



BGP Techniques for Providers

Philip Smith <pfs@cisco.com>

QUESTnet 2006

11th - 14th July

Presentation Slides

- **Are available on**

<ftp://ftp-eng.cisco.com>

[/pfs/seminars/QUESTnet2006-BGP-Tutorial.pdf](#)

And will be on the QUESTnet 2006 website

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

BGP Techniques for Providers

- **BGP Basics**
- **Scaling BGP**
- **Deploying BGP**
- **Multihoming Basics**
- **BGP “Traffic Engineering”**
- **BGP Configuration Tips**



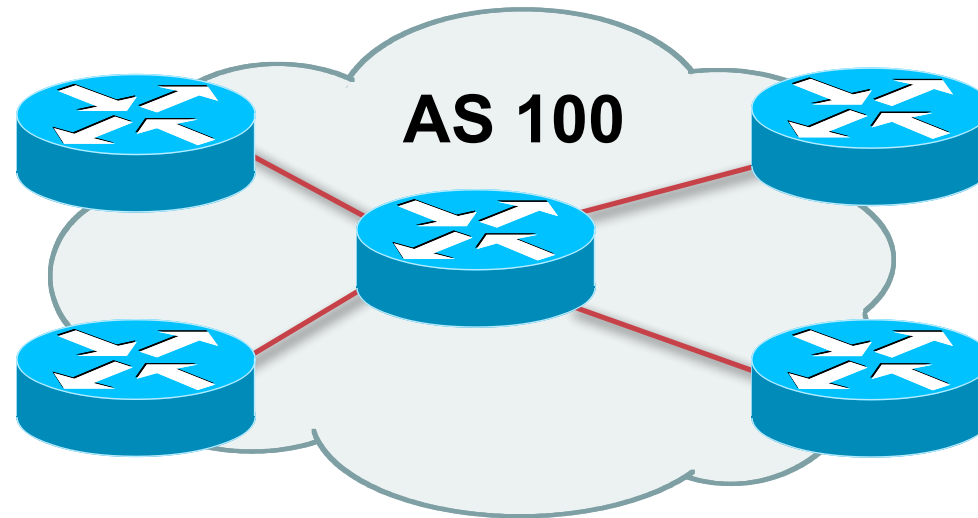
BGP Basics

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

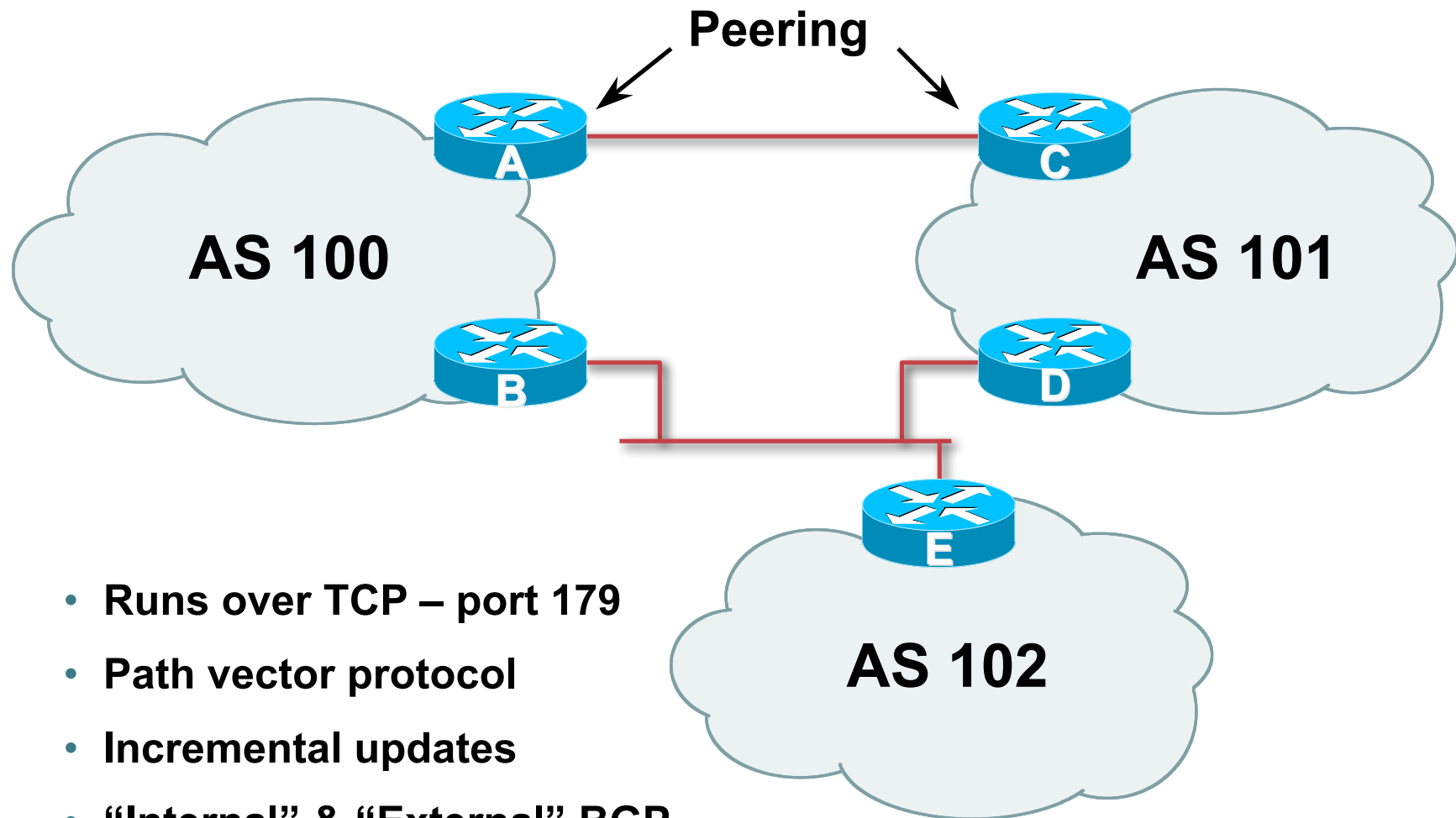
Autonomous System Number (ASN)

- **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 41983 have been made to the RIRs

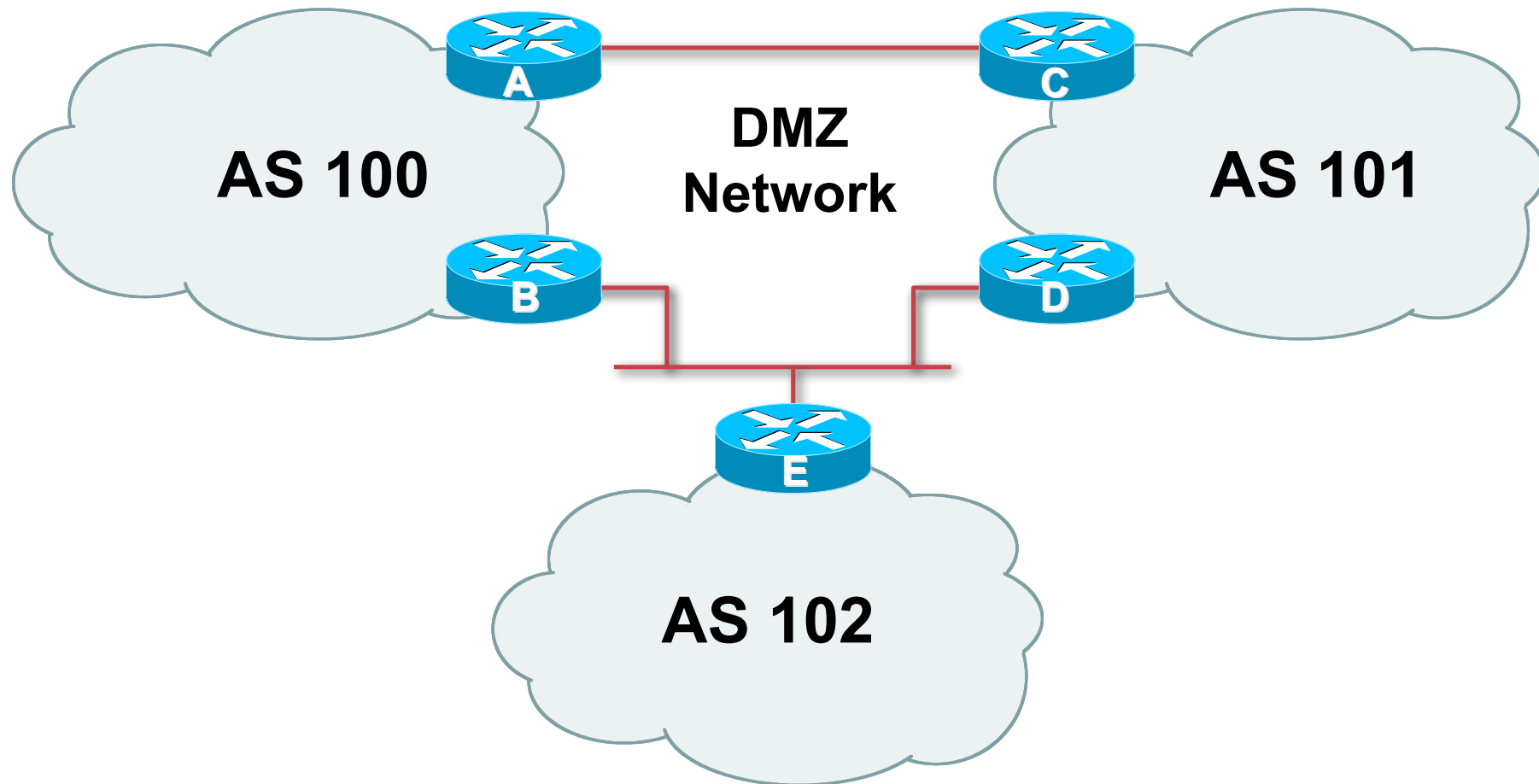
Of these, around 22500 are visible on the Internet

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

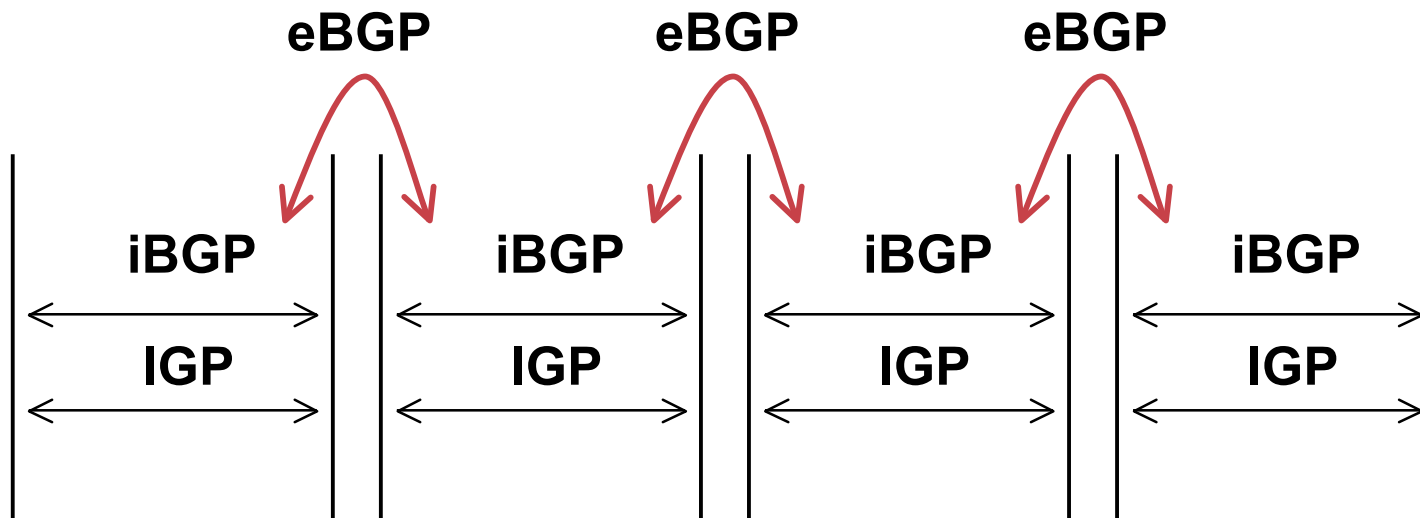
- **Learns multiple paths via internal and external BGP speakers**
- **Picks the best path and installs in the forwarding table**
- **Best path is sent to external BGP neighbours**
- **Policies applied by influencing the best path selection**

eBGP & iBGP

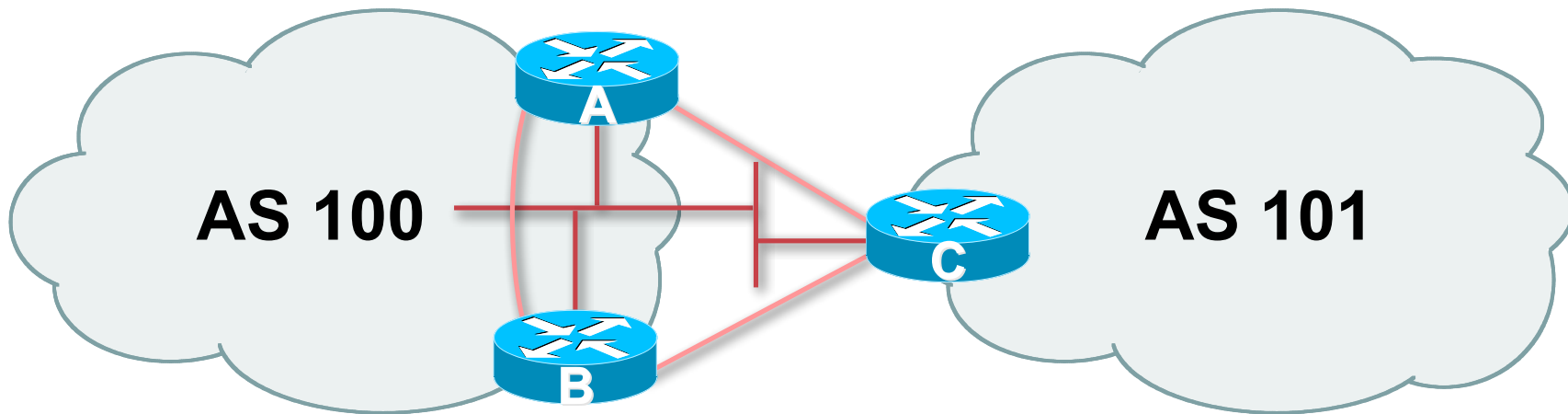
- **BGP used internally (iBGP) and externally (eBGP)**
- **iBGP used to carry**
 - some/all Internet prefixes across ISP backbone**
 - ISP's customer prefixes**
- **eBGP used to**
 - exchange prefixes with other ASes**
 - implement routing policy**

BGP/IGP model used in ISP networks

- **Model representation**



External BGP Peering (eBGP)

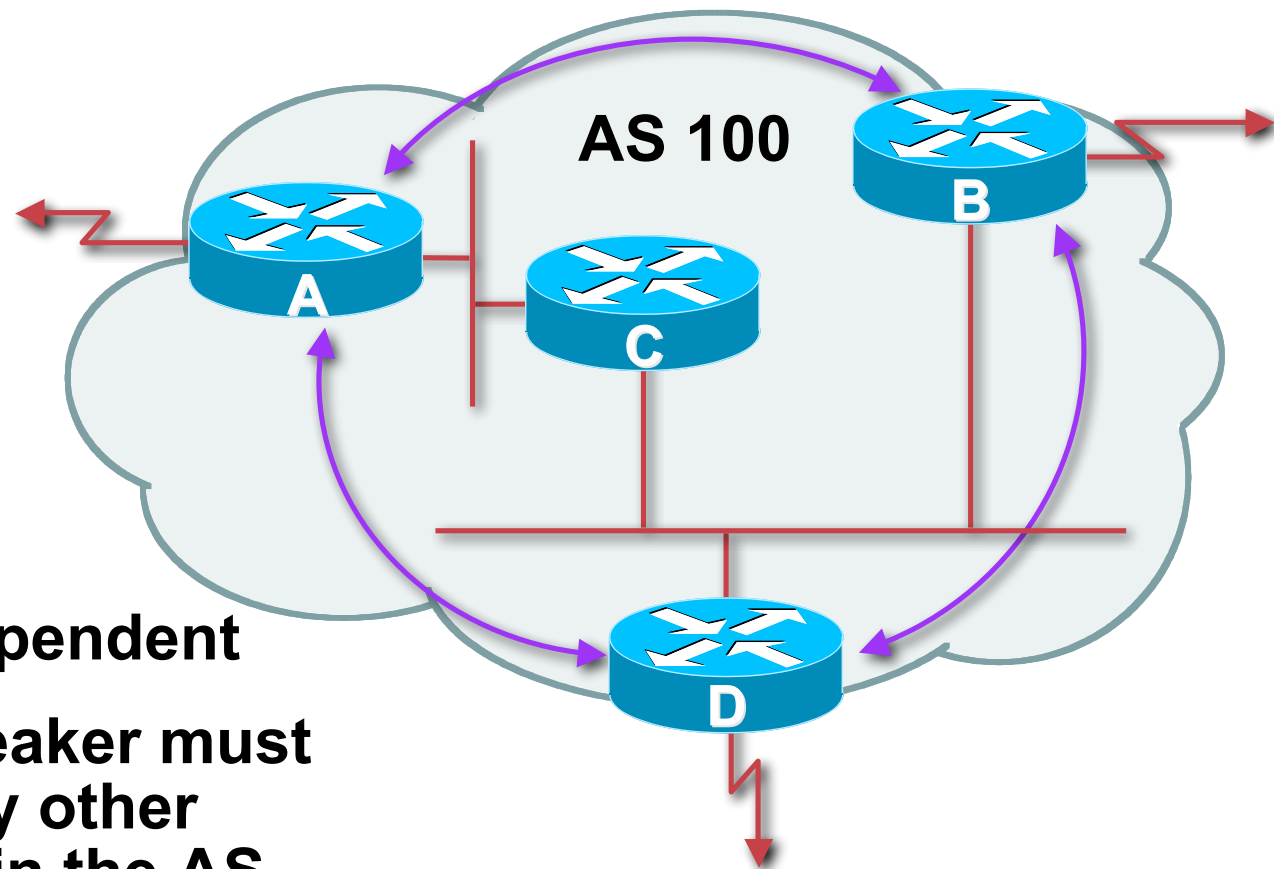


- Between BGP speakers in different AS
- Should be directly connected
- **Never** run an IGP between eBGP peers

Internal BGP (iBGP)

