BGP for Internet Service Providers

Philip Smith <pfs@cisco.com>

NANOG 25, Toronto, Canada

Presentation Slides

- **Will be available on**

  www.cisco.com/public/cons/seminars/NANOG25

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

BGP for Internet Service Providers

- BGP Basics (quick recap)
- Scaling BGP
- Deploying BGP in an ISP network
- Multihoming Examples
- Using Communities

BGP Basics

What is this BGP thing?

Border Gateway Protocol

Cisco.com

- **Routing Protocol used to exchange routing information between networks**

  exterior gateway protocol

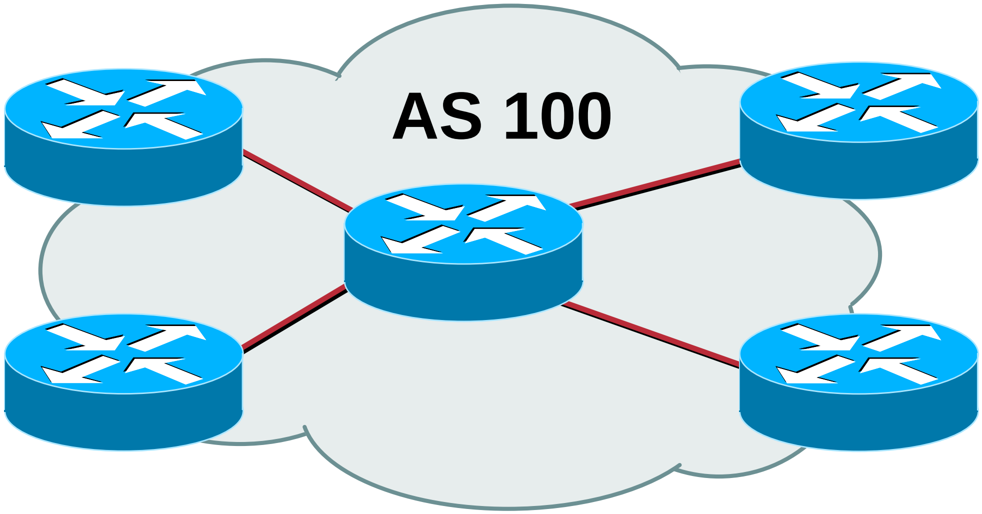
- **RFC1771**

  work in progress to update

  `draft-ietf-idr-bgp4-17.txt`

Autonomous System (AS)

