
Unique prefixes in Internet	67146
Prefixes smaller than registry alloc	62012
/24s announced	75003
only 5523 /24s are from 192.0.0.0/8	
ASes in use	17267

“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**

“The New Swamp” July 2000

- **192/3 space contributed 69000 networks – rest of Internet contributed 16000 networks**

Block	Networks	Block	Networks	Block	Networks	Block	Networks
192/8	6352	204/8	4694	217/8	0	65/8	0
193/8	2746	205/8	3210	218/8	0	66/8	0
194/8	2963	206/8	4206	219/8	0	67/8	0
195/8	1689	207/8	3943	220/8	0	68/8	0
196/8	525	208/8	4804	221/8	0	69/8	0
198/8	4481	209/8	4755	222/8	0	80/8	0
199/8	4084	210/8	1375	24/8	1122	81/8	0
200/8	2436	211/8	532	61/8	80	82/8	0
201/8	0	212/8	1859	62/8	428	83/8	0
202/8	3712	213/8	635	63/8	2198		
203/8	5494	216/8	4177	64/8	1439		

"The New Swamp" April 2004

- **192/3 space contributes 90000 networks – rest of Internet contributes 46500 networks**

Block	Networks	Block	Networks	Block	Networks	Block	Networks
192/8	6676	204/8	4224	217/8	2036	65/8	3234
193/8	4082	205/8	2762	218/8	985	66/8	5698
194/8	3229	206/8	3625	219/8	781	67/8	1018
195/8	2853	207/8	3906	220/8	855	68/8	2836
196/8	752	208/8	3604	221/8	251	69/8	1713
198/8	4674	209/8	4604	222/8	155	80/8	1252
199/8	3881	210/8	2989	24/8	2508	81/8	807
200/8	5346	211/8	1785	61/8	1592	82/8	697
201/8	125	212/8	2507	62/8	1458	83/8	230
202/8	7122	213/8	2559	63/8	2679		
203/8	7735	216/8	6046	64/8	3852		

"The New Swamp" Summary

- 192/3 space shows creeping increase in bad aggregation

    e.g. 193/8, 200/8, 202/7, 208/8 and 216/8 show major changes not consistent with fresh RIR allocations

- Rest of address space is showing similar increase too

    New RIR blocks in former A space are showing deaggregation

    Other nets in former A and B space are also being deaggregated

- Why??

    Excuses usually are traffic engineering

    Real reason tends to be lack of knowledge and laziness

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

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

Status Summary

Table History

Date	Prefixes	CIDR Aggregated
17-05-04	134431	94339
18-05-04	134557	94505
19-05-04	134683	94655
20-05-04	134815	94861
21-05-04	134981	94909
22-05-04	135027	94796
23-05-04	135200	94941
24-05-04	136041	94926

Plot: BGP Table Size

AS Summary

17183	Number of ASes in routing system
6951	Number of ASes announcing only one prefix
1429	Largest number of prefixes announced by an AS AS7018: ATTW AT&T WorldNet Services
73561344	Largest address span announced by an AS (/32s) AS568: DISOUN DISO-UNRRA

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

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

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').