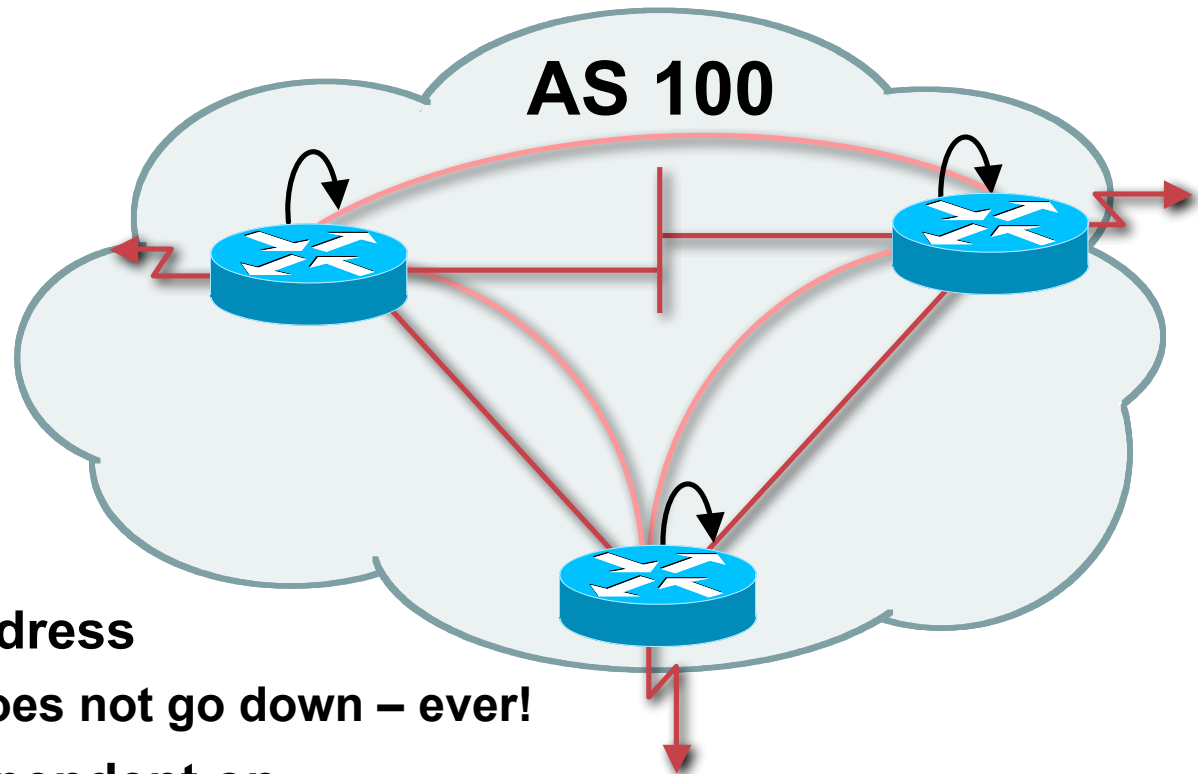
- **BGP peer within the same AS**
- **Not required to be directly connected**
IGP takes care of inter-BGP speaker connectivity
- **iBGP speakers need to be fully meshed**
they originate connected networks
they do not pass on prefixes learned from other iBGP speakers

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

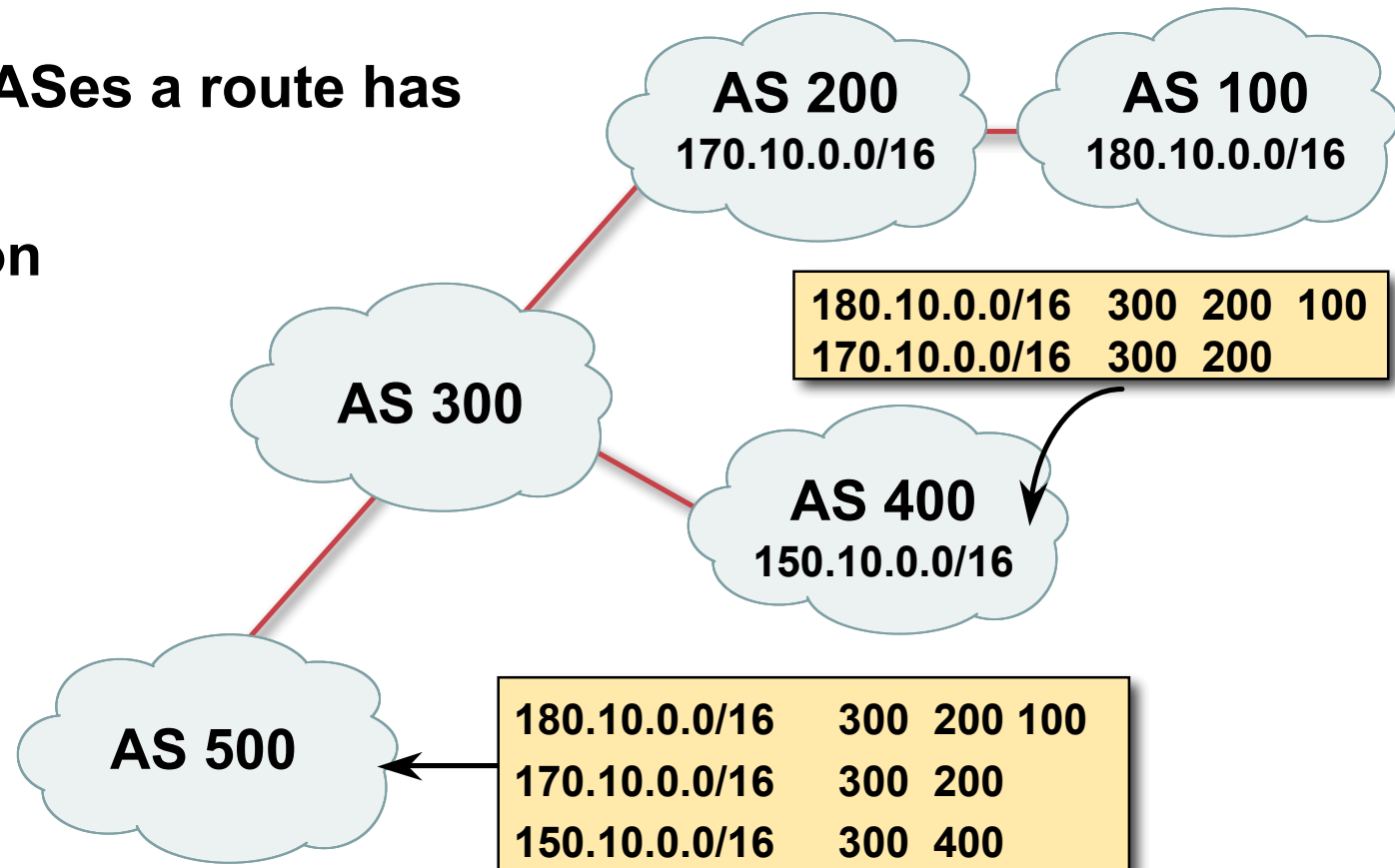


BGP Attributes

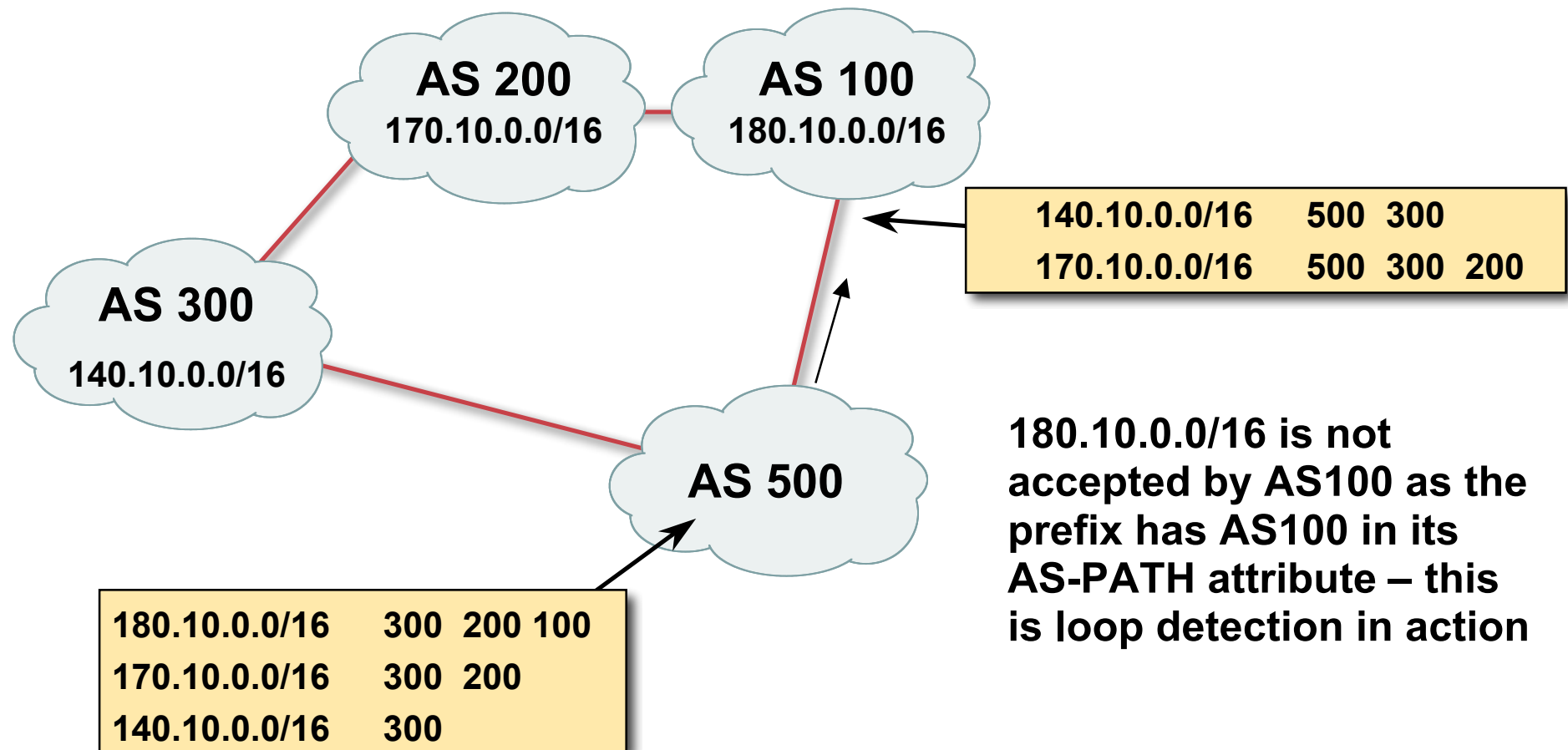
Information about BGP

AS-Path

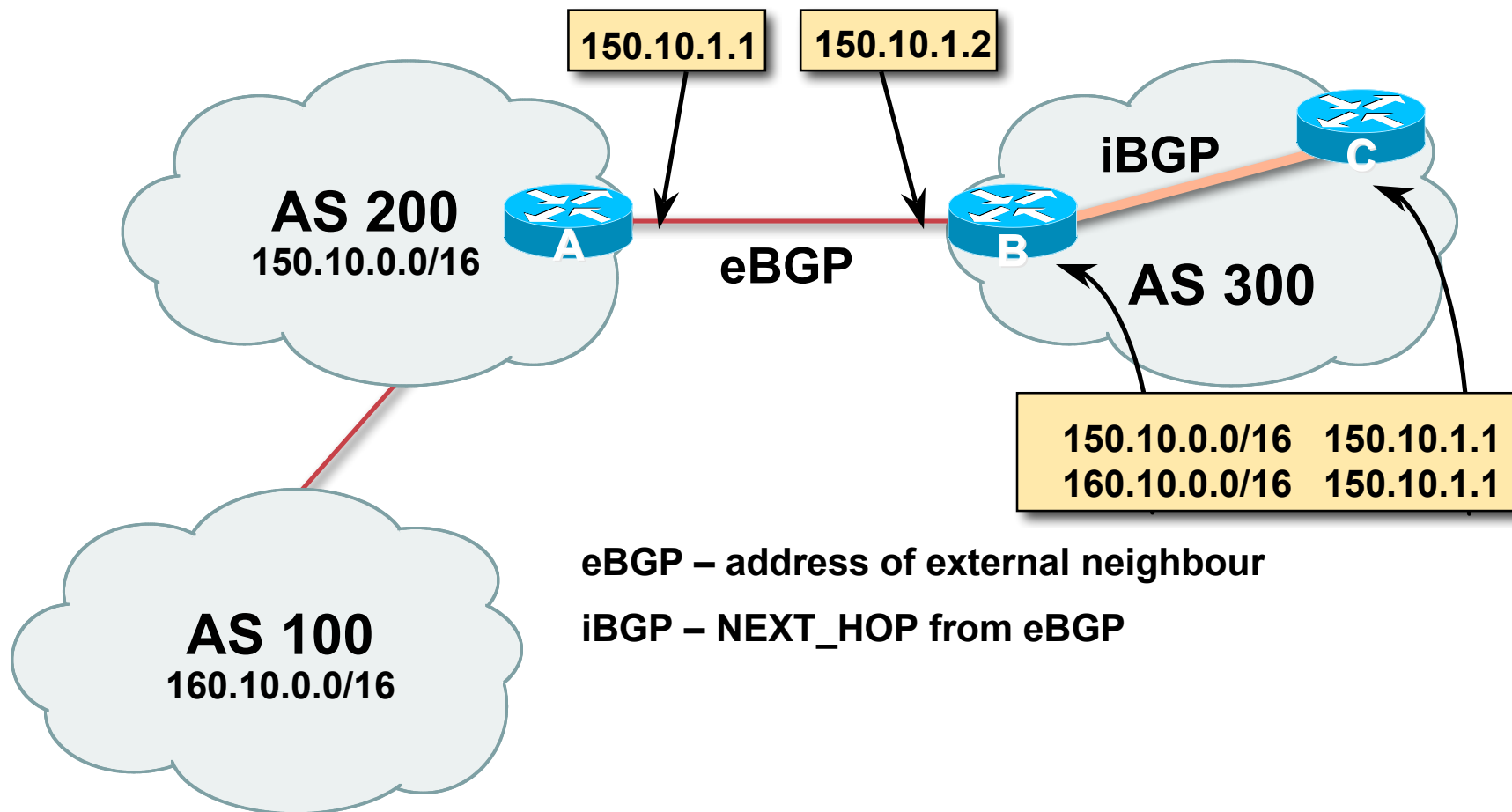
- Sequence of ASes a route has traversed
- Loop detection
- Apply policy



AS-Path loop detection



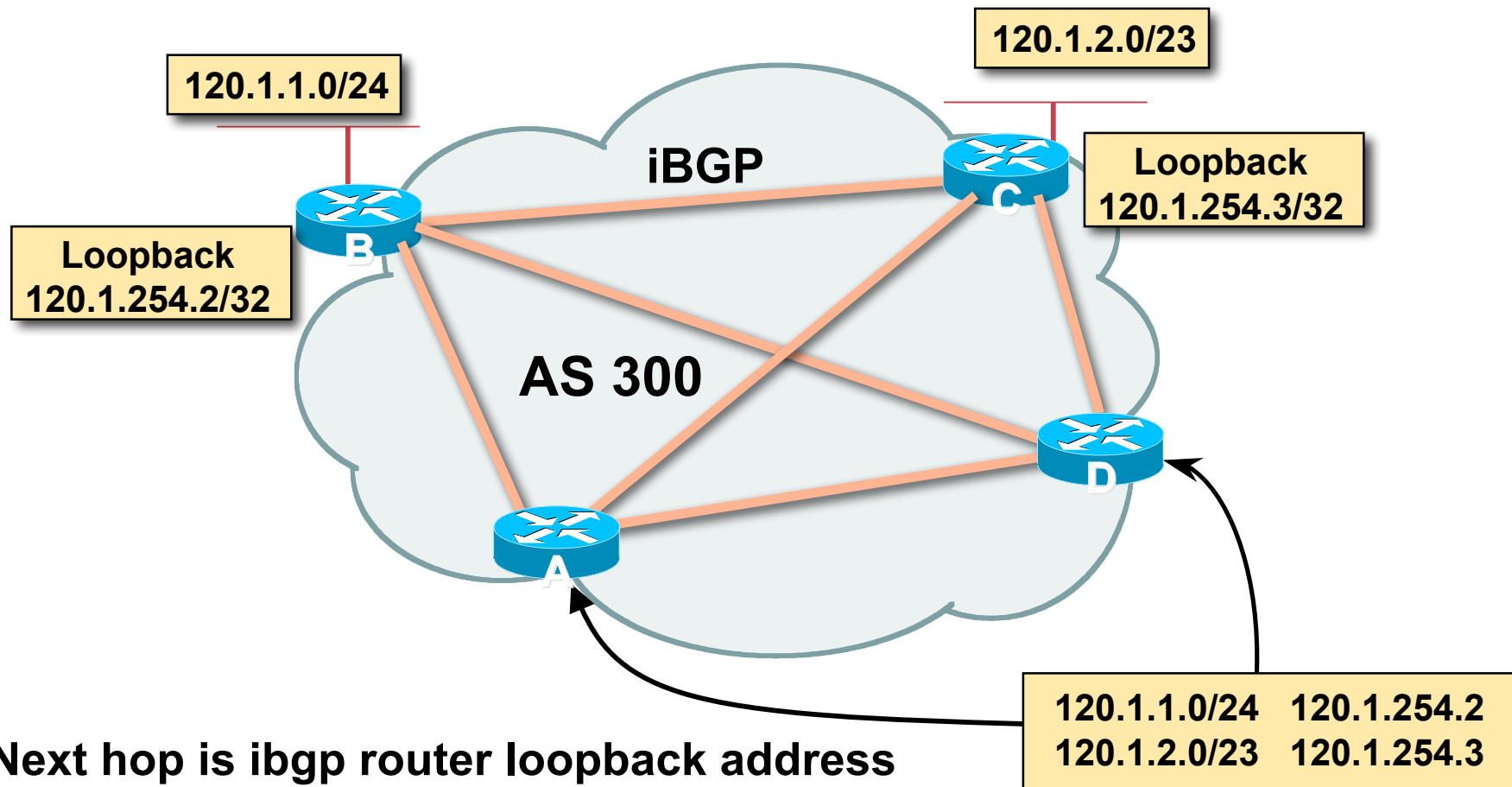
Next Hop



eBGP – address of external neighbour

iBGP – NEXT_HOP from eBGP

iBGP Next Hop



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**

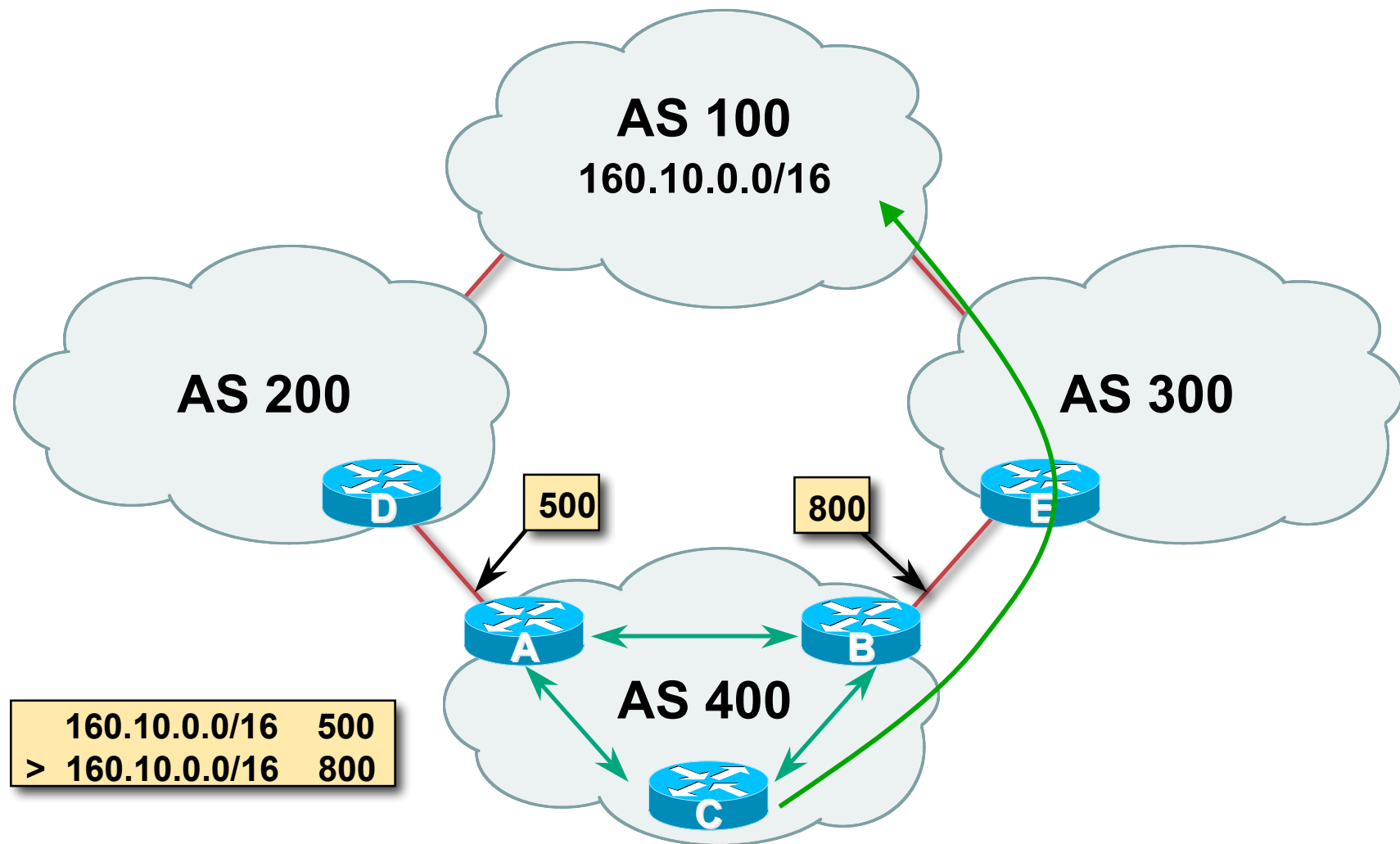
Origin

- **Conveys the origin of the prefix**
- **Historical** attribute
 - Was used in transition from EGP to BGP
- **Influences best path selection**
- **Three values: IGP, EGP, incomplete**
 - IGP – generated by BGP network statement
 - EGP – generated by EGP
 - incomplete – redistributed from another routing protocol