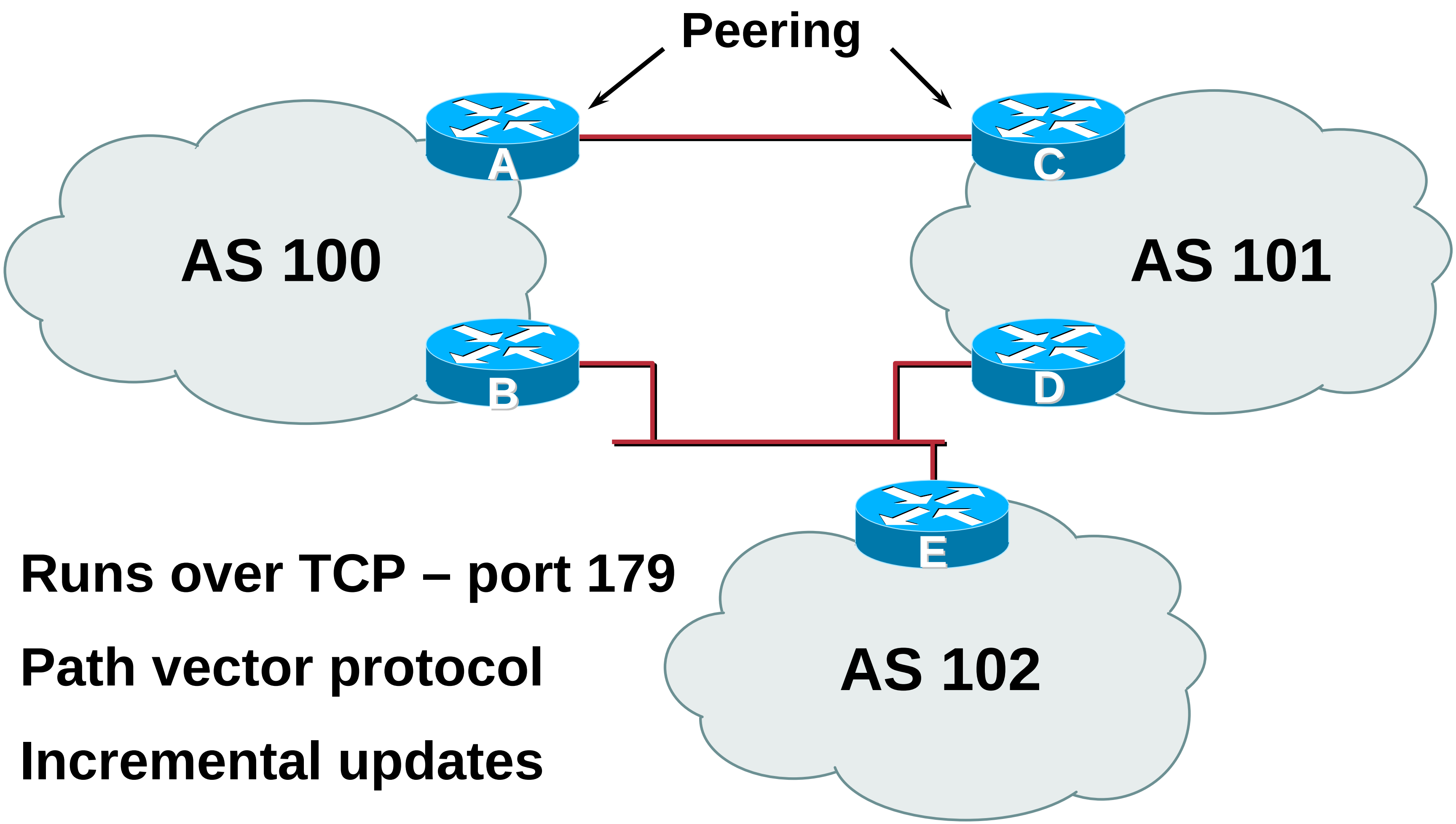
Cisco.com



- Collection of networks with same routing policy
- Single routing protocol
- Usually under single ownership, trust and administrative control