--- 24May04 ---

ASnum	NetsNow	NetsAggr	NetGain	% Gain	Description
Table	136002	94906	41096	30.2%	All ASes
AS4134	751	153	598	79.6%	CHINANET-BACKBONE No.31,Jin-rong Street
AS18566	704	163	541	76.8%	CVAD Covad Communications
AS4323	725	199	526	72.6%	TWTC Time Warner Telecom
AS9583	475	36	439	92.4%	SATYAMNET-AS Satyam Infoway Ltd.,
AS7018	1429	992	437	30.6%	ATTW AT&T WorldNet Services
AS6197	698	314	384	55.0%	BNS-14 BellSouth Network Solutions, Inc
AS7843	496	115	381	76.8%	ADELPH-13 Adelphia Corp.
AS701	1293	930	363	28.1%	UU UUNET Technologies, Inc.
AS22909	387	37	350	90.4%	CMCS Comcast Cable Communications, Inc.
AS6198	555	225	330	59.5%	BNS-14 BellSouth Network Solutions, Inc
AS22773	372	52	320	86.0%	CXAB Cox Communications Inc. Atlanta
AS27364	358	40	318	88.8%	ARMC Armstrong Cable Services
AS9929	334	33	301	90.1%	CNCNET-CN China Netcom Corp.
AS11172	355	55	300	84.5%	Servicios Alestra S.A de C.V
AS1239	940	644	296	31.5%	SPRN Sprint
AS17676	339	50	289	85.3%	JPNIC-JP-ASN-BLOCK Japan Network Information Center
AS4355	381	99	282	74.0%	ERSD EARTHLINK, INC
AS6140	386	121	265	68.7%	IMPSA ImpSat
AS6478	304	48	256	84.2%	ATTW AT&T WorldNet Services
AS6347	401	150	251	62.6%	SAVV SAVVIS Communications Corporation
AS1221	857	619	238	27.8%	ASN-TELSTRA Telstra Pty Ltd
AS209	735	502	233	31.7%	QWEST-4 Qwest
AS25844	243	16	227	93.4%	SASMFL-2 Skadden, Arps, Slate, Meagher & Flom LLP
AS14654	230	5	225	97.8%	WAYPOR-3 Wayport
AS3356	894	678	216	24.2%	LEVEL3 Level 3 Communications
AS4766	474	263	211	44.5%	KIX Korea Internet Exchange for "96 World Internet Exposition
AS9443	358	155	203	56.7%	INTERNETPRIMUS-AS-AP Primus Telecommunications
AS2386	427	240	187	43.8%	ADCS-1 AT&T Data Communications Services
AS5668	380	197	183	48.2%	CIH-12 CenturyTel Internet Holdings, Inc.
AS6327	208	28	180	86.5%	SHAWC-2 Shaw Communications Inc.
Total	16489	7159	9330	56.6%	Top 30 total

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

Top 20 Added Routes this week per Originating AS

Prefixes	ASnum	AS Description
694	AS18566	CVAD Covad Communications
144	AS11172	Servicios Alestra S.A de C.V
118	AS10036	PARNET-AS C&M Communication Co. Ltd.
62	AS27257	WAIR Webair Internet Development Inc
39	AS1591	DNIC DoD Network Information Center
32	AS16814	NSS S.A.
30	AS20115	CC04 Charter Communications
28	AS9225	LEVEL3-AP Reach Networks HK Ltd.
28	AS27046	DNIC DoD Network Information Center
26	AS10113	DATAFAST-AP DATAFAST TELECOMMUNICATIONS LTD
25	AS17854	CABLELINE-AS-KR BANDOCABLELINE
23	AS8866	BTC-AS Bulgarian Telecommunication Company
22	AS812	ROCB Rogers Cable Inc.
21	AS6467	ACSI e.spire Communications, Inc.
20	AS5979	DNIC DoD Network Information Center
20	AS20889	STELLAR-AS Stellar-PCS GmbH Germany
18	AS7018	ATTW AT&T WorldNet Services
17	AS5180	DNIC DoD Network Information Center
16	AS17536	PRODIGY-AS-AP Prodidgy Telecommunications
16	AS3243	TELEPAC Telepac - Comunicacoes Interactivas, SA

Top 20 Withdrawn Routes this week per Originating AS

Prefixes	ASnum	AS Description
-56	AS17964	DXTNET Beijing Dian-Xin-Tong Network Technologies Co., Ltd.
-34	AS9782	WOOSONGEDU Woosong University
-25	AS10223	UECOMM-AU Uecomm Ltd
-25	AS7586	PDOX-AS-AP Paradox Digital Pty Ltd
-20	AS13237	LAMBDANET-AS European Backbone of LambdaNet Germany
-20	AS30981	HSS-CGN-AS Horizon Satellite Services FZ LLC
-17	AS6198	BNS-14 BellSouth Network Solutions, Inc
-16	AS32065	VTC1 Vortech Inc.
-15	AS8406	AS8406 PIPEX Communications
-14	AS2548	ATCW Allegiance Telecom Companies Worldwide
-12	AS20115	CC04 Charter Communications
-12	AS4452	ACCESS-3 Access America
-11	AS21882	PRIORI-26 Priority Networks Inc.
-11	AS3356	LEVEL3 Level 3 Communications
-11	AS9051	IDM Autonomous System
-11	AS3043	AMC-92 Amphibian Media Corporation