Aggregator

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

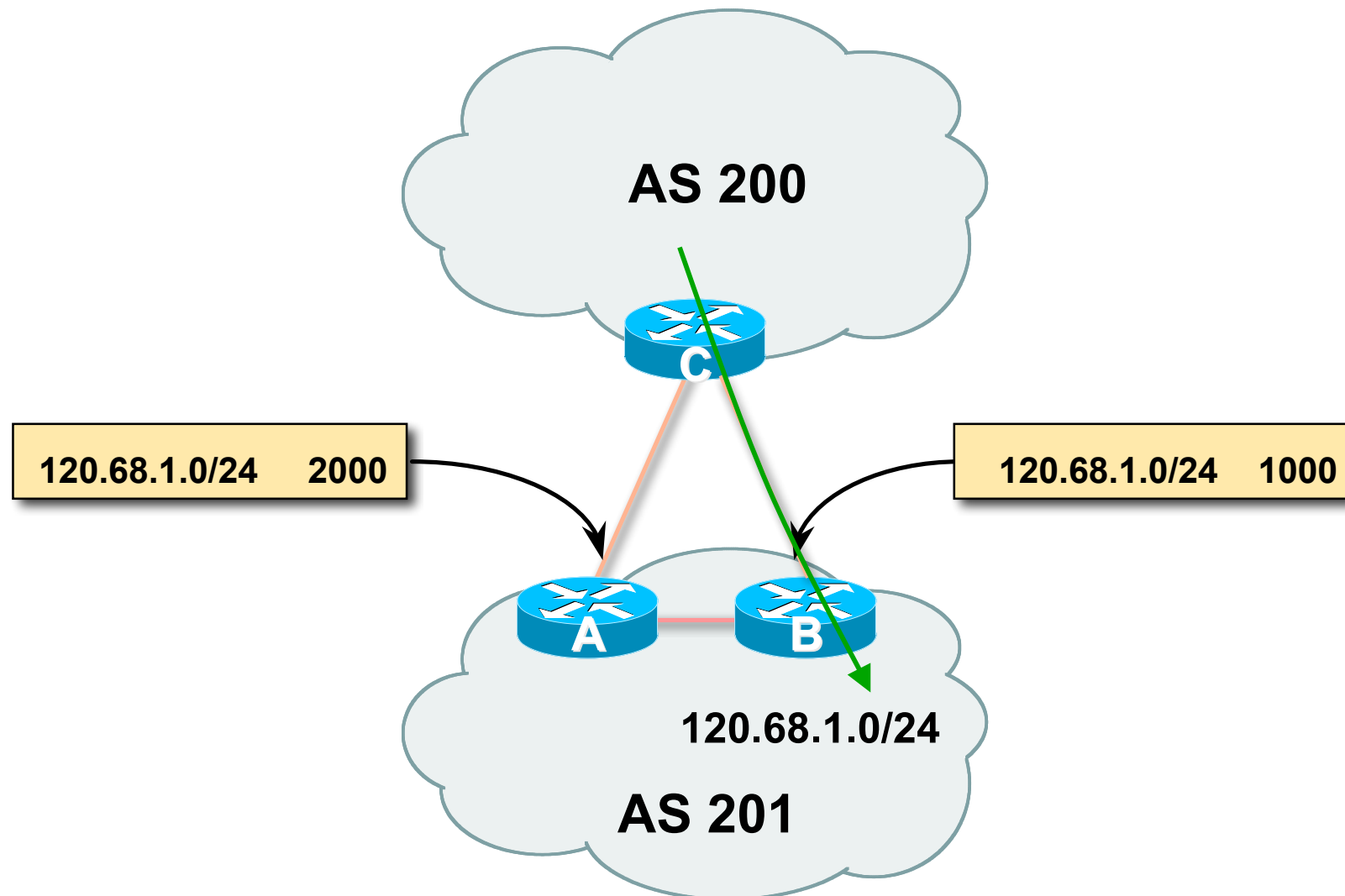
Local Preference



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

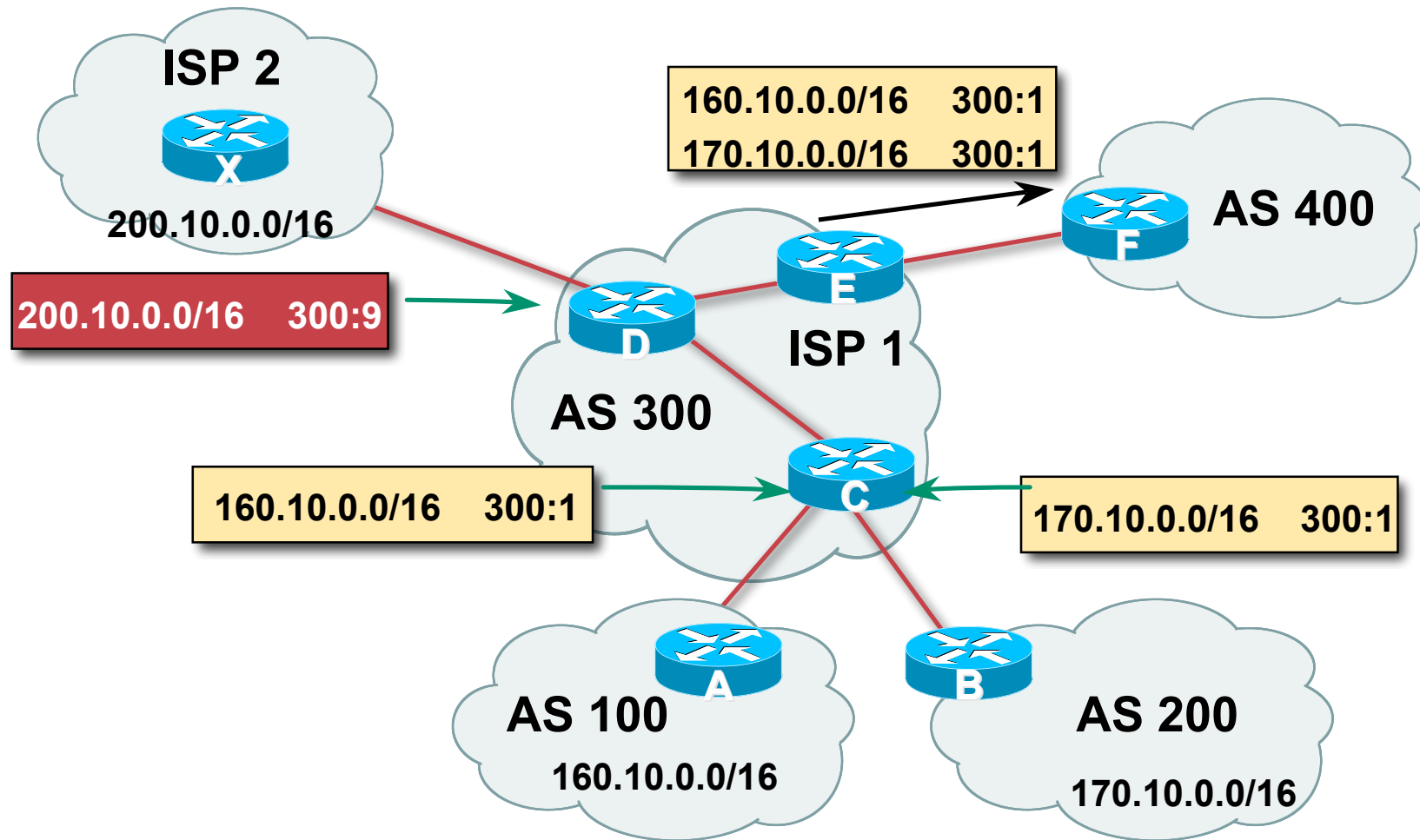
“metric confusion”

- **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



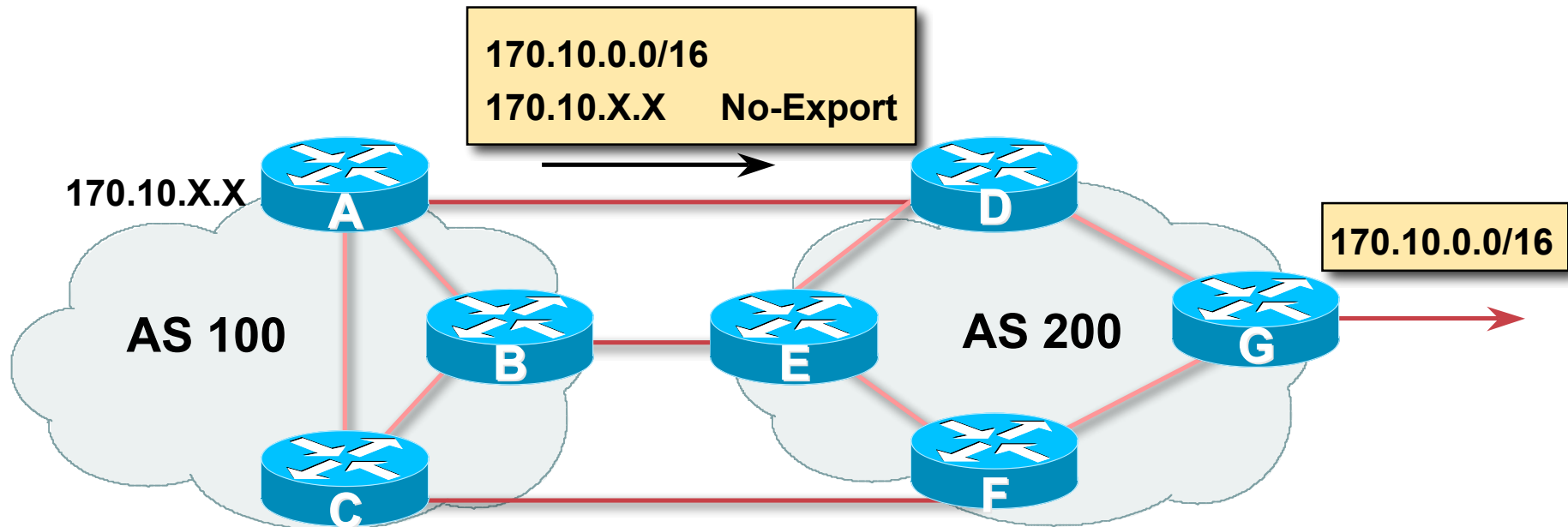
Well-Known Communities

- **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

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 Techniques for Providers

- BGP Basics
- **Scaling BGP**
- Deploying BGP
- Multihoming Basics
- BGP “Traffic Engineering”
- BGP Configuration Tips



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?**
- **Route Refresh**
- **Route Reflectors**
- **(Confederations)**



Dynamic Reconfiguration

Route Refresh

Route Refresh

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

Route Refresh Capability

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

Dynamic Reconfiguration

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

Consider the impact to be equivalent to a router reboot



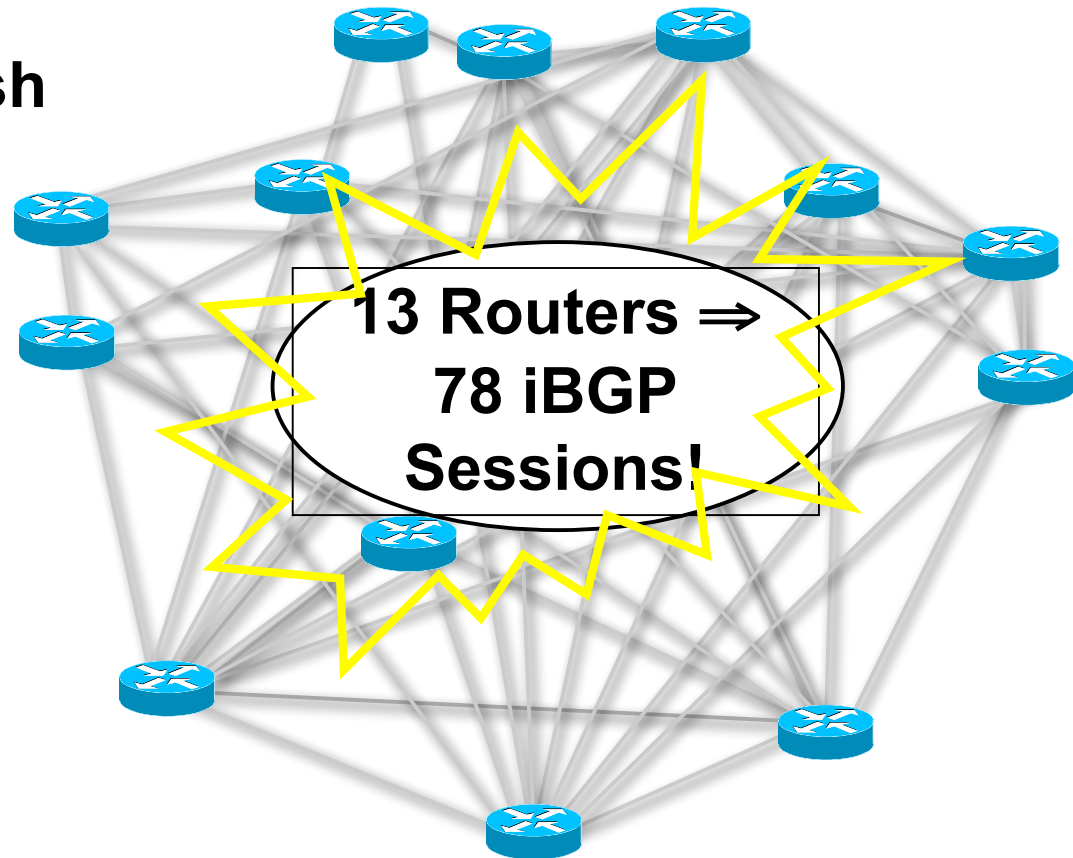
Route Reflectors

Scaling the iBGP mesh

Scaling iBGP mesh

Avoid $\frac{1}{2}n(n-1)$ iBGP mesh

**$n=1000 \Rightarrow$ nearly
half a million
ibgp sessions!**

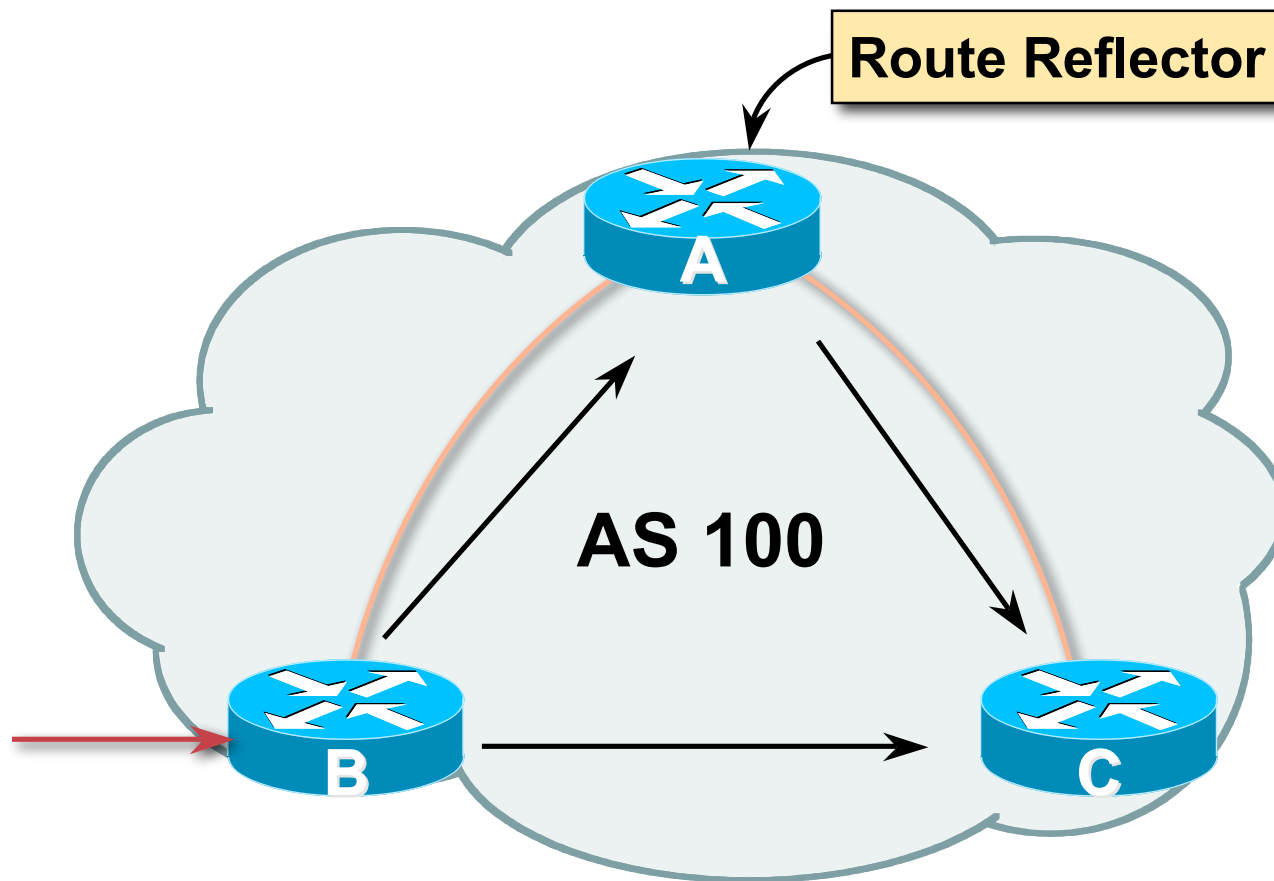


Two solutions

Route reflector – simpler to deploy and run

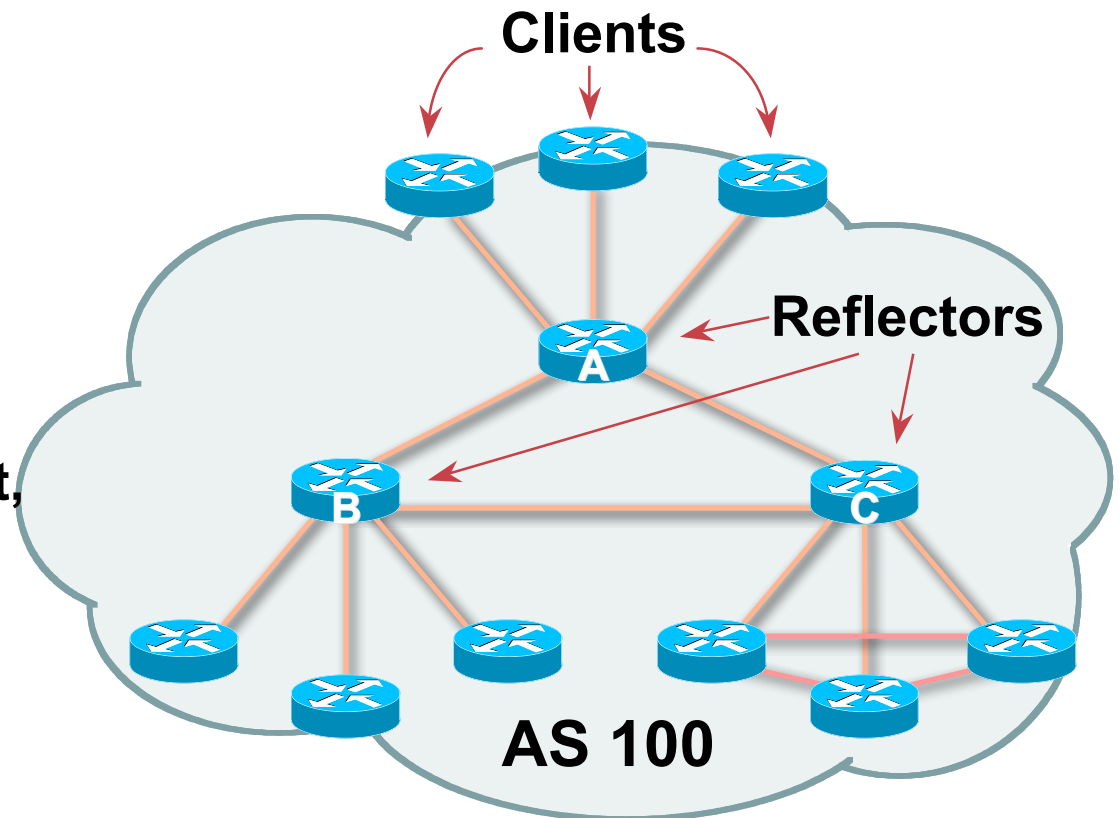
Confederation – more complex, has corner case advantages