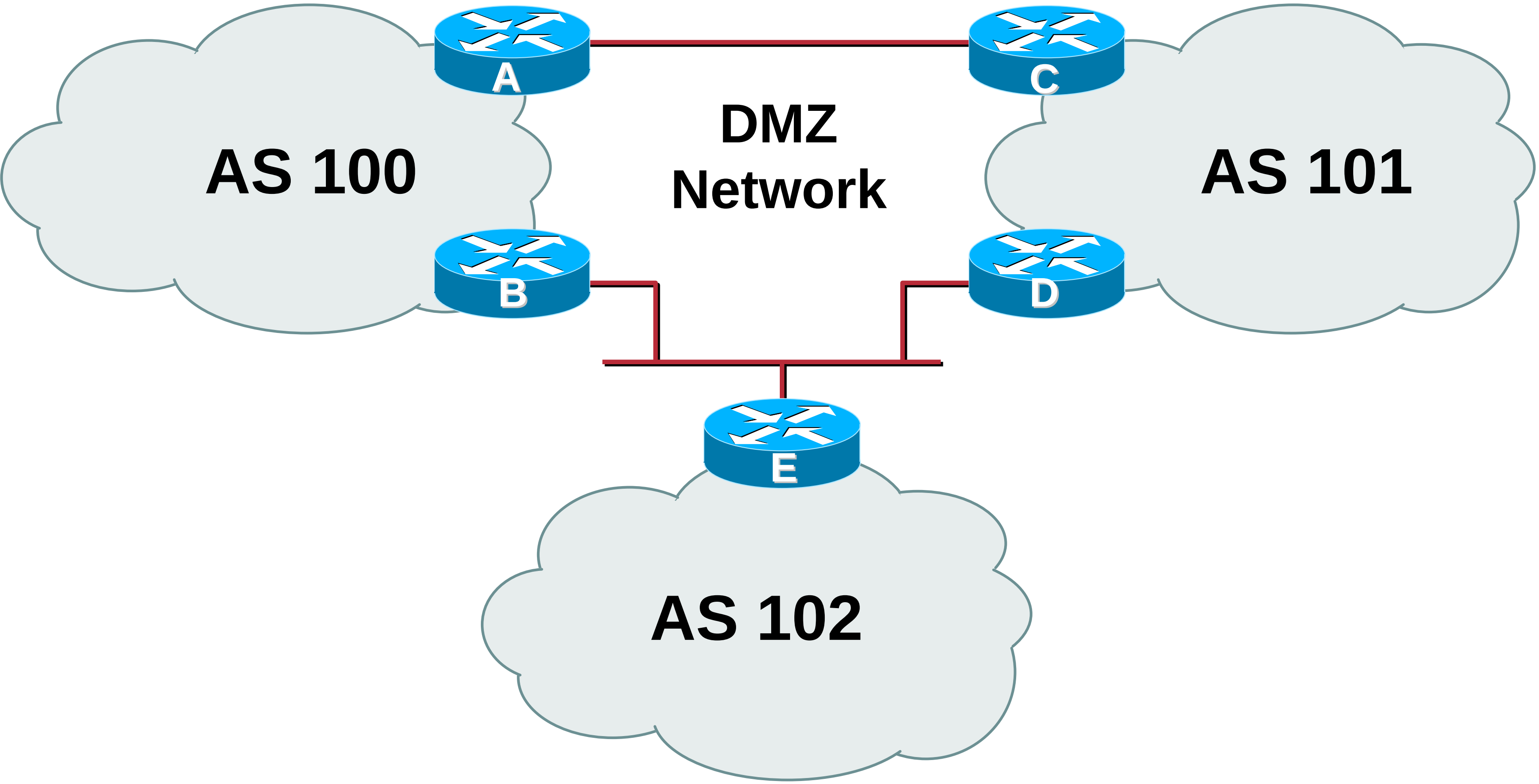
BGP Basics

Cisco.com



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

Demarcation Zone (DMZ)

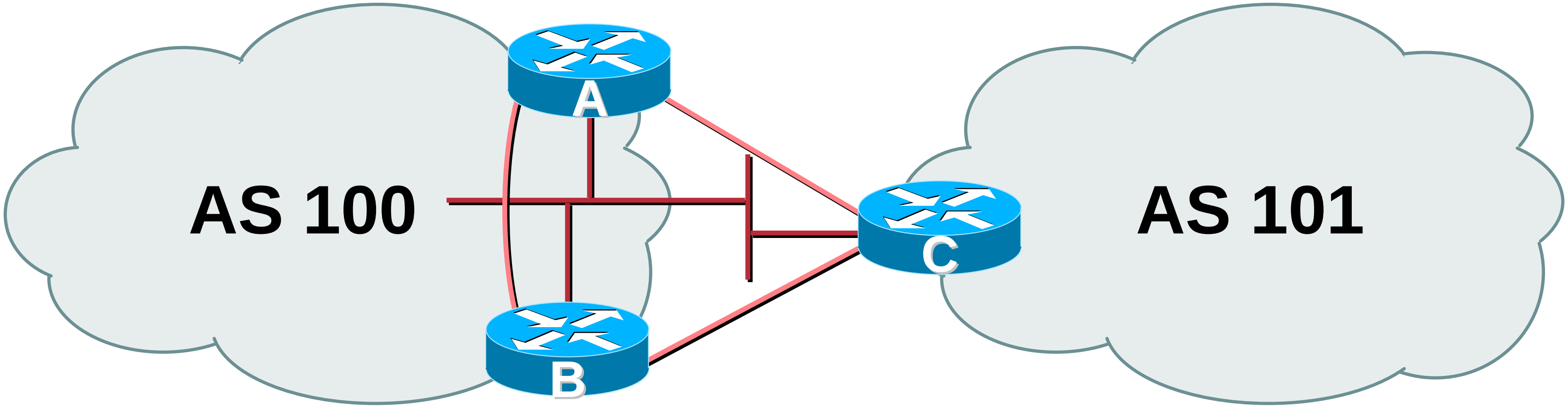


- **Shared network between ASes**

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**

External BGP Peering (eBGP)