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

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
1022	1429	AS7018	ATTW AT&T WorldNet Services
853	1293	AS701	UU UUNET Technologies, Inc.
697	704	AS18566	CVAD Covad Communications
682	698	AS6197	BNS-14 BellSouth Network Solutions, Inc
657	940	AS1239	SPRN Sprint
649	857	AS1221	ASN-TELSTRA Telstra Pty Ltd
638	751	AS4134	CHINANET-BACKBONE No.31,Jin-rong Street
632	894	AS3356	LEVEL3 Level 3 Communications
616	725	AS4323	TWTC Time Warner Telecom
549	554	AS20115	CC04 Charter Communications
544	555	AS6198	BNS-14 BellSouth Network Solutions, Inc
492	496	AS7843	ADELPH-13 Adelphia Corp.
472	475	AS9583	SATYAMNET-AS Satyam Infoway Ltd.,
456	735	AS209	QWEST-4 Qwest
440	474	AS4766	KIX Korea Internet Exchange for "96 World Internet Exposition
387	387	AS22909	CMCS Comcast Cable Communications, Inc.
363	380	AS5668	CIH-12 CenturyTel Internet Holdings, Inc.
363	401	AS6347	SAVV SAVVIS Communications Corporation
358	372	AS22773	CXAB Cox Communications Inc. Atlanta
358	651	AS702	AS702 MCI EMEA

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=AS1221&view=4637

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      %
   22 AS1221    ASN-TELSTRA Telstra Pty Ltd               857    324     86    619    238  27.77%