Route Reflector: Principle



Route Reflector

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



Route Reflector Topology

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

Route 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!**

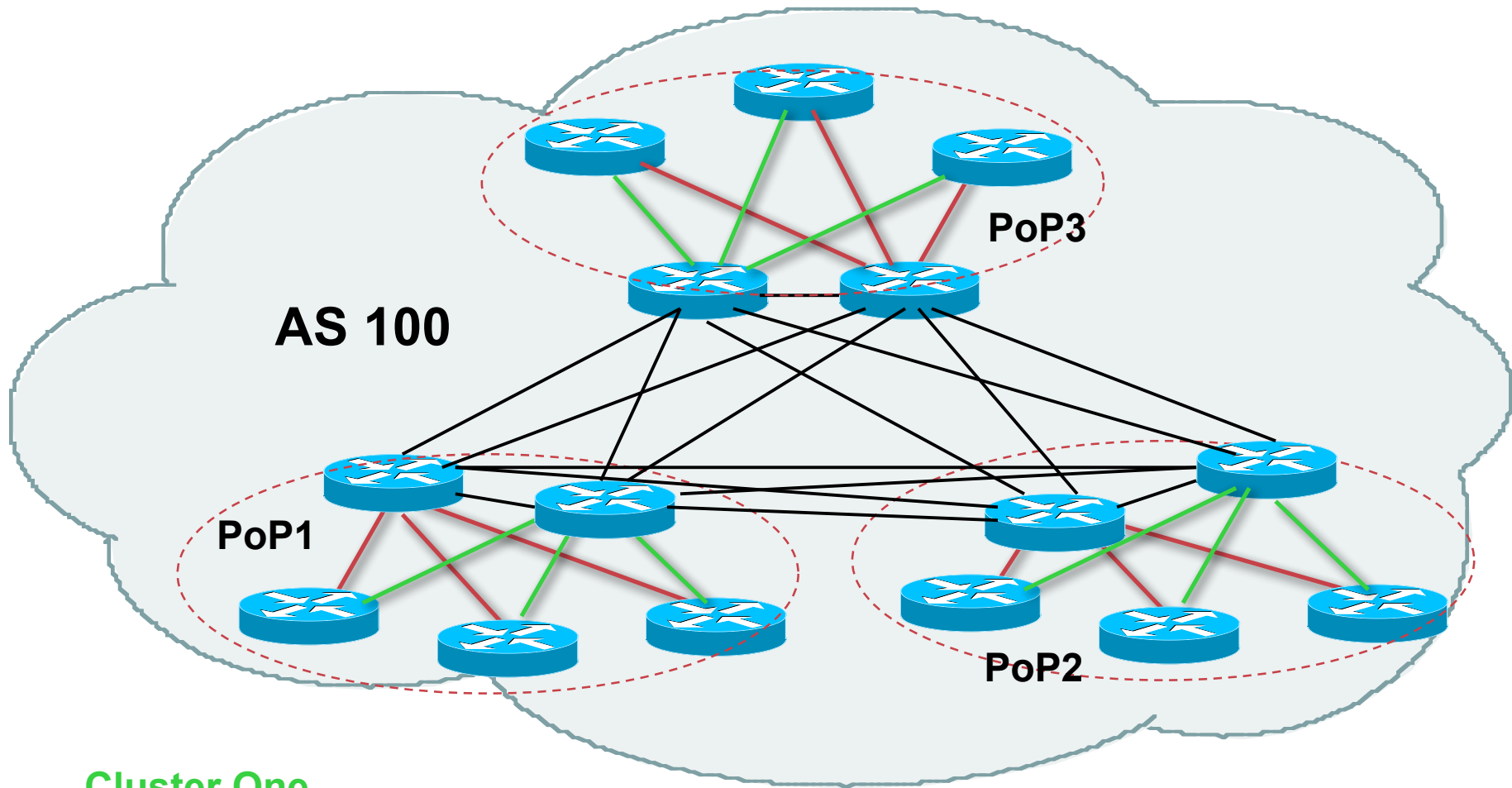
All RRs in the cluster **must** have the same cluster-id (otherwise it is a different cluster)

- **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

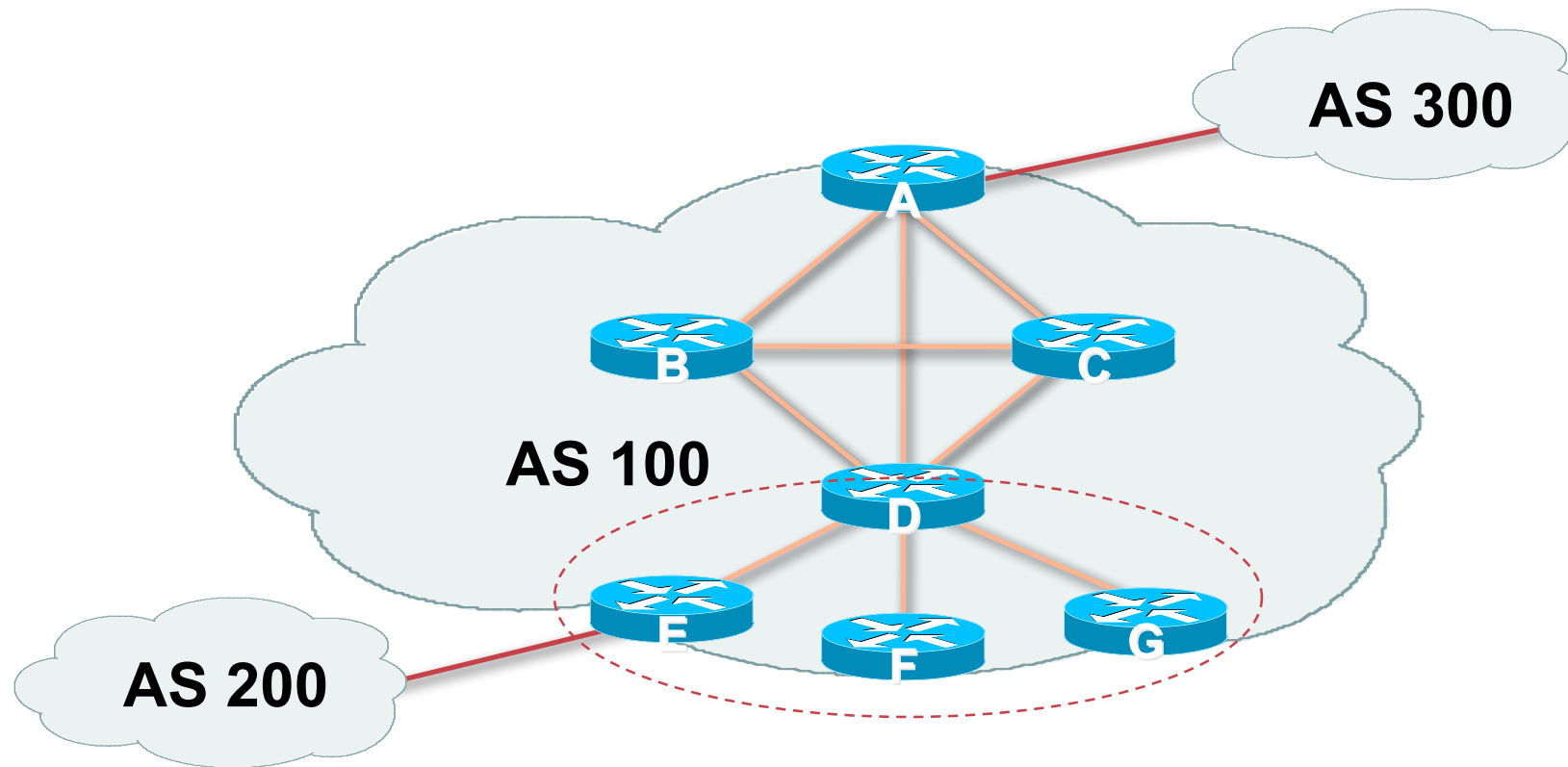
Core routers are RR for the PoP

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 Scaling Techniques

- **Route Refresh**
Use should be mandatory
- **Route Reflectors**
The only way to scale iBGP mesh

BGP Techniques for Providers

- **BGP Basics**
- **Scaling BGP**
- **Deploying BGP**
- **Multihoming Basics**
- **BGP “Traffic Engineering”**
- **BGP Configuration Tips**



Deploying BGP

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**



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

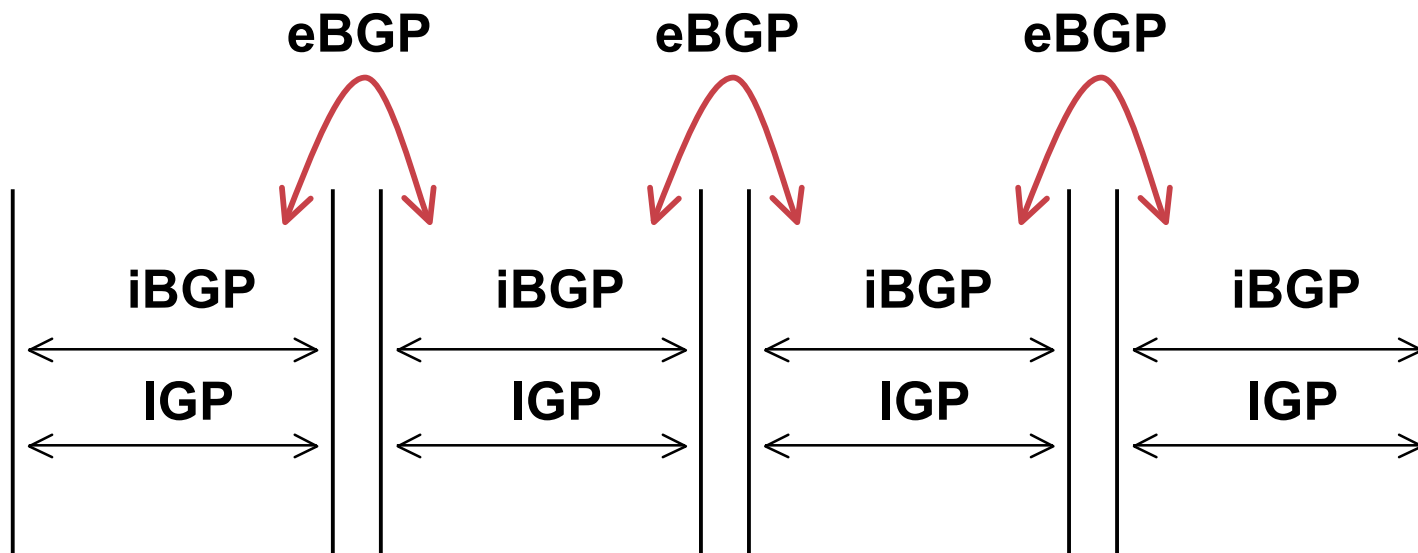
design goal is to **minimise** number of prefixes in IGP to aid scalability and rapid convergence

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**

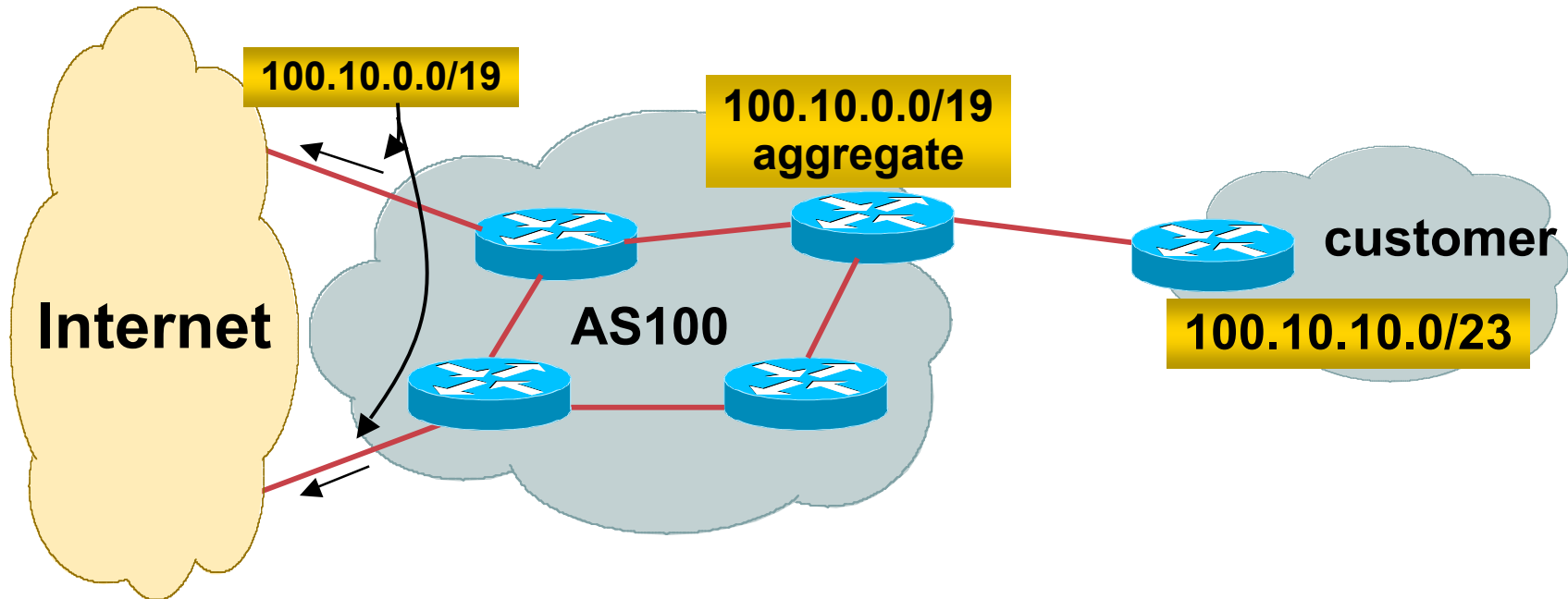
Aggregation

- 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

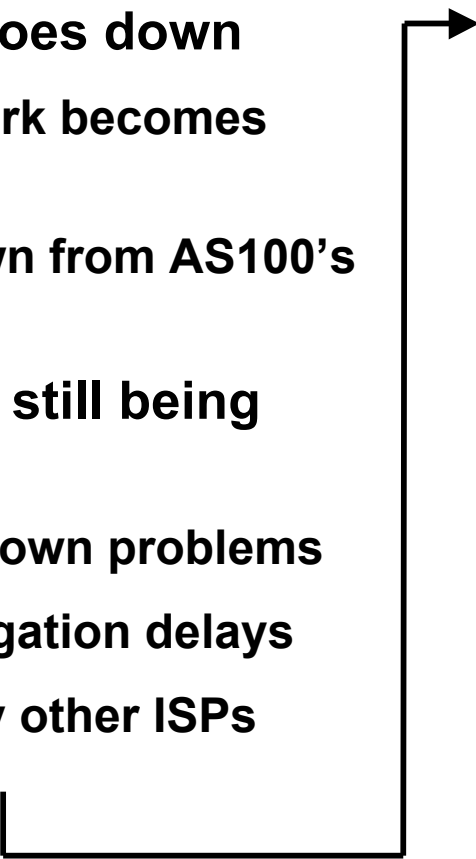
- **ISPs who don't and won't aggregate are held in poor regard by community**
- **Registries publish their minimum allocation size**
Either a /21 or a /22 depending on RIR
- **No real reason to see anything longer than a /22 prefix in the Internet**
BUT there are currently >102000 /24s!