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

Configuring External BGP

Cisco.com

Router A in AS100

```
interface ethernet 5/0
 ip address 222.222.10.2 255.255.255.240
router bgp 100
 network 220.220.8.0 mask 255.255.252.0
 neighbor 222.222.10.1 remote-as 101
 neighbor 222.222.10.1 prefix-list RouterC in
 neighbor 222.222.10.1 prefix-list RouterC out
```

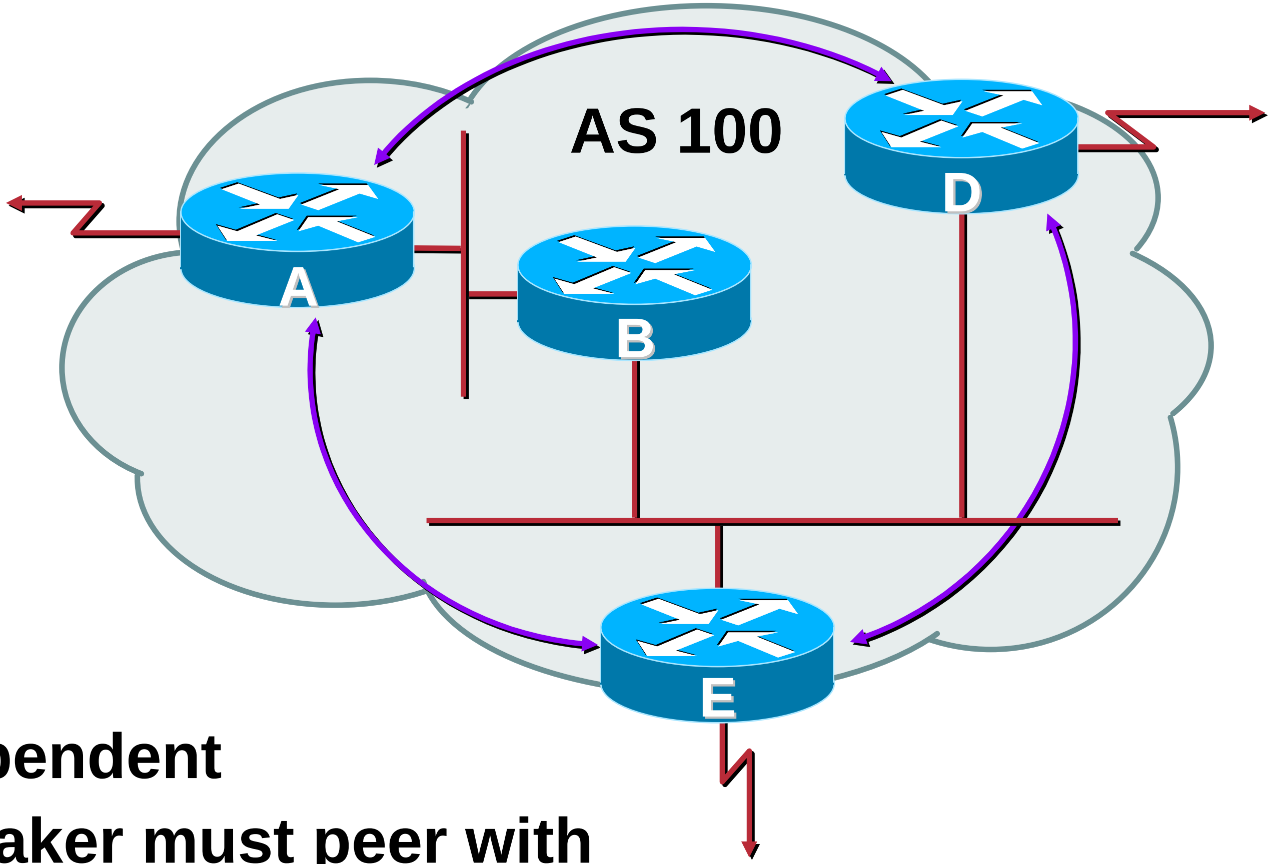
Router C in AS101

```
interface ethernet 1/0/0
 ip address 222.222.10.1 255.255.255.240
router bgp 101
 network 220.220.16.0 mask 255.255.240.0
 neighbor 222.222.10.2 remote-as 100
 neighbor 222.222.10.2 prefix-list RouterA in
 neighbor 222.222.10.2 prefix-list RouterA out
```