AS 1221: ASN-TELSTRA Telstra Pty Ltd
  Prefix  (AS Path)                    Aggregation Action
47.153.192.0/18       4637 1221
61.9.128.0/17         4637 1221
129.223.0.0/16        4637 1221
129.223.0.0/18        4637 1221           - Withdrawn - matching aggregate 129.223.0.0/16 4637 1221
129.223.64.0/19       4637 1221           - Withdrawn - matching aggregate 129.223.0.0/16 4637 1221
129.223.131.0/24      4637 1221           - Withdrawn - matching aggregate 129.223.0.0/16 4637 1221
129.223.160.0/19      4637 1221           - Withdrawn - matching aggregate 129.223.0.0/16 4637 1221
129.223.192.0/19      4637 1221           - Withdrawn - matching aggregate 129.223.0.0/16 4637 1221
129.223.224.0/19      4637 1221           - Withdrawn - matching aggregate 129.223.0.0/16 4637 1221
129.226.0.0/17        4637 1221
134.144.72.0/21       4637 1221
136.153.0.0/16        4637 1221
137.76.3.0/24         4637 1221
137.76.6.0/24         4637 1221
137.76.8.0/24         4637 1221
137.76.28.0/24        4637 1221
137.76.31.0/24        4637 1221
137.76.60.0/24        4637 1221
137.76.81.0/24        4637 1221
137.147.0.0/16        4637 1221
138.7.32.0/19         4637 1221           + Announce - aggregate of 138.7.32.0/20 (4637 1221) and 138.7.48.0/20 (4637 12
138.7.32.0/21         4637 1221           - Withdrawn - aggregated with 138.7.40.0/21 (4637 1221)
138.7.40.0/21         4637 1221           - Withdrawn - aggregated with 138.7.32.0/21 (4637 1221)
138.7.48.0/21         4637 1221           - Withdrawn - aggregated with 138.7.56.0/21 (4637 1221)
138.7.56.0/21         4637 1221           - Withdrawn - aggregated with 138.7.48.0/21 (4637 1221)
138.7.64.0/21         4637 1221
138.7.80.0/21         4637 1221
138.7.96.0/20         4637 1221           + Announce - aggregate of 138.7.96.0/21 (4637 1221) and 138.7.104.0/21 (4637 1
138.7.96.0/21         4637 1221           - Withdrawn - aggregated with 138.7.104.0/21 (4637 1221)
138.7.104.0/21        4637 1221           - Withdrawn - aggregated with 138.7.96.0/21 (4637 1221)
138.7.120.0/21        4637 1221
138.7.128.0/20        4637 1221           + Announce - aggregate of 138.7.128.0/21 (4637 1221) and 138.7.136.0/21 (4637
138.7.128.0/21        4637 1221           - Withdrawn - aggregated with 138.7.136.0/21 (4637 1221)
```

http://www.cidr-report.org/cgi-bin/as-report?as=AS109&view=4637

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.

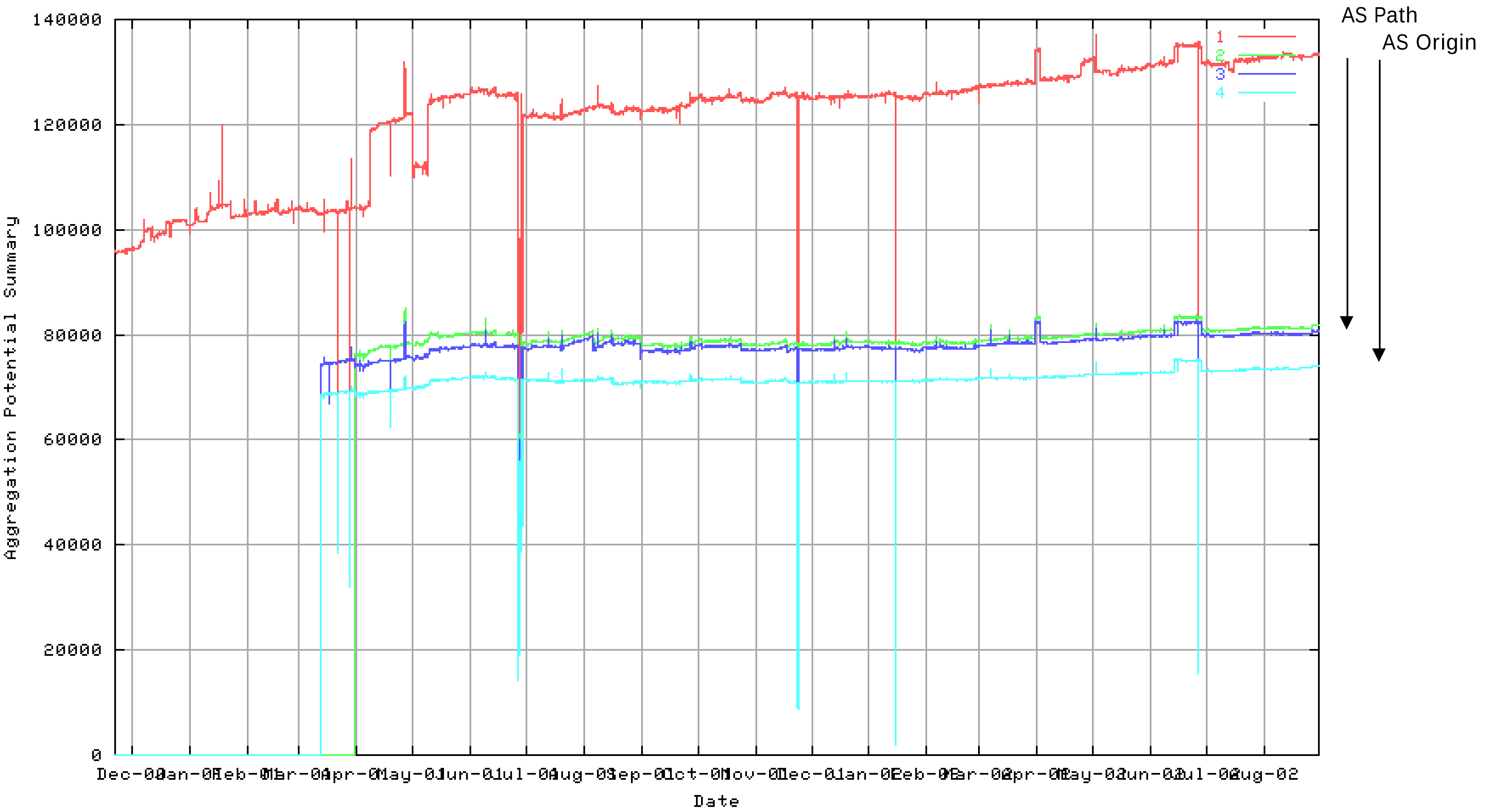
Rank	AS	AS Name	Current	Wthdw	Aggte	Annce	Redctn	%
4286	AS109	CISCO-EU-109 Cisco Systems Global ASN -	29	0	0	29	0	0.00%