Aggregation – Example

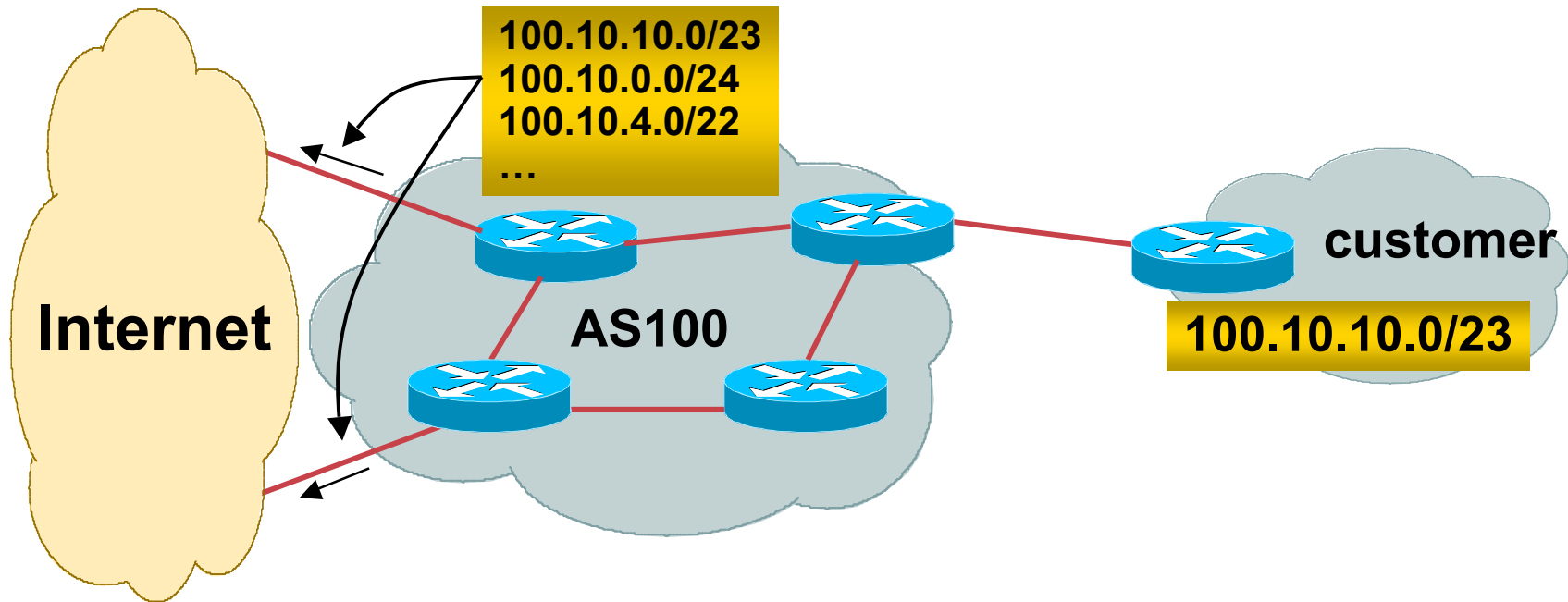


- **Customer has /23 network assigned from AS100's /19 address block**
- **AS100 announced /19 aggregate to the Internet**

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 (July 2006)

- **Current Internet Routing Table Statistics**

BGP Routing Table Entries	191458
----------------------------------	---------------

Prefixes after maximum aggregation	105432
---	---------------

Unique prefixes in Internet	93726
------------------------------------	--------------

Prefixes smaller than registry alloc	94718
---	--------------

/24s announced	103595
-----------------------	---------------

only 5729 /24s are from 192.0.0.0/8

ASes in use	22583
--------------------	--------------



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

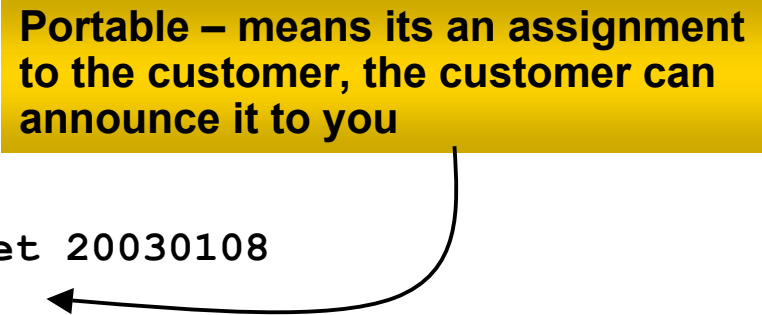
The tool: **whois** -h whois.apnic.net x.x.x.0/24

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



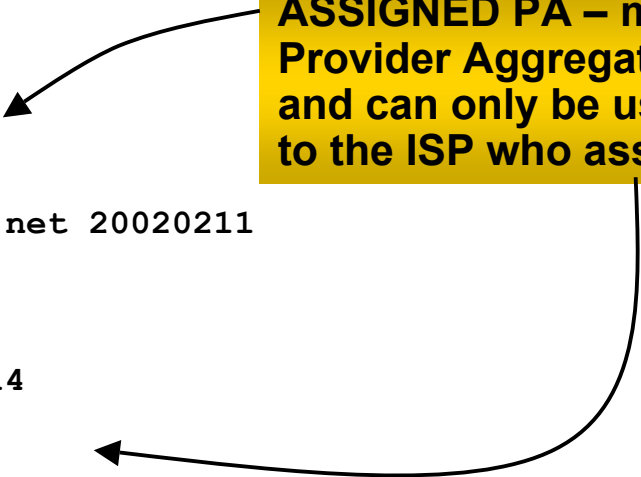
Receiving Prefixes: From Customers

- 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.com/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

Initial Assumptions

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

Preparing the Network

First Step: IGP

- **Decide on 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

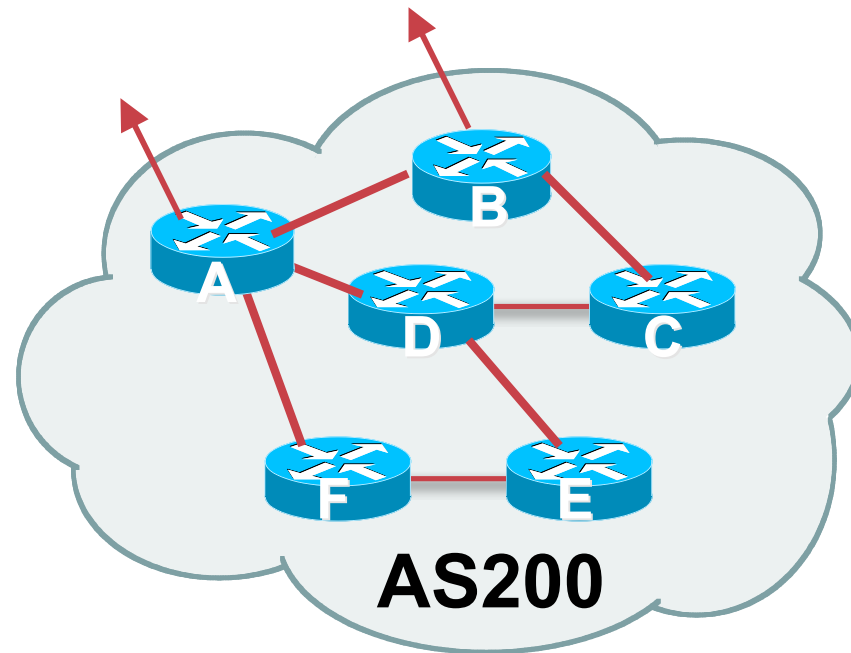
IGP (cont)

- **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

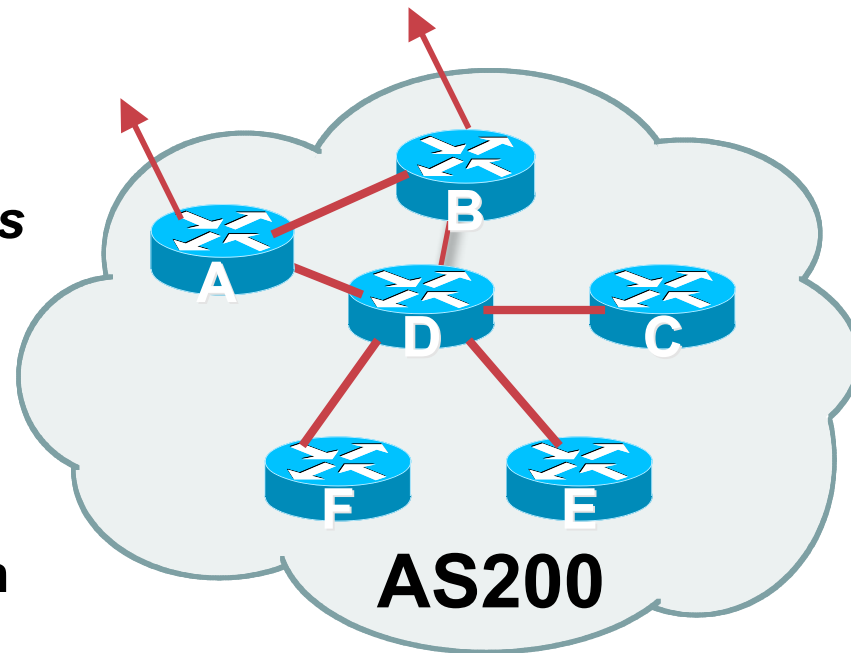
- **Second step is to configure the local network to use iBGP**
- **iBGP can run on**
 - all routers, or
 - a subset of routers, or
 - just on the upstream edge
- ***iBGP must run on all routers which are in the transit path between external connections***



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 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**

BGP Techniques for Providers

- BGP Basics
- Scaling BGP
- Deploying BGP
- **Multihoming Basics**
- BGP “Traffic Engineering”
- BGP Configuration Tips



Multihoming: Definitions & Options

What does it mean, what do we need, and how do we do it?

Multihoming Definition

- **More than one link external to the local network**
 - two or more links to the same ISP
 - two or more links to different ISPs
- **Usually *two* external facing routers**
 - one router gives link and provider redundancy only

AS Numbers

- **An Autonomous System Number is required by BGP**
- **Obtained from upstream ISP or Regional Registry (RIR)**
 - AfriNIC, APNIC, ARIN, LACNIC, RIPE NCC**
- **Necessary when you have links to more than one ISP or to an exchange point**
- **16 bit integer, ranging from 1 to 65534**
 - Zero and 65535 are reserved**
 - 64512 through 65534 are called Private ASNs**

Private-AS – Application

- **Applications**

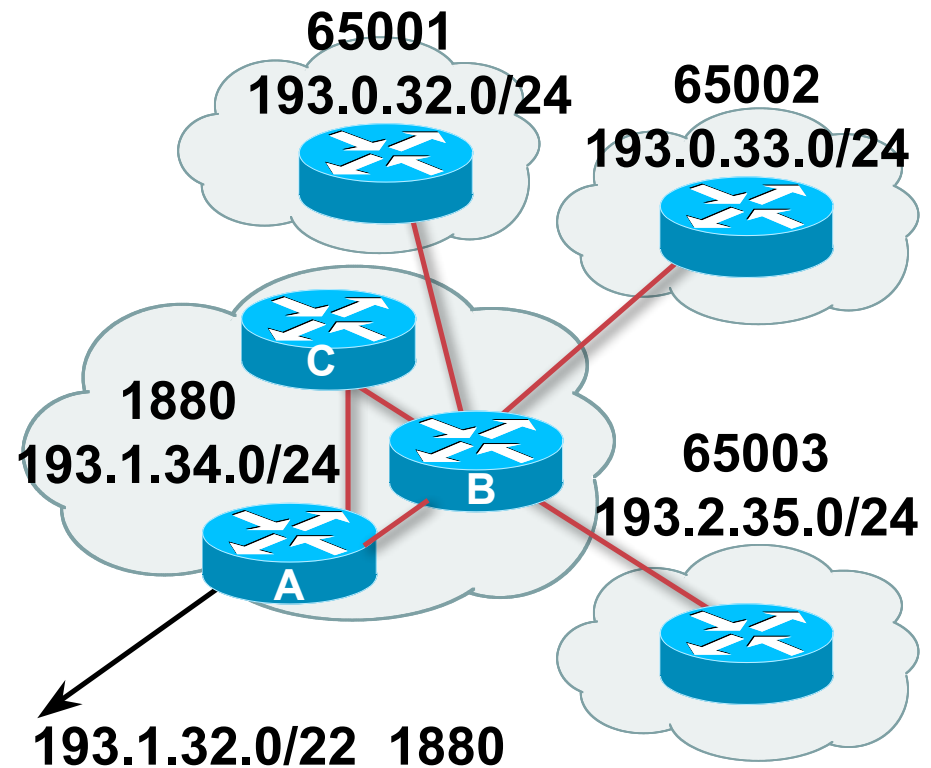
An ISP with customers multihomed on their backbone (RFC2270)

-or-

A corporate network with several regions but connections to the Internet only in the core

-or-

Within a BGP Confederation



Private-AS – Removal

- **Private ASNs MUST be removed from all prefixes announced to the public Internet**
 - Include configuration to remove private ASNs in the eBGP template
- **As with RFC1918 address space, private ASNs are intended for internal use**
 - They should not be leaked to the public Internet
- **Cisco IOS**
 - neighbor x.x.x.x remove-private-AS**

Policy Tools

- **Local preference**
outbound traffic flows
- **Metric (MED)**
inbound traffic flows (local scope)
- **AS-PATH prepend**
inbound traffic flows (Internet scope)
- **Communities**
specific inter-provider peering

Originating Prefixes: Assumptions

- **MUST** announce assigned address block to Internet
- **MAY** also announce subprefixes – reachability is not guaranteed
- **Current RIR minimum allocation is /21**

Several ISPs filter RIR blocks on this boundary

Several ISPs filter the rest of address space according to the IANA assignments

This activity is called “Net Police” by some

Originating Prefixes

- **RIRs publish the minimum allocation sizes per /8 address block**

AfriNIC: www.afrinic.net/docs/policies/afpol-v4200407-000.htm

APNIC: www.apnic.net/db/min-alloc.html

ARIN: www.arin.net/reference/ip_blocks.html

LACNIC: lacnic.net/en/registro/index.html

RIPE NCC: www.ripe.net/ripe/docs/smallest-alloc-sizes.html

Note that AfriNIC only publishes its current minimum allocation size, not the allocation size for its address blocks

- **IANA publishes the address space it has assigned to end-sites and allocated to the RIRs:**

www.iana.org/assignments/ipv4-address-space

- **Several ISPs use this published information to filter prefixes on:**

What should be routed (from IANA)

The minimum allocation size from the RIRs

“Net Police” prefix list issues

- meant to “punish” ISPs who pollute the routing table with specifics rather than announcing aggregates
- impacts legitimate multihoming especially at the Internet’s edge
- impacts regions where domestic backbone is unavailable or costs \$\$\$ compared with international bandwidth
- hard to maintain – requires updating when RIRs start allocating from new address blocks
- **don’t do it unless consequences understood and you are prepared to keep the list current**

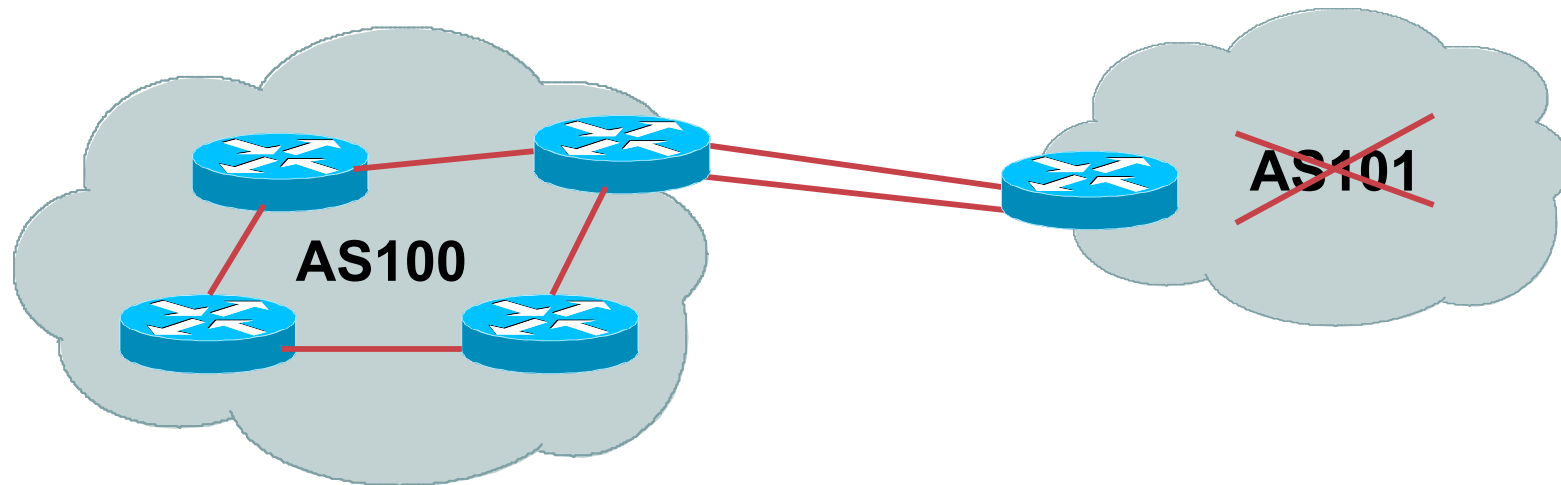
Consider using the Project Cymru bogon BGP feed

<http://www.cymru.com/BGP/bogon-rs.html>

Multihoming Scenarios

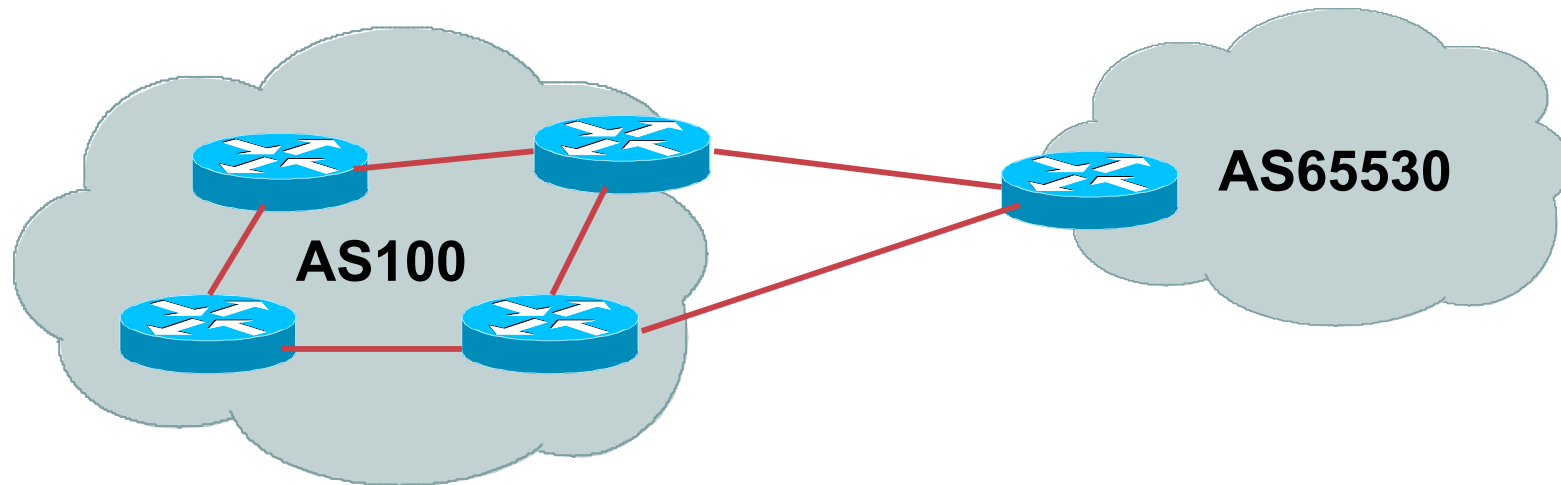
- **Stub network**
- **Multi-homed stub network**
- **Multi-homed network**
- **Load-balancing**

Stub Network



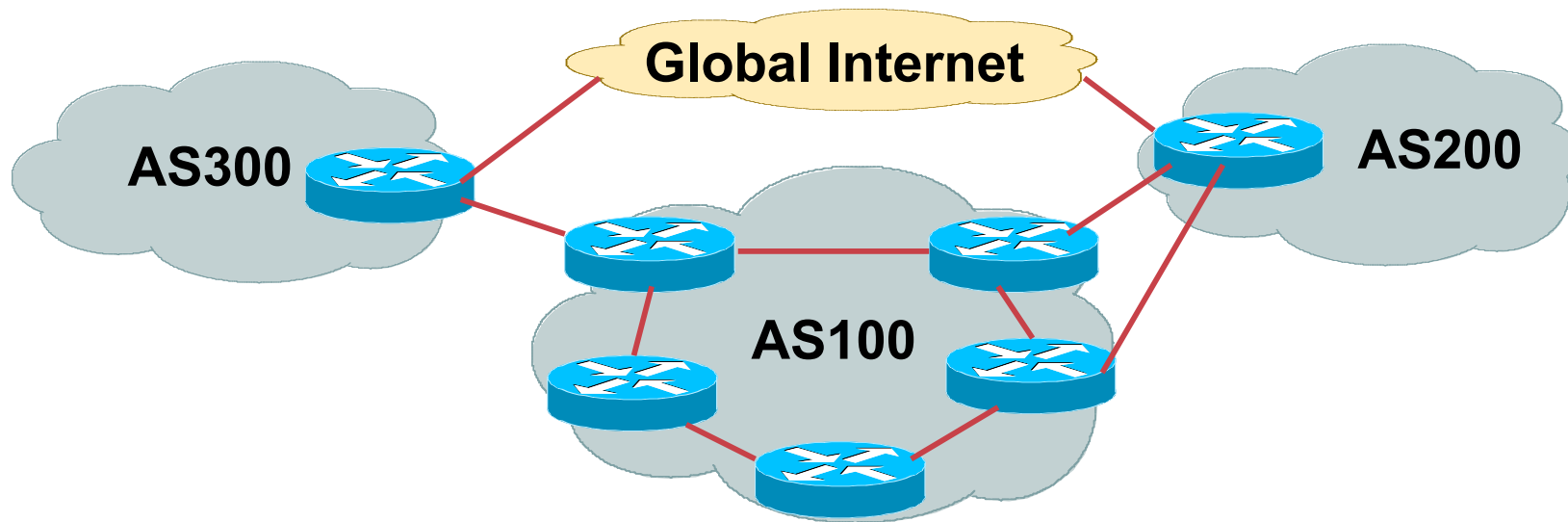
- **No need for BGP**
- **Point static default to upstream ISP**
- **Router will load share on the two parallel circuits**
- **Upstream ISP advertises stub network**
- **Policy confined within upstream ISP's policy**