Internal BGP (iBGP)

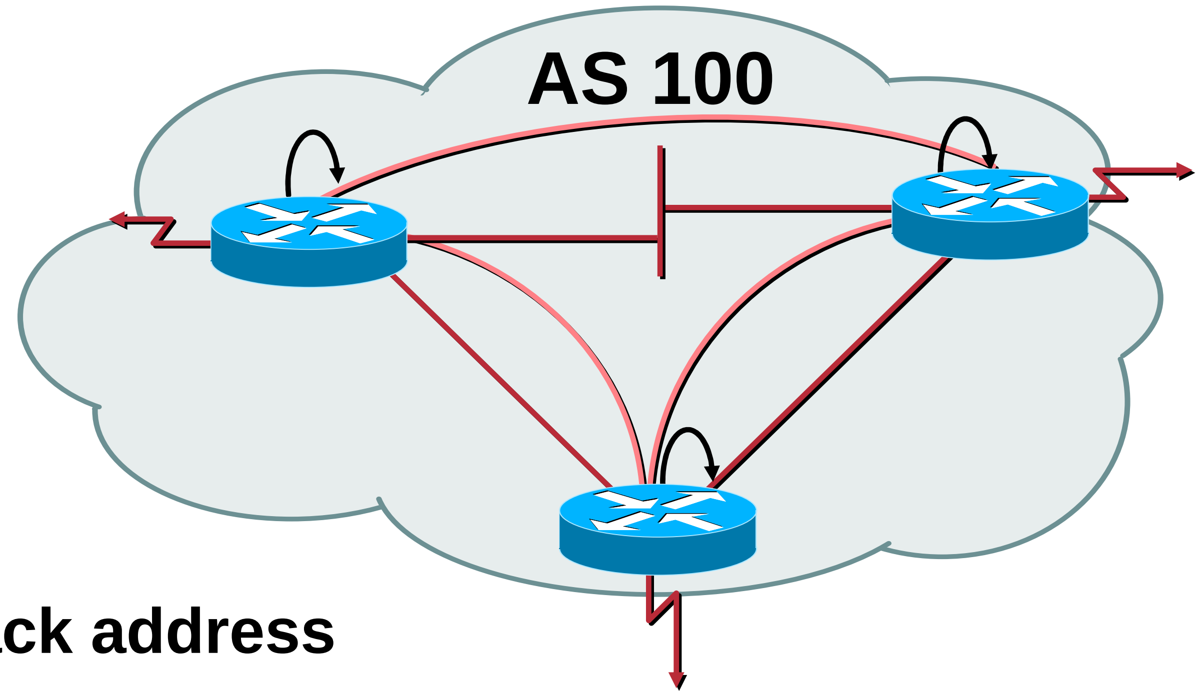
- BGP peer within the same AS
- Not required to be directly connected
- iBGP speakers need to be fully meshed
    - they originate connected networks
    - they do not pass on prefixes learned from other iBGP speakers

Internal BGP Peering (iBGP)



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

Peering to Loop-back Address



- **Peer with loop-back address**

  Loop-back interface does not go down – ever!

- **iBGP session is not dependent on state of a single interface**
- **iBGP session is not dependent on physical topology**

Configuring Internal BGP

Cisco.com

Router A

```
interface loopback 0
 ip address 215.10.7.1 255.255.255.255
router bgp 100
  network 220.220.1.0
  neighbor 215.10.7.2 remote-as 100
  neighbor 215.10.7.2 update-source loopback0
  neighbor 215.10.7.3 remote-as 100
  neighbor 215.10.7.3 update-source loopback0
```