```
AS  109: CISCO-EU-109 Cisco Systems Global ASN - ARIN Assigned
  Prefix   (AS Path)                    Aggregation Action
64.100.128.0/17      4637 1239 109
64.101.0.0/17        4637 701 109
64.101.128.0/18      4637 701 109
64.101.192.0/19      4637 701 109
64.101.240.0/20      4637 701 109
64.102.0.0/16        4637 701 109
64.103.0.0/17        4637 1239 109
64.104.0.0/16        4637 3356 109
64.104.0.0/18        4637 4694 4713 2914 109
64.104.64.0/19       4637 4694 4713 2914 109
64.104.96.0/19       4637 109
64.104.142.0/24      4637 109
64.104.160.0/19      4637 109
64.104.192.0/18      4637 109
128.107.0.0/16       4637 3356 109
144.254.0.0/16       4637 1239 109
161.44.0.0/16        4637 701 109
171.68.0.0/14        4637 3356 109
192.31.7.0/24        4637 3356 109
192.118.76.0/22      4637 3491 9116 109
192.122.173.0/24     4637 3356 109
192.122.174.0/24     4637 3356 109
192.135.240.0/21     4637 3356 109
192.135.250.0/24     4637 701 109
198.92.0.0/18        4637 3356 109
198.133.219.0/24     4637 3356 109
198.135.4.0/22       4637 3356 109
204.69.198.0/23      4637 3356 109
204.69.200.0/24      4637 3356 109
```

Advertisements that are fragments of the original RIR allocation (more specifics) originated by this AS.