Multi-homed Stub Network



- **Use BGP (not IGP or static) to loadshare**
- **Use private AS (ASN > 64511)**
- **Upstream ISP advertises stub network**
- **Policy confined within upstream ISP's policy**

Multi-Homed Network

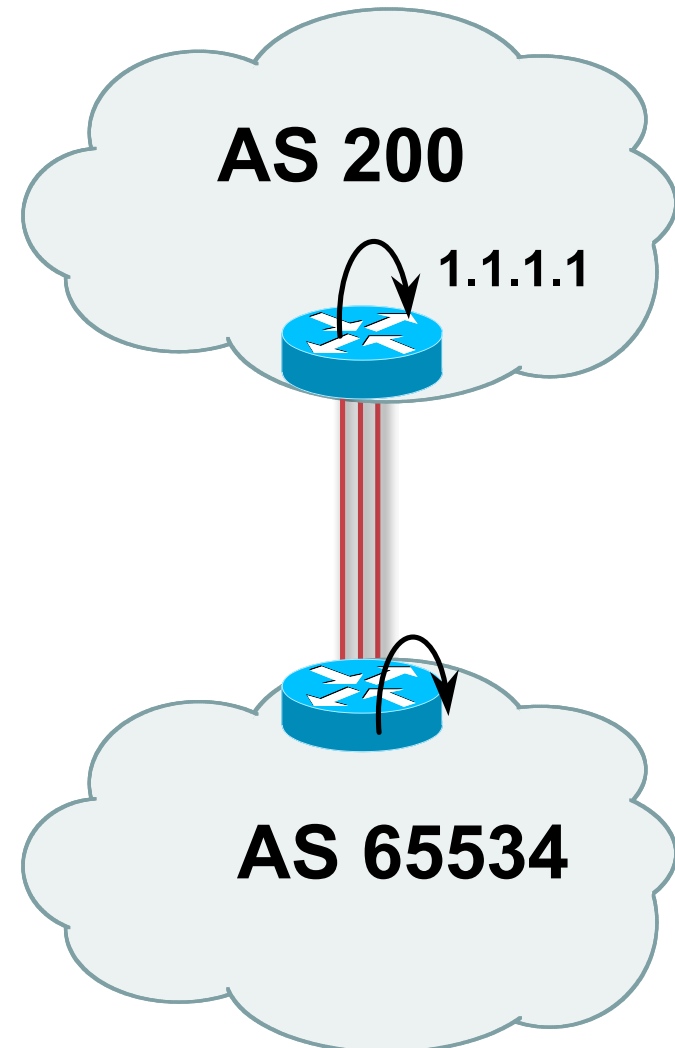


- **Many situations possible**
 - multiple sessions to same ISP
 - secondary for backup only
 - load-share between primary and secondary
 - selectively use different ISPs

Multiple Sessions to an ISP

- Use eBGP multihop
 - eBGP to loopback addresses
 - eBGP prefixes learned with loopback address as next hop
- Cisco IOS

```
router bgp 65534
  neighbor 1.1.1.1 remote-as 200
  neighbor 1.1.1.1 ebgp-multihop 2
!
ip route 1.1.1.1 255.255.255.255 serial 1/0
ip route 1.1.1.1 255.255.255.255 serial 1/1
ip route 1.1.1.1 255.255.255.255 serial 1/2
```



Multiple Sessions to an ISP

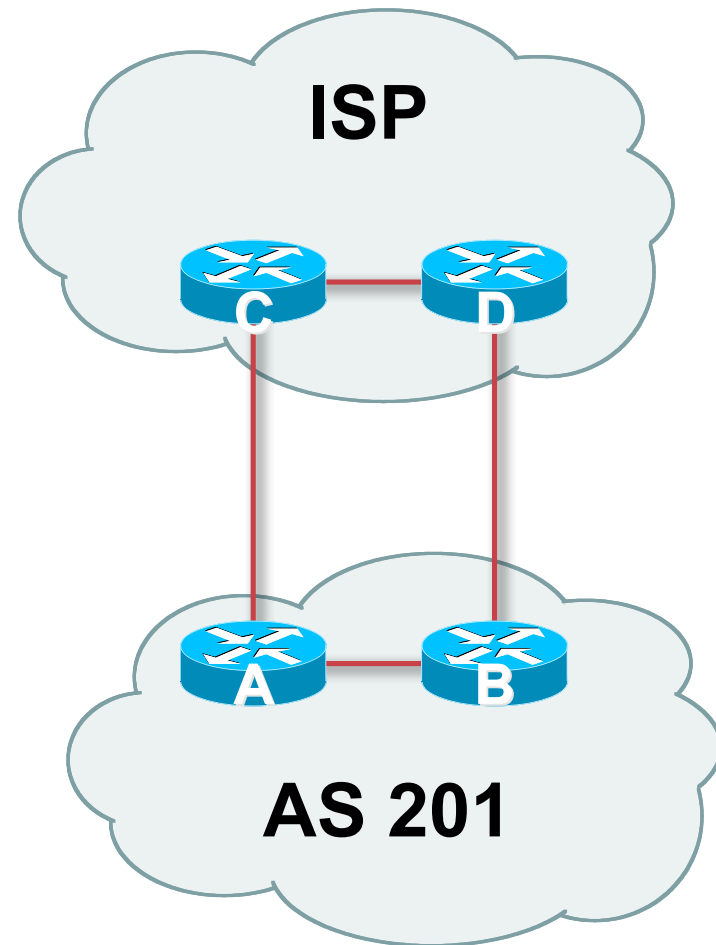
- Try and avoid use of ebgp-multihop unless:
 - It's absolutely necessary **–or–**
 - Loadsharing across multiple links
- Many ISPs discourage its use, for example:

We will run eBGP multihop, but do not support it as a standard offering because customers generally have a hard time managing it due to:

- routing loops
- failure to realise that BGP session stability problems are usually due connectivity problems between their CPE and their BGP speaker

Multiple Sessions to an ISP

- Simplest scheme is to use defaults
- Learn/advertise prefixes for better control
- Planning and some work required to achieve loadsharing
 - Point default towards one ISP
 - Learn selected prefixes from second ISP
 - Modify the number of prefixes learnt to achieve acceptable load sharing
- No magic solution





Basic Multihoming

Learning to walk before we try running

Basic Multihoming

- **No frills multihoming**
- **Will look at two cases:**
 - Multihoming with the same ISP**
 - Multihoming to different ISPs**
- **Will keep the examples easy**
 - Understanding easy concepts will make the more complex scenarios easier to comprehend**
 - All assume that the site multihoming has a /19 address block**

Basic Multihoming

- **This type is most commonplace at the edge of the Internet**
 - Networks here are usually concerned with inbound traffic flows**
 - Outbound traffic flows being “nearest exit” is usually sufficient**
- **Can apply to the leaf ISP as well as Enterprise networks**



Basic Multihoming

Multihoming to the Same ISP

Basic Multihoming:

Multihoming to the same ISP

- **Use BGP for this type of multihoming**

use a private AS (ASN > 64511)

There is no need or justification for a public ASN

Making the nets of the end-site visible gives no useful information to the Internet

- **Upstream ISP proxy aggregates**

in other words, announces only your address block to the Internet from their AS (as would be done if you had one statically routed connection)



Two links to the same ISP

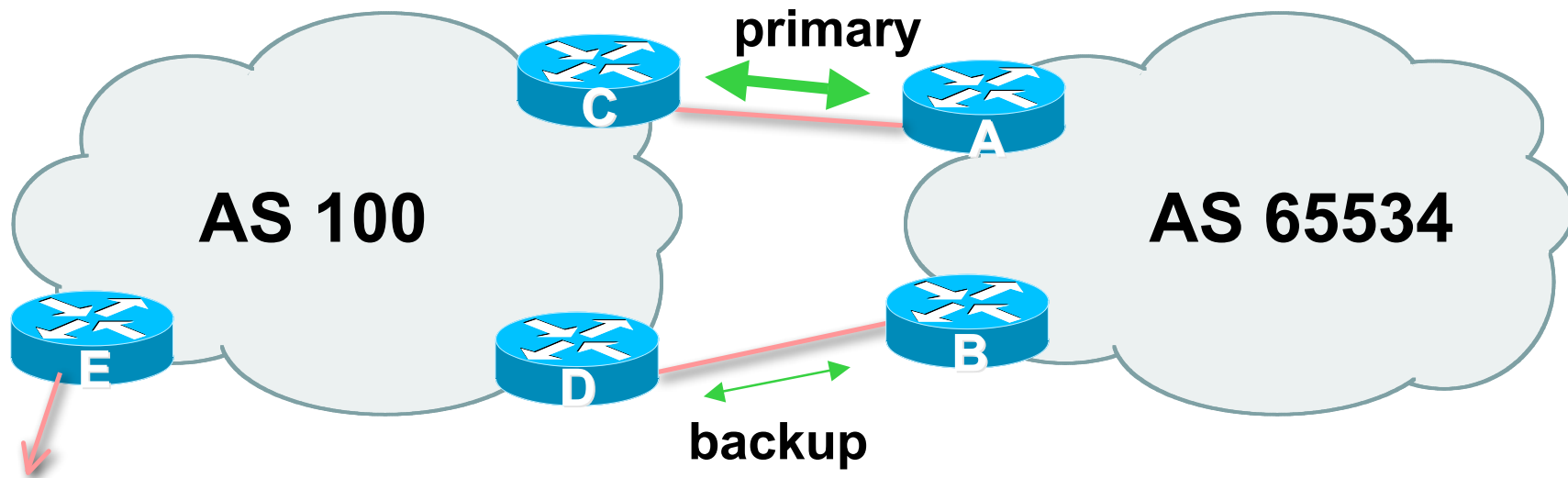
One link primary, the other link backup only

Two links to the same ISP (one as backup only)

- **Applies when end-site has bought a large primary WAN link to their upstream a small secondary WAN link as the backup**

For example, primary path might be an E1, backup might be 64kbps

Two links to the same ISP (one as backup only)



- **Border router E in AS100 removes private AS and any customer subprefixes from Internet announcement**

Two links to the same ISP (one as backup only)

- **Announce /19 aggregate on each link**
 - primary link:**
 - Outbound – announce /19 unaltered**
 - Inbound – receive default route**
 - backup link:**
 - Outbound – announce /19 with increased metric**
 - Inbound – received default, and reduce local preference**
- **When one link fails, the announcement of the /19 aggregate via the other link ensures continued connectivity**



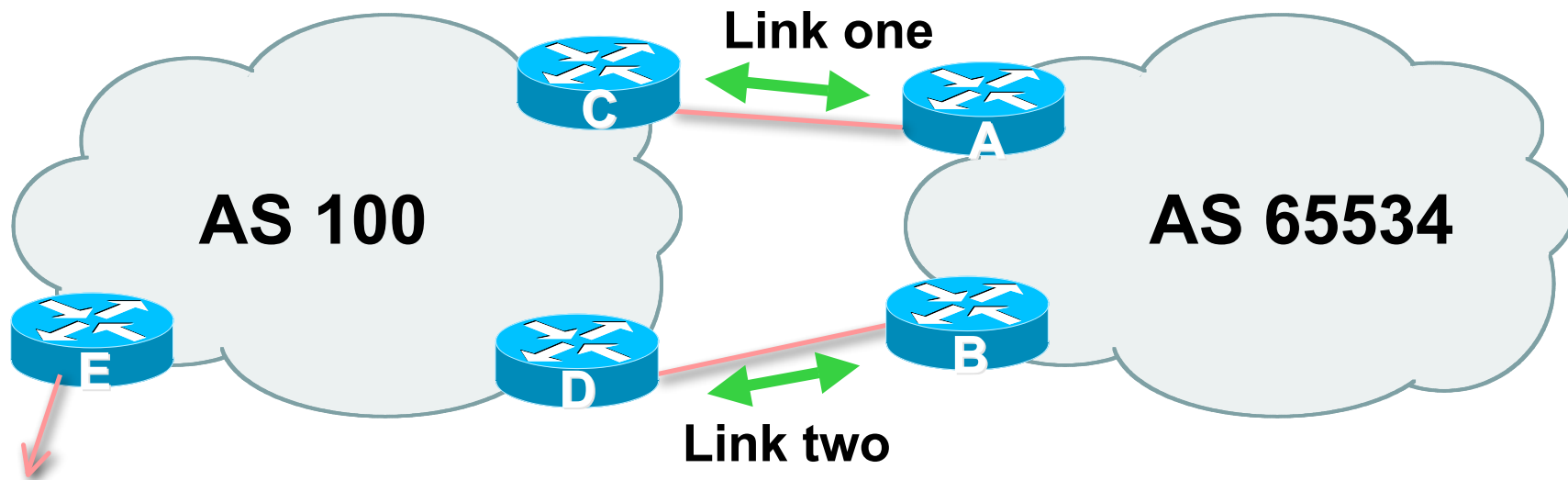
Two links to the same ISP

With Loadsharing

Loadsharing to the same ISP

- **More common case**
- **End sites tend not to buy circuits and leave them idle, only used for backup as in previous example**
- **This example assumes equal capacity circuits**
Unequal capacity circuits requires more refinement – see later

Loadsharing to the same ISP



- **Border router E in AS100 removes private AS and any customer subprefixes from Internet announcement**

Loadsharing to the same ISP

- **Announce /19 aggregate on each link**
- **Split /19 and announce as two /20s, one on each link**
 - basic inbound loadsharing**
 - assumes equal circuit capacity and even spread of traffic across address block**
- **Vary the split until “perfect” loadsharing achieved**
- **Accept the default from upstream**
 - basic outbound loadsharing by nearest exit**
 - okay in first approx as most ISP and end-site traffic is inbound**

Loadsharing to the same ISP

- **Loadsharing configuration is only on customer router**
- **Upstream ISP has to**
 - remove customer subprefixes from external announcements**
 - remove private AS from external announcements**
- **Could also use BGP communities**



Basic Multihoming

Multihoming to different ISPs

Two links to different ISPs

- **Use a Public AS**

Or use private AS if agreed with the other ISP

But some people don't like the "inconsistent-AS" which results from use of a private-AS

- **Address space comes from**

both upstreams **or**

Regional Internet Registry

- **Configuration concepts very similar**

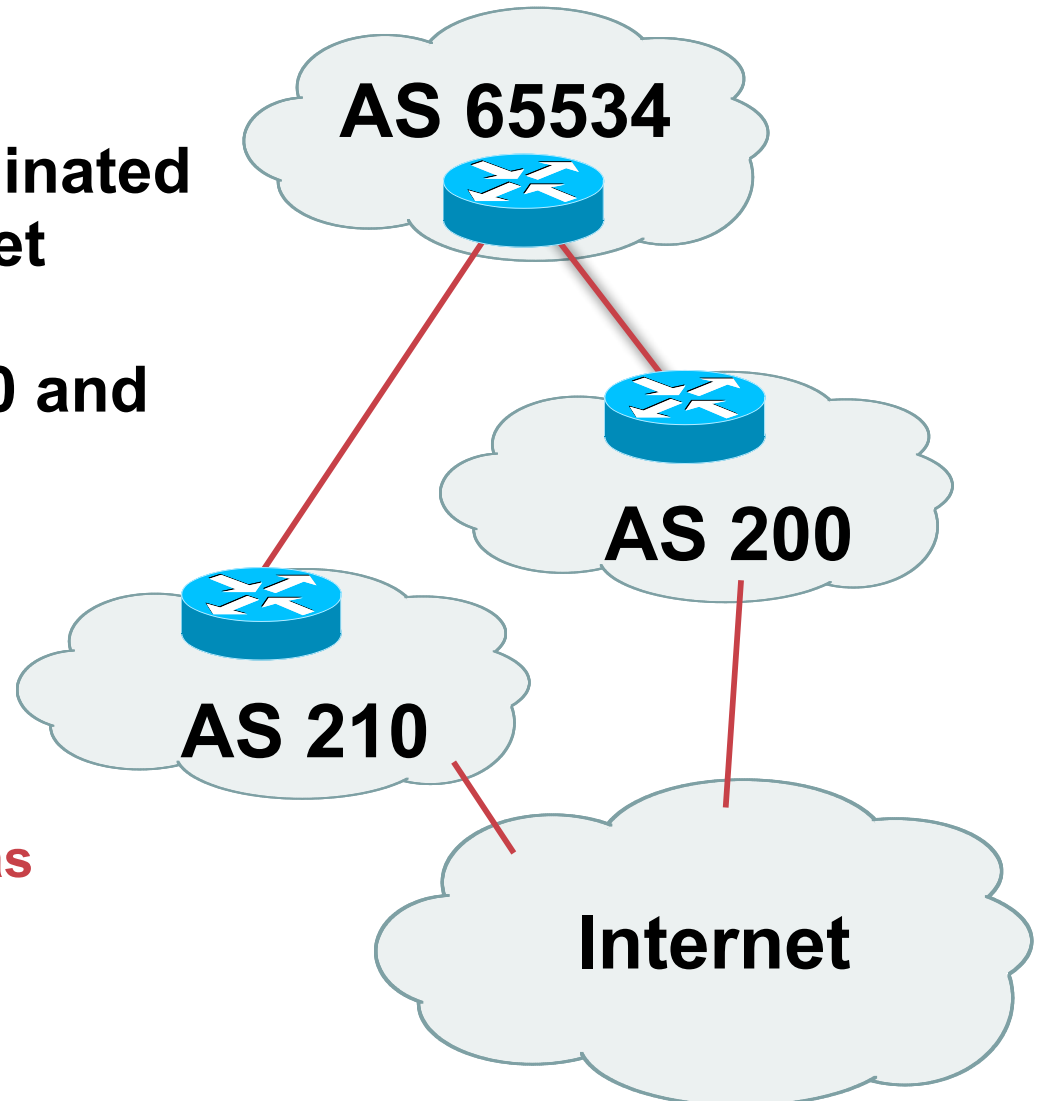
Inconsistent-AS?