Router B

```
interface loopback 0
 ip address 215.10.7.2 255.255.255.255
router bgp 100
  network 220.220.5.0
  neighbor 215.10.7.1 remote-as 100
  neighbor 215.10.7.1 update-source loopback0
  neighbor 215.10.7.3 remote-as 100
  neighbor 215.10.7.3 update-source loopback0
```

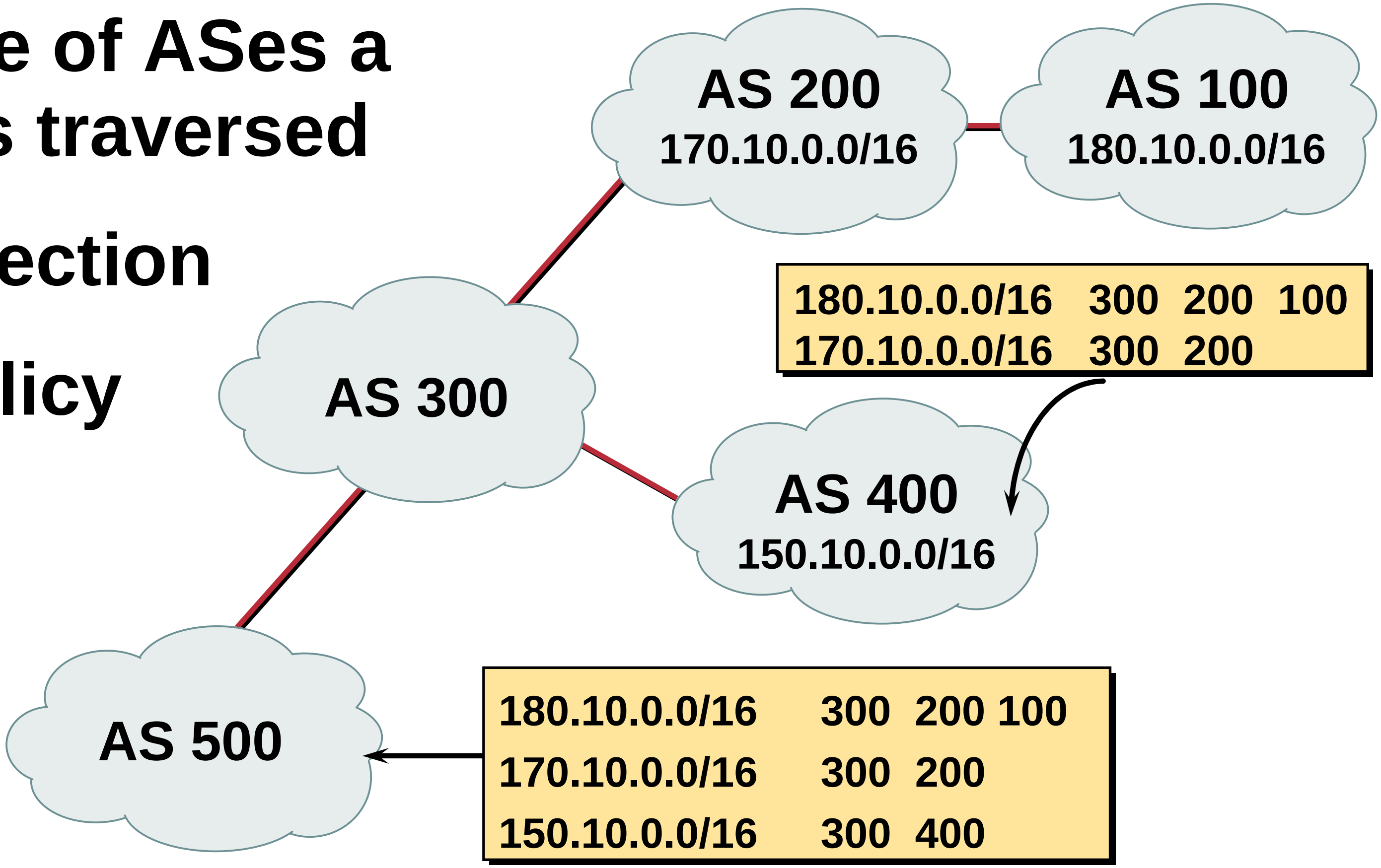
BGP Attributes

Recap

AS-Path