```
AS109   18 More Specifics   29 Total Advertisements   CISCO-EU-109 Cisco Systems Global ASN - ARIN Assigned
```

Aggregation Potential



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 router 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.ripe.net/ripencc/pub-services/db/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

Preparing the Network

- We want to deploy BGP now…
- BGP will be used therefore an ASN is required
- If multihoming to different ISPs is intended in the near future, a public ASN should be obtained:

    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

- **Will look at two examples of BGP deployment:**
  - **Example One: network is only static routes**
  - **Example Two: network is currently running an IGP**

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
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

  Backbone 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)

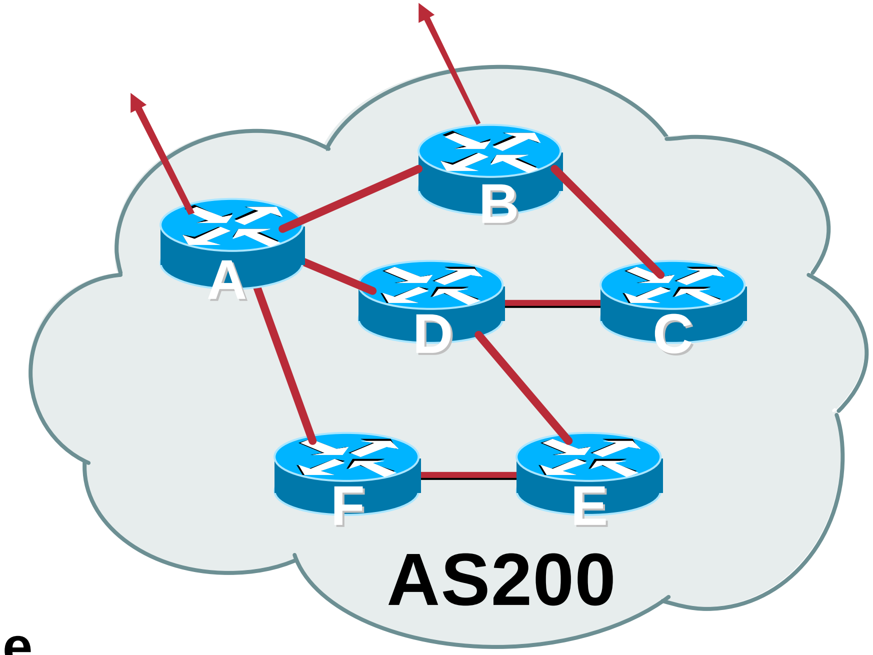
  Static/Hosting LANs

  Customer assigned address space

  Anything else not listed in the previous slide

Preparing the Network
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***

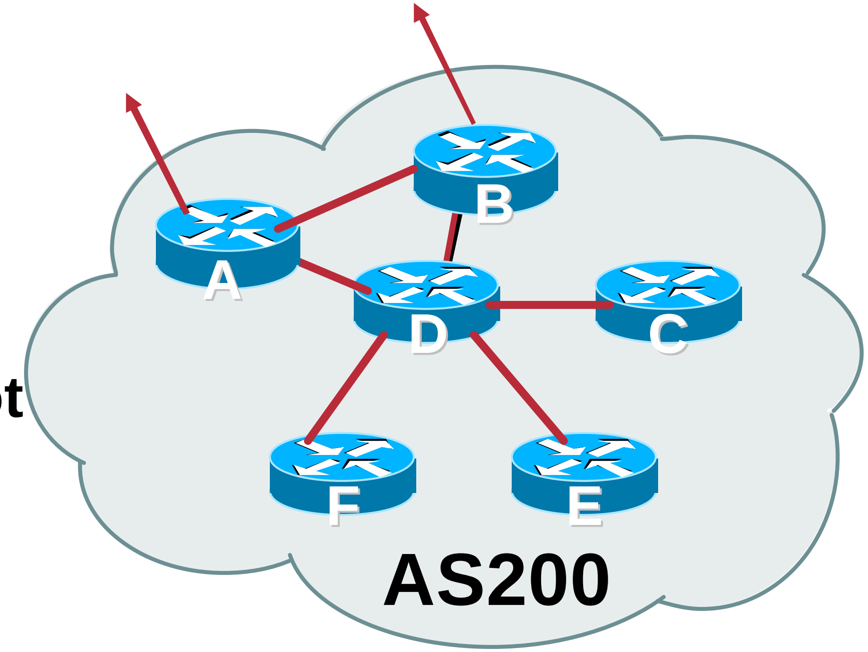


Preparing the Network
iBGP (Transit Path)