- Viewing the prefixes originated by AS65534 in the Internet shows they appear to be originated by both AS210 and AS200

This is NOT bad

Nor is it illegal

- Cisco IOS command is
show ip bgp inconsistent-as

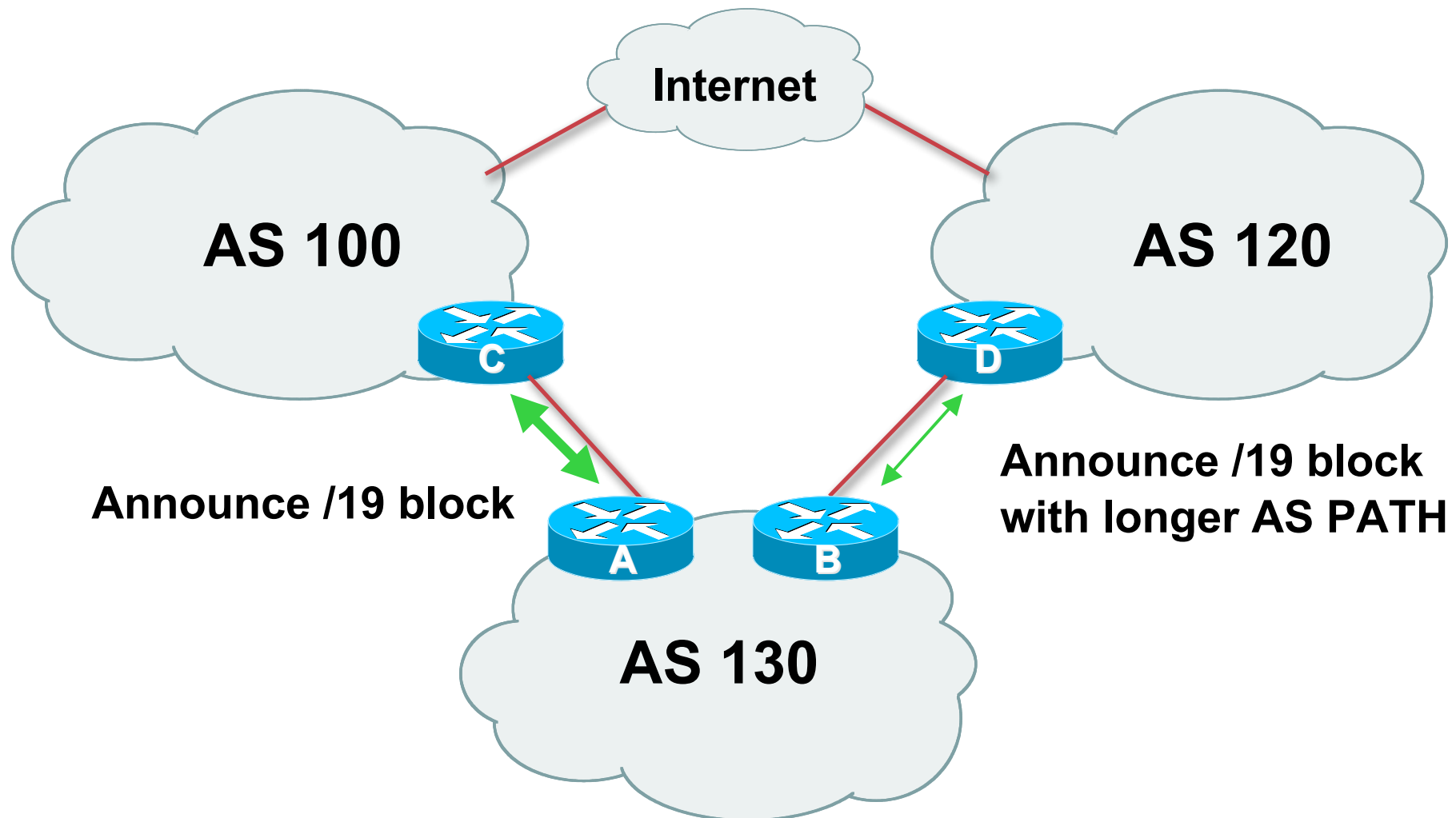




Two links to different ISPs

One link primary, the other link backup only

Two links to different ISPs (one as backup only)



Two links to different ISPs (one as backup only)

- **Announce /19 aggregate on each link**
 - primary link makes standard announcement
 - backup link lengthens the AS PATH by using AS PATH prepend
- **When one link fails, the announcement of the /19 aggregate via the other link ensures continued connectivity**

Two links to different ISPs (one as backup only)

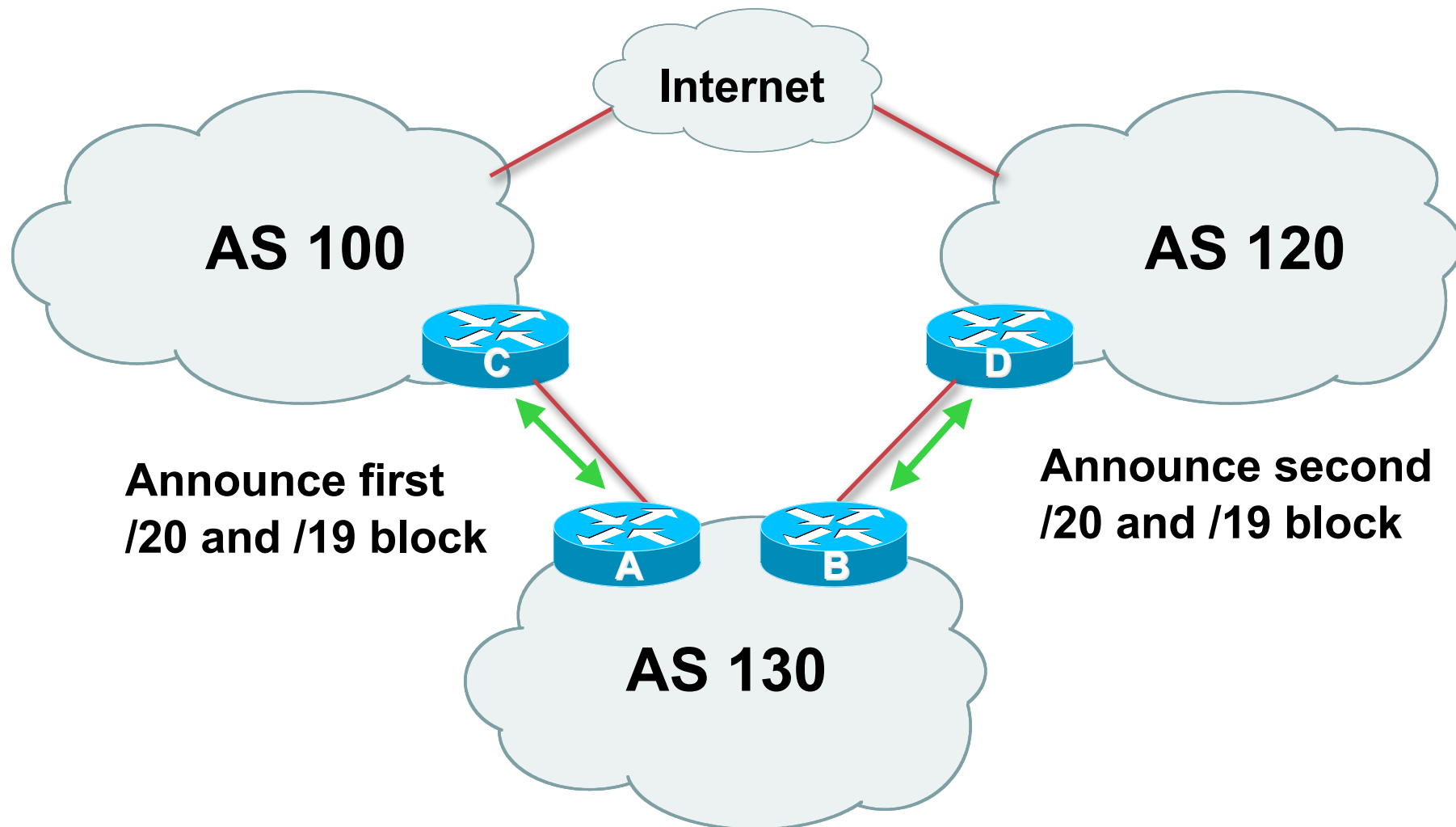
- **Not a common situation as most sites tend to prefer using whatever capacity they have**
- **But it shows the basic concepts of using local-prefs and AS-path prepends for engineering traffic in the chosen direction**



Two links to different ISPs

With Loadsharing

Two links to different ISPs (with loadsharing)



Two links to different ISPs (with loadsharing)

- **Announce /19 aggregate on each link**
- **Split /19 and announce as two /20s, one on each link**
basic inbound loadsharing
- **When one link fails, the announcement of the /19 aggregate via the other ISP ensures continued connectivity**

Two links to different ISPs (with loadsharing)

- **Loadsharing in this case is very basic**
- **But shows the first steps in designing a load sharing solution**

Start with a simple concept

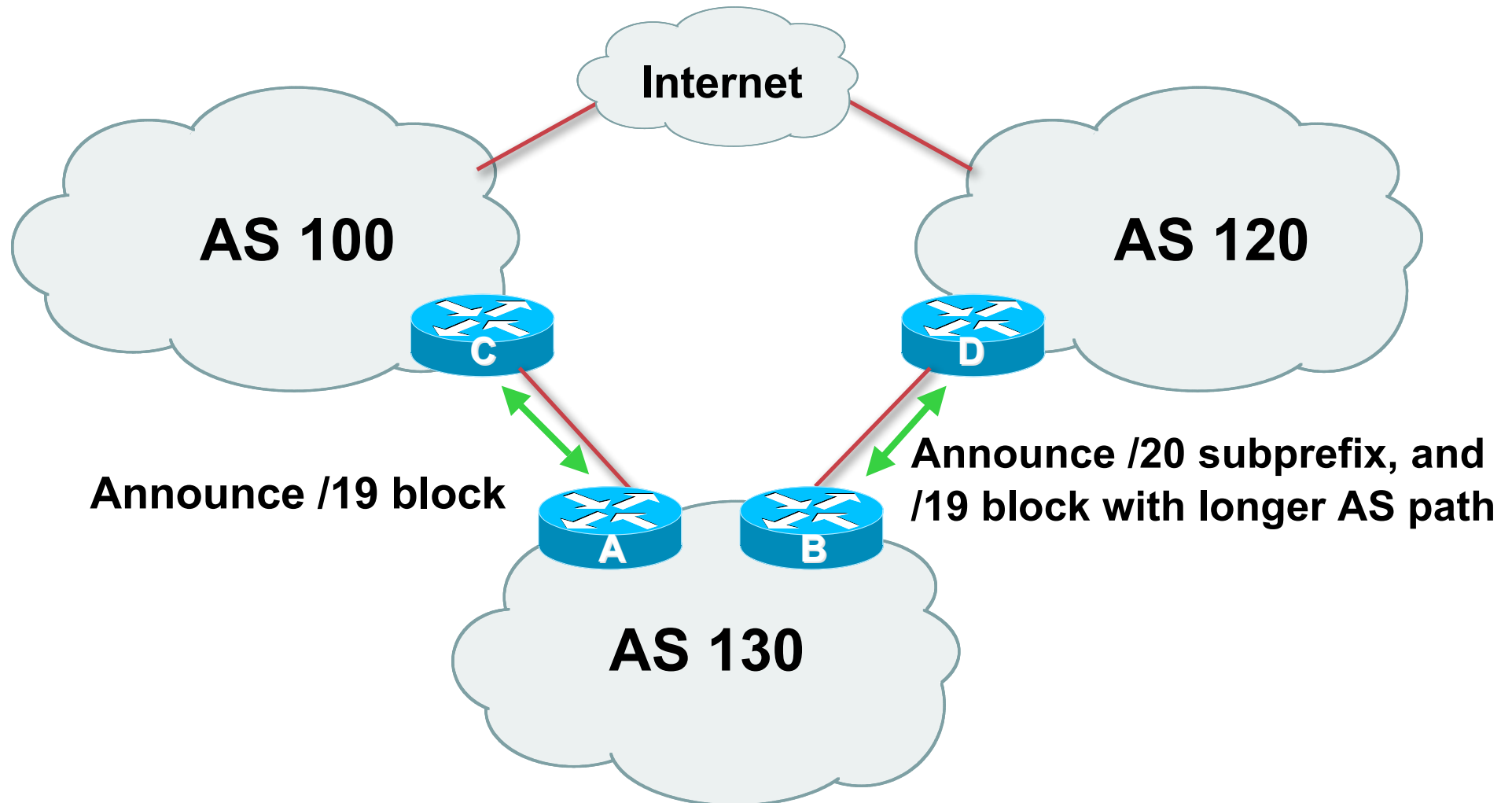
And build on it...!



Two links to different ISPs

More Controlled Loadsharing

Loadsharing with different ISPs



Loadsharing with different ISPs

- **Announce /19 aggregate on each link**
 - On first link, announce /19 as normal
 - On second link, announce /19 with longer AS PATH, and announce one /20 subprefix
 - controls loadsharing between upstreams and the Internet
- **Vary the subprefix size and AS PATH length until “perfect” loadsharing achieved**
- **Still require redundancy!**

Loadsharing with different ISPs

- **This example is more commonplace**
- **Shows how ISPs and end-sites subdivide address space frugally, as well as use the AS-PATH prepend concept to optimise the load sharing between different ISPs**
- **Notice that the /19 aggregate block is ALWAYS announced**

BGP Techniques for Providers

- **BGP Basics**
- **Scaling BGP**
- **Deploying BGP**
- **Multihoming Basics**
- **BGP “Traffic Engineering”**
- **BGP Configuration Tips**



Service Provider Multihoming

BGP Traffic Engineering

Service Provider Multihoming

- **Previous examples dealt with loadsharing inbound traffic**
 - Of primary concern at Internet edge
 - What about outbound traffic?
- **Transit ISPs strive to balance traffic flows in both directions**
 - Balance link utilisation
 - Try and keep most traffic flows symmetric
 - Some edge ISPs try and do this too
- **The original “Traffic Engineering”**

Service Provider Multihoming

- **Balancing outbound traffic requires inbound routing information**

Common solution is “full routing table”

Rarely necessary

Why use the “routing mallet” to try solve loadsharing problems?

“Keep It Simple” is often easier (and \$\$\$ cheaper) than carrying N-copies of the full routing table

Service Provider Multihoming MYTHS!!

- **Common MYTHS**
- **1: You need the full routing table to multihome**

People who sell router memory would like you to believe this
Only true if you are a transit provider
Full routing table can be a significant hindrance to multihoming
- **2: You need a BIG router to multihome**

Router size is related to data rates, not running BGP
In reality, to multihome, your router needs to:
Have two interfaces,
Be able to talk BGP to at least two peers,
Be able to handle BGP attributes,
Handle at least one prefix
- **3: BGP is complex**

In the wrong hands, yes it can be! Keep it Simple!

Service Provider Multihoming: Some Strategies

- **Take the prefixes you need to aid traffic engineering**
Look at NetFlow data for popular sites
- **Prefixes originated by your immediate neighbours and their neighbours will do more to aid load balancing than prefixes from ASNs many hops away**
Concentrate on local destinations
- **Use default routing as much as possible**
Or use the full routing table with care

Service Provider Multihoming

- **Examples**

- One upstream, one local peer

- Two upstreams, one local peer

- **Require BGP and a public ASN**

- **Examples assume that the local network has their own /19 address block**



Service Provider Multihoming

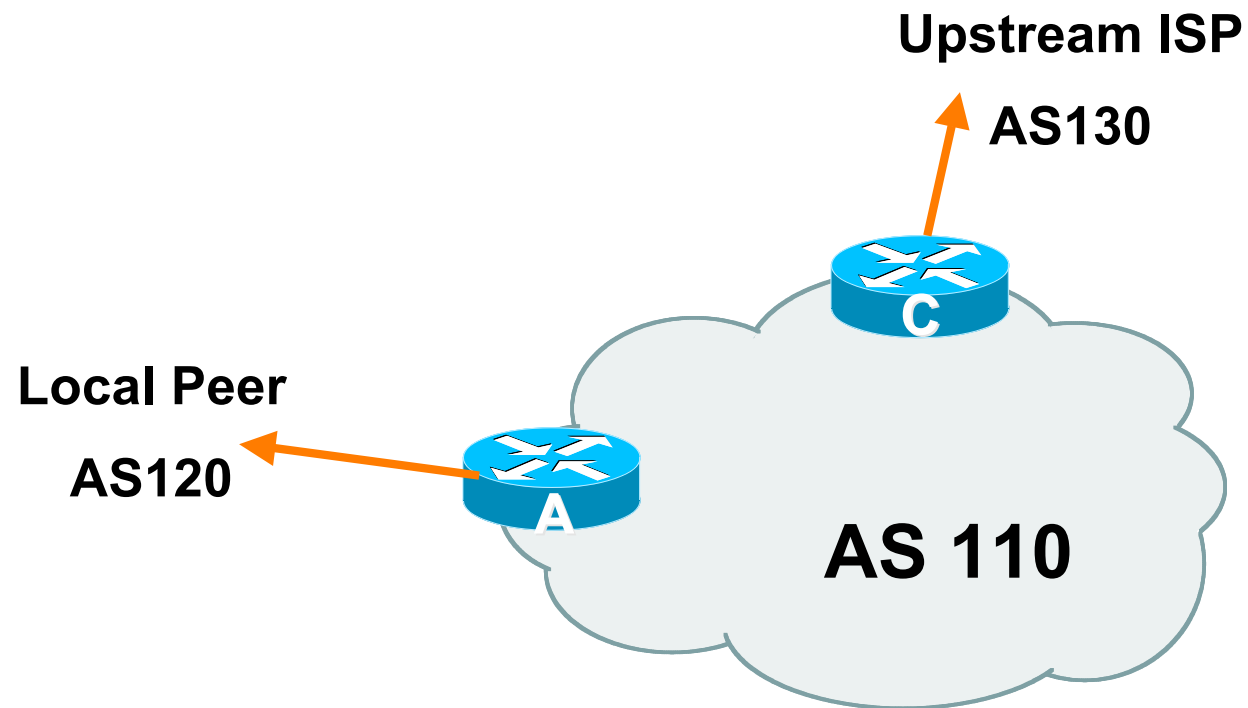
One upstream, one local peer

One Upstream, One Local Peer

- **Very common situation in many regions of the Internet**
- **Connect to upstream transit provider to see the “Internet”**
- **Connect to the local competition so that local traffic stays local**

Saves spending valuable \$ on upstream transit costs for local traffic

One Upstream, One Local Peer



One Upstream, One Local Peer

- **Announce /19 aggregate on each link**
- **Accept default route only from upstream**
Either 0.0.0.0/0 or a network which can be used as default
- **Accept all routes from local peer**

One Upstream, One Local Peer

- **Two configurations possible for Router A**
 - Use of AS Path Filters assumes peer knows what they are doing
 - Prefix Filters are higher maintenance, but safer
 - Some ISPs use **both**
- **Local traffic goes to and from local peer, everything else goes to upstream**



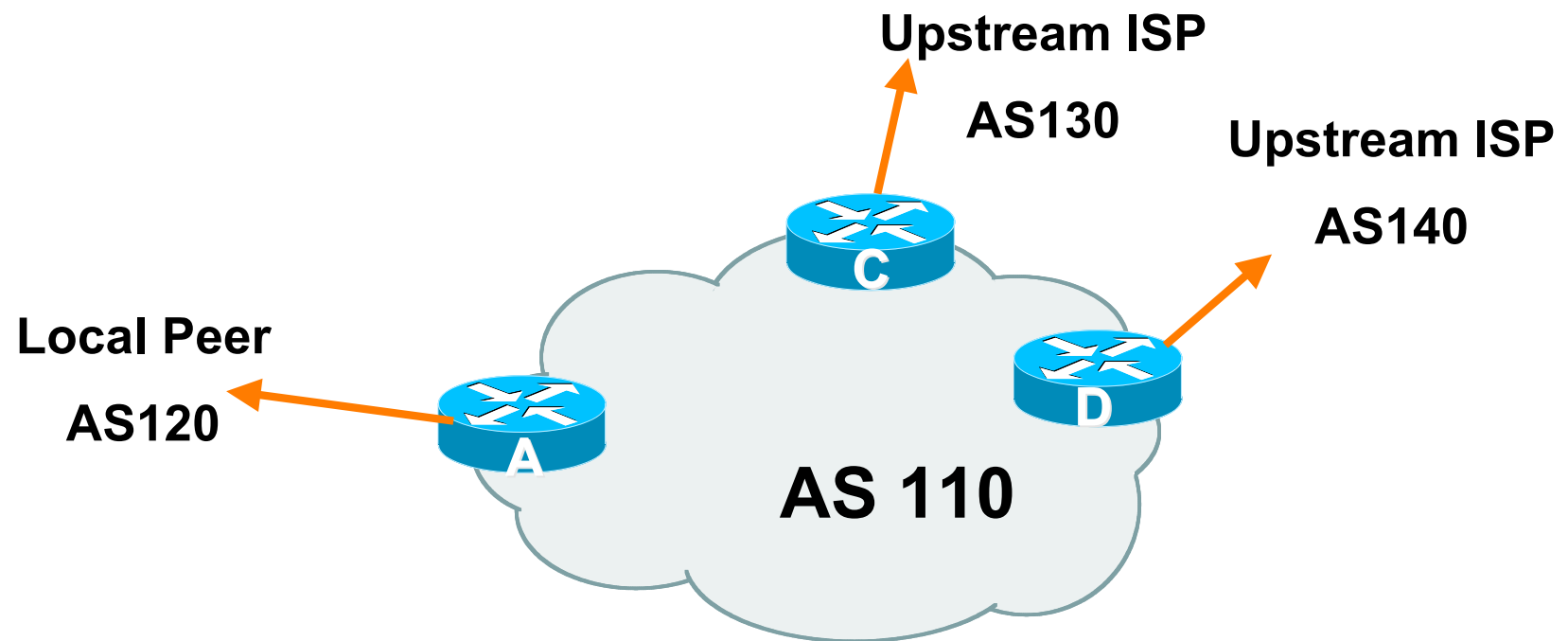
Service Provider Multihoming

Two Upstreams, One local peer

Two Upstreams, One Local Peer

- **Connect to both upstream transit providers to see the “Internet”**
Provides external redundancy and diversity – the reason to multihome
- **Connect to the local peer so that local traffic stays local**
Saves spending valuable \$ on upstream transit costs for local traffic

Two Upstreams, One Local Peer



Two Upstreams, One Local Peer

- **Announce /19 aggregate on each link**
- **Accept default route only from upstreams**
 - Either 0.0.0.0/0 or a network which can be used as default
- **Accept all routes from local peer**

Two Upstreams, One Local Peer

- **Router A has same routing configuration as in example with one upstream and one local peer**
- **Two configuration options for Routers C and D:**
 - Accept full routing from both upstreams**
 - Expensive & unnecessary!**
 - Accept default from one upstream and some routes from the other upstream**
 - The way to go!**

Two Upstreams, One Local Peer Full Routes

- **Router C configuration:**
 - Accept full routes from AS130**
 - Tag prefixes originated by AS130 and AS130's neighbouring ASes with local preference 120**
 - Traffic to those ASes will go over AS130 link**
 - Remaining prefixes tagged with local preference of 80**
 - Traffic to other all other ASes will go over the link to AS140**
- **Router D configuration same as Router C without setting any preferences**

Two Upstreams, One Local Peer

Full Routes

- **Full routes from upstreams**

Expensive – needs lots of memory and CPU

Need to play preference games

Previous example is only an example – real life will need improved fine-tuning!