- ***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 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, the 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 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 routes 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 either by IGP or by 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 routes 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 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**
  - **Use RIPE-229 parameters, or something even weaker**
  - **Don't 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**

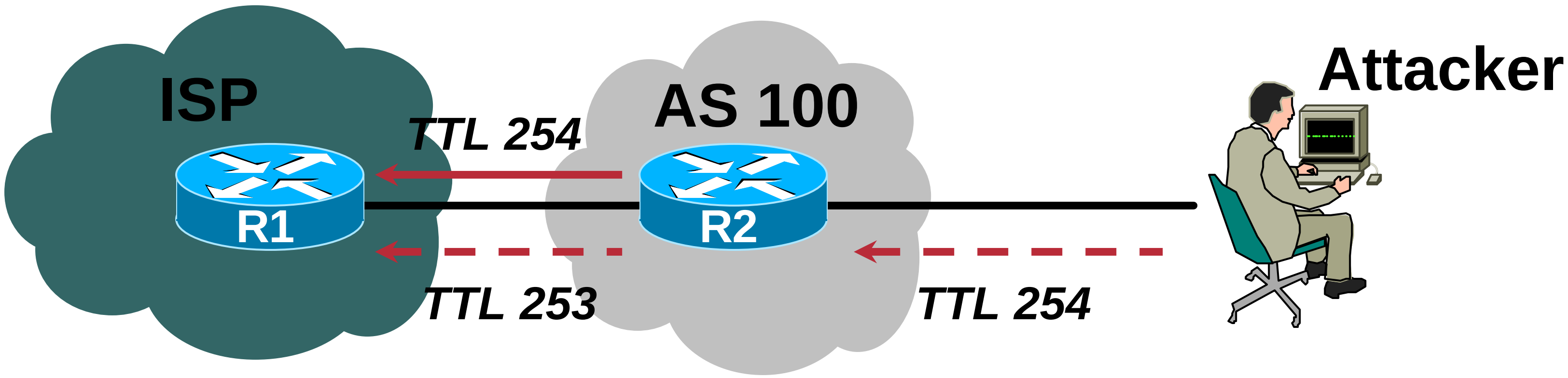
BGP TTL "hack"

- **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

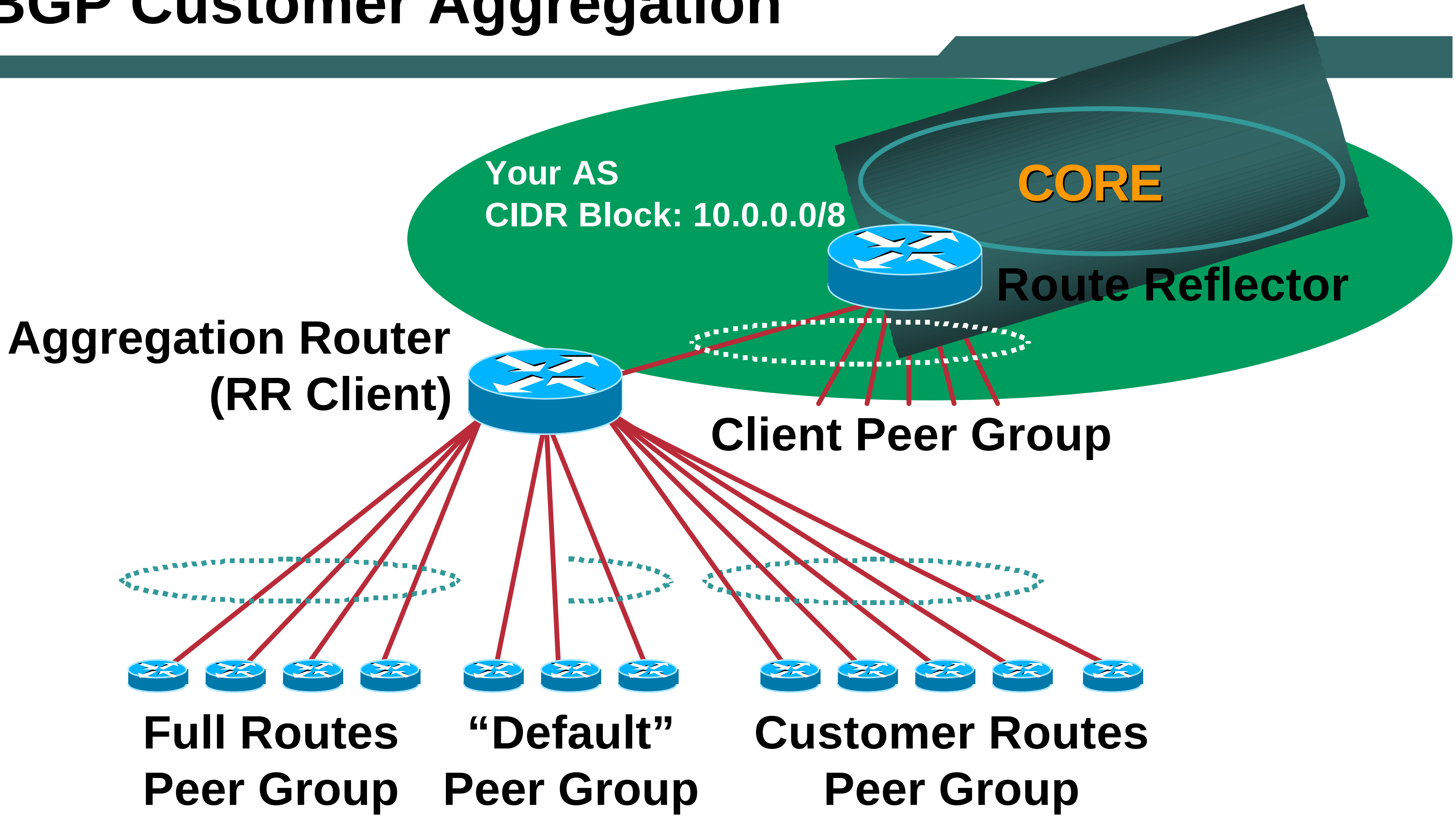
  cust-full —send full Internet routes

- **Identify routes via communities e.g.**

  100:4100=customers; 100:4500=peers

- **Apply passwords per neighbour**
- **Apply inbound & outbound prefix filters per neighbour**

Using Communities: BGP Customer Aggregation



Apply passwords and in/outbound prefix filters 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

BGP for Internet Service Providers

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

BGP Techniques for Internet Service Providers

Philip Smith <pfs@cisco.com>

NANOG 31

San Francisco

23-25 May 2004