Previous example doesn't consider inbound traffic – see earlier in presentation for examples

Two Upstreams, One Local Peer

Partial Routes

- **Strategy:**

- Ask one upstream for a default route**

- Easy to originate default towards a BGP neighbour**

- Ask other upstream for a full routing table**

- Then filter this routing table based on neighbouring ASN**

- E.g. want traffic to their neighbours to go over the link to that ASN**

- Most of what upstream sends is thrown away**

- Easier than asking the upstream to set up custom BGP filters for you**

Two Upstreams, One Local Peer

Partial Routes

- **Router C configuration:**
 - Accept full routes from AS130**
(or get them to send less)
 - Filter ASNs so only AS130 and AS130's neighbouring ASes are accepted**
 - Allow default, and set it to local preference 80**
 - Traffic to those ASes will go over AS130 link**
 - Traffic to other all other ASes will go over the link to AS140**
 - If AS140 link fails, backup via AS130 – and vice-versa**
- **Router D configuration:**
 - Accept only the default route**

Two Upstreams, One Local Peer

Partial Routes

- **Partial routes from upstreams**

Not expensive – only carry the routes necessary for loadsharing

Need to filter on AS paths

Previous example is only an example – real life will need improved fine-tuning!

Previous example doesn't consider inbound traffic – see earlier in presentation for examples

Two Upstreams, One Local Peer

- **When upstreams cannot or will not announce default route**

Because of operational policy against using “default-originate” on BGP peering

Solution is to use IGP to propagate default from the edge/peering routers

BGP Techniques for Providers

- **BGP Basics**
- **Scaling BGP**
- **Deploying BGP**
- **Multihoming Basics**
- **BGP “Traffic Engineering”**
- **BGP Configuration Tips**



Configuration Tips

Of templates, passwords, tricks, and more templates

iBGP and IGP 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 the impact
 - Do NOT use the vendor defaults** without thinking
- **Remove private ASes from announcements**
 - Common omission today
- **Use extensive filters, with “backup”**
 - Use as-path filters to backup prefix filters
 - Keep policy language for implementing policy, rather than basic filtering
- **Use password agreed between you and peer on eBGP session**

eBGP Template

Example continued

- **Use maximum-prefix tracking**
Router will warn you if there are sudden increases in BGP table size, bringing down eBGP if desired
- **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**

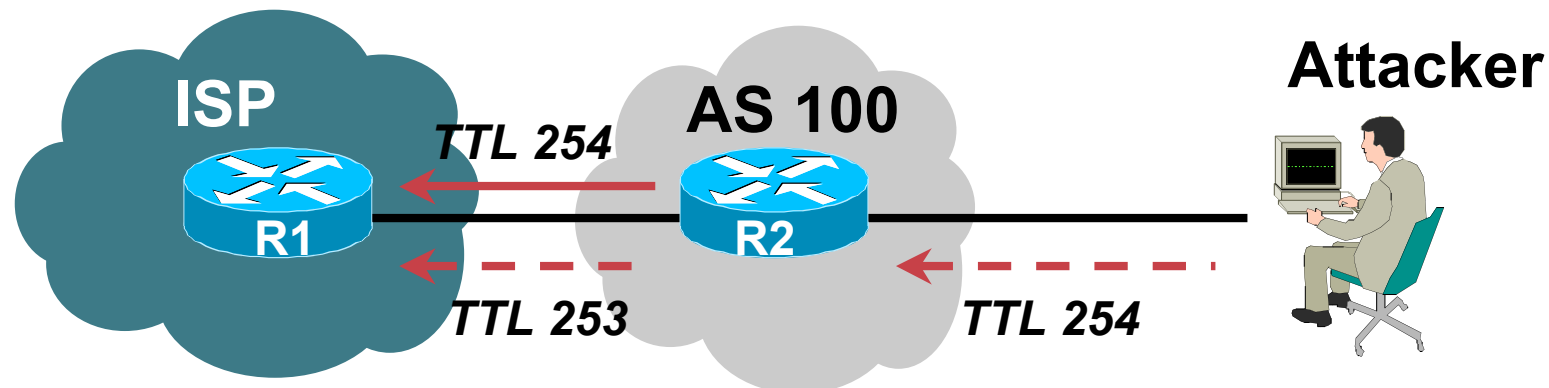
BGP TTL “hack”

- 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"**

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**

Presentation Slides

- **Are available on**

<ftp://ftp-eng.cisco.com>

[/pfs/seminars/QUESTnet2006-BGP-Tutorial.pdf](#)

And will be on the QUESTnet 2006 website

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



BGP Techniques for Providers

Philip Smith <pfs@cisco.com>

QUESTnet 2006

11th - 14th July



Supplementary Materials



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**

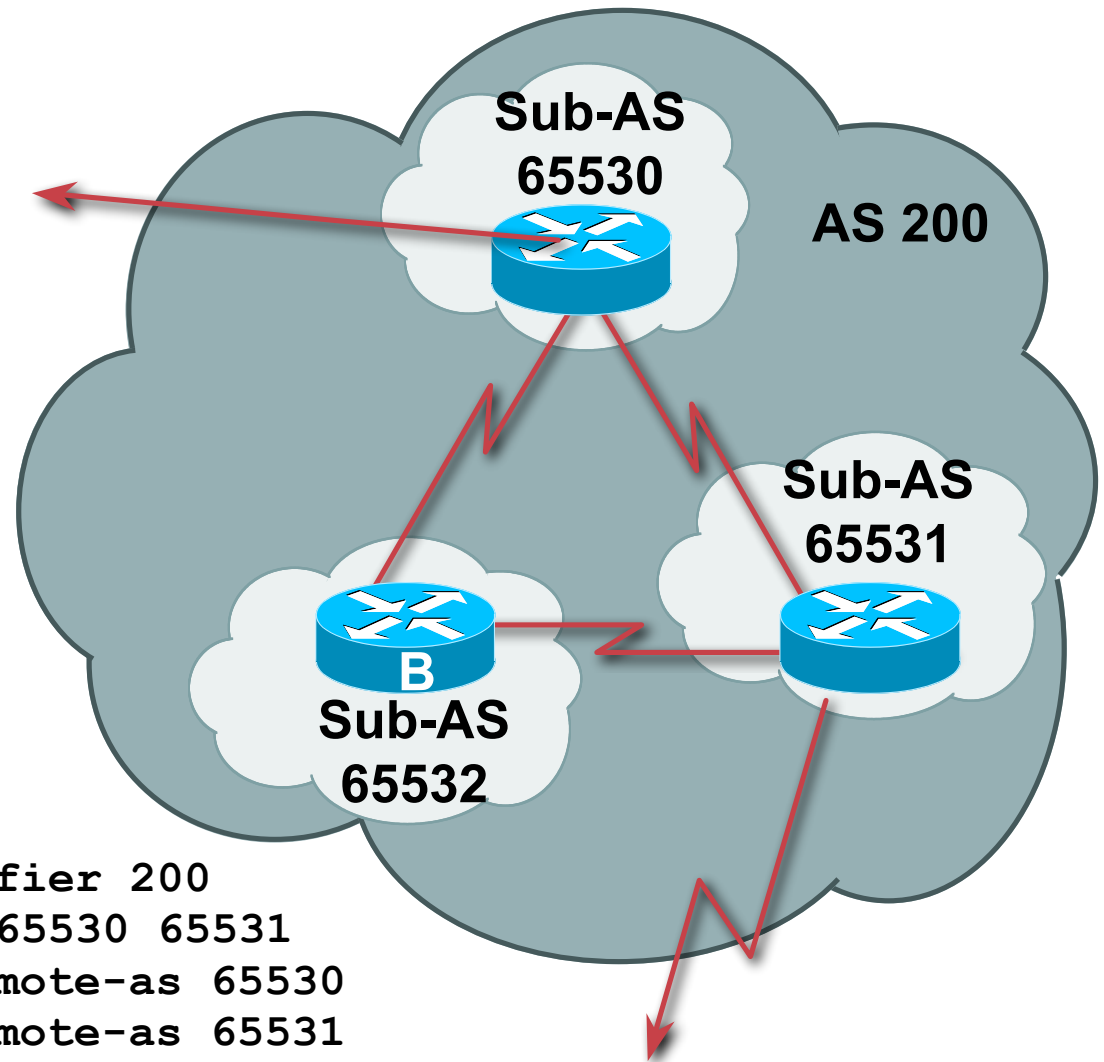
The total number of neighbours is reduced by limiting the full mesh requirement to only the peers in the sub-AS

Can also use Route-Reflector within sub-AS

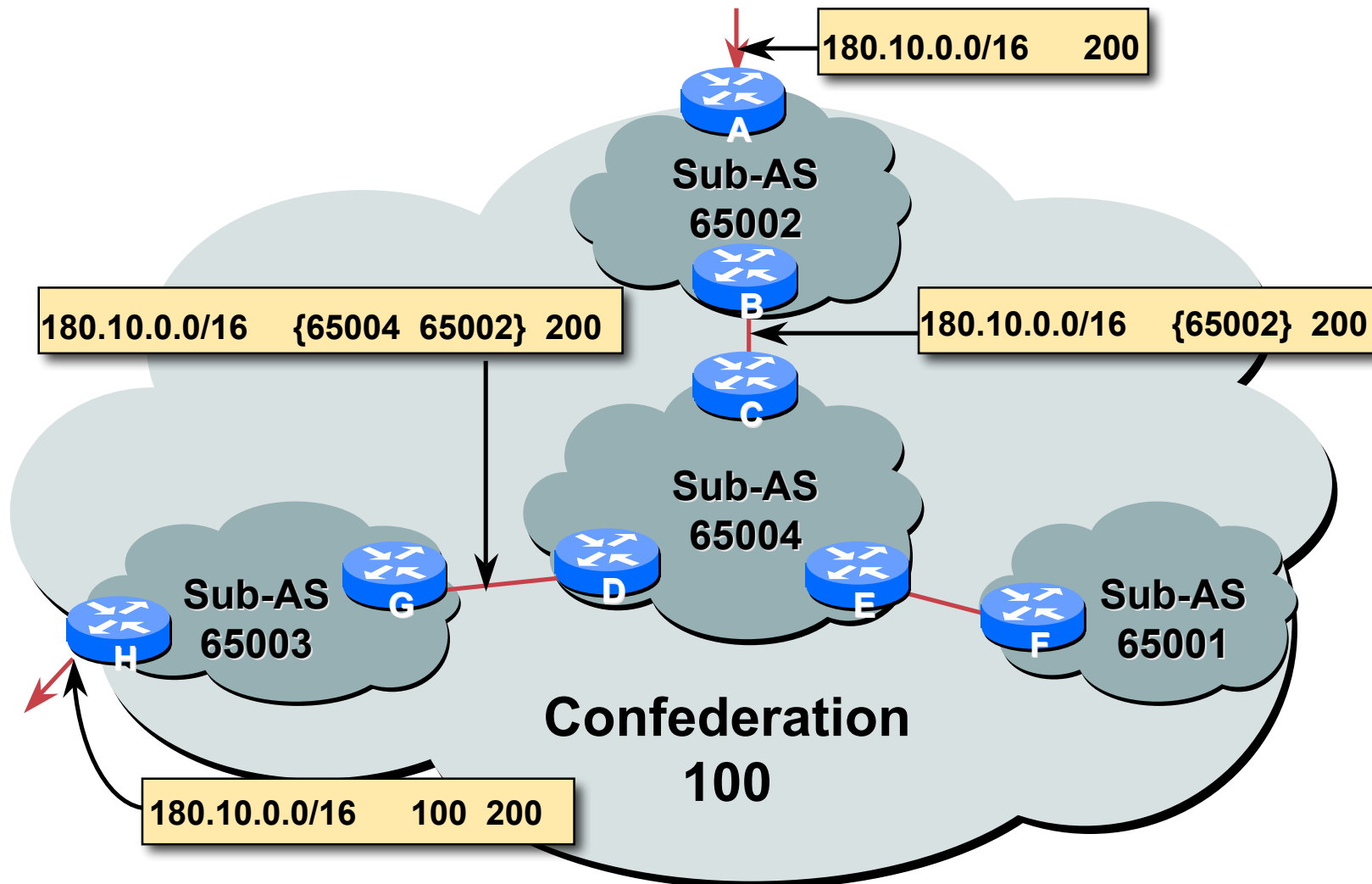
Confederations

- **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**



Route Flap Damping

Network Stability for the 1990s

Network Instability for the 21st Century!

Route Flap Damping

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

Route Flap Damping

- **Route flap**

Going up and down of path or change in attribute

BGP WITHDRAW followed by UPDATE = 1 flap

eBGP neighbour going down/up is NOT a flap

Ripples through the entire Internet

Wastes CPU

- **Damping aims to reduce scope of route flap propagation**

Route Flap Damping (continued)

- **Requirements**

- Fast convergence for normal route changes**

- History predicts future behaviour**

- Suppress oscillating routes**

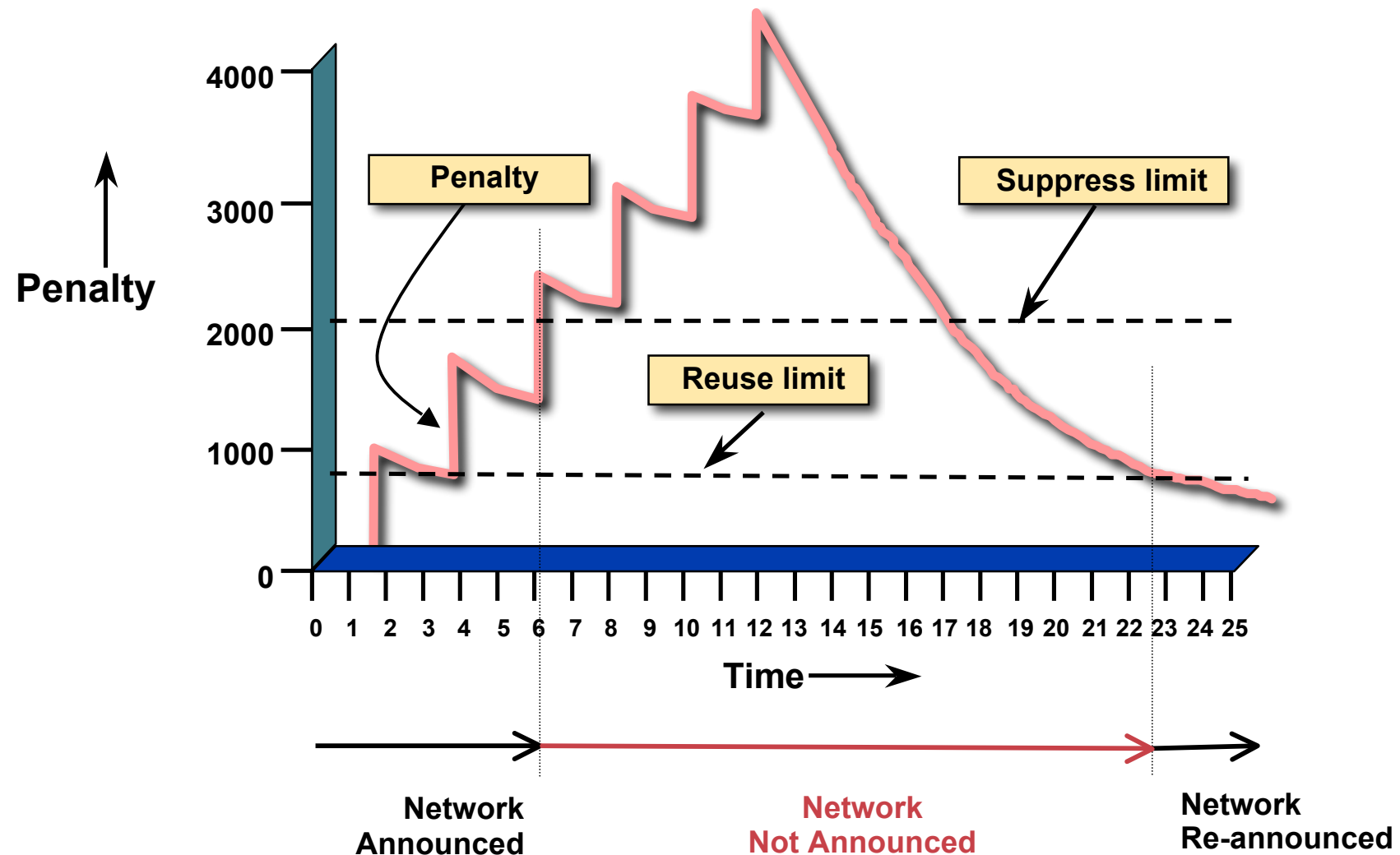
- Advertise stable routes**

- **Implementation described in RFC 2439**

Operation

- **Add penalty (1000) for each flap**
Change in attribute gets penalty of 500
- **Exponentially decay penalty**
half life determines decay rate
- **Penalty above suppress-limit**
do not advertise route to BGP peers
- **Penalty decayed below reuse-limit**
re-advertise route to BGP peers
penalty reset to zero when it is half of reuse-limit

Operation



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**

Route Flap Damping History

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

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

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

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

Serious Problems:

- **"Route Flap Damping Exacerbates Internet Routing Convergence"**

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

- **"What is the sound of one route flapping?"**

Tim Griffin, June 2002

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

- **"Happy Packets"**

Closely related work by Randy Bush *et al*

Problem 1:

- **One path flaps:**

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

Those peers see change in best path, flap counter incremented

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

Problem 2:

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

Solution:

- Do **NOT** use Route Flap Damping whatever you do!
- RFD will unnecessarily impair access
to your network and
to the Internet
- More information contained in RIPE Routing
Working Group recommendations:
[www.ripe.net/ripe/docs/ripe-378.\[pdf,html,txt\]](http://www.ripe.net/ripe/docs/ripe-378.[pdf,html,txt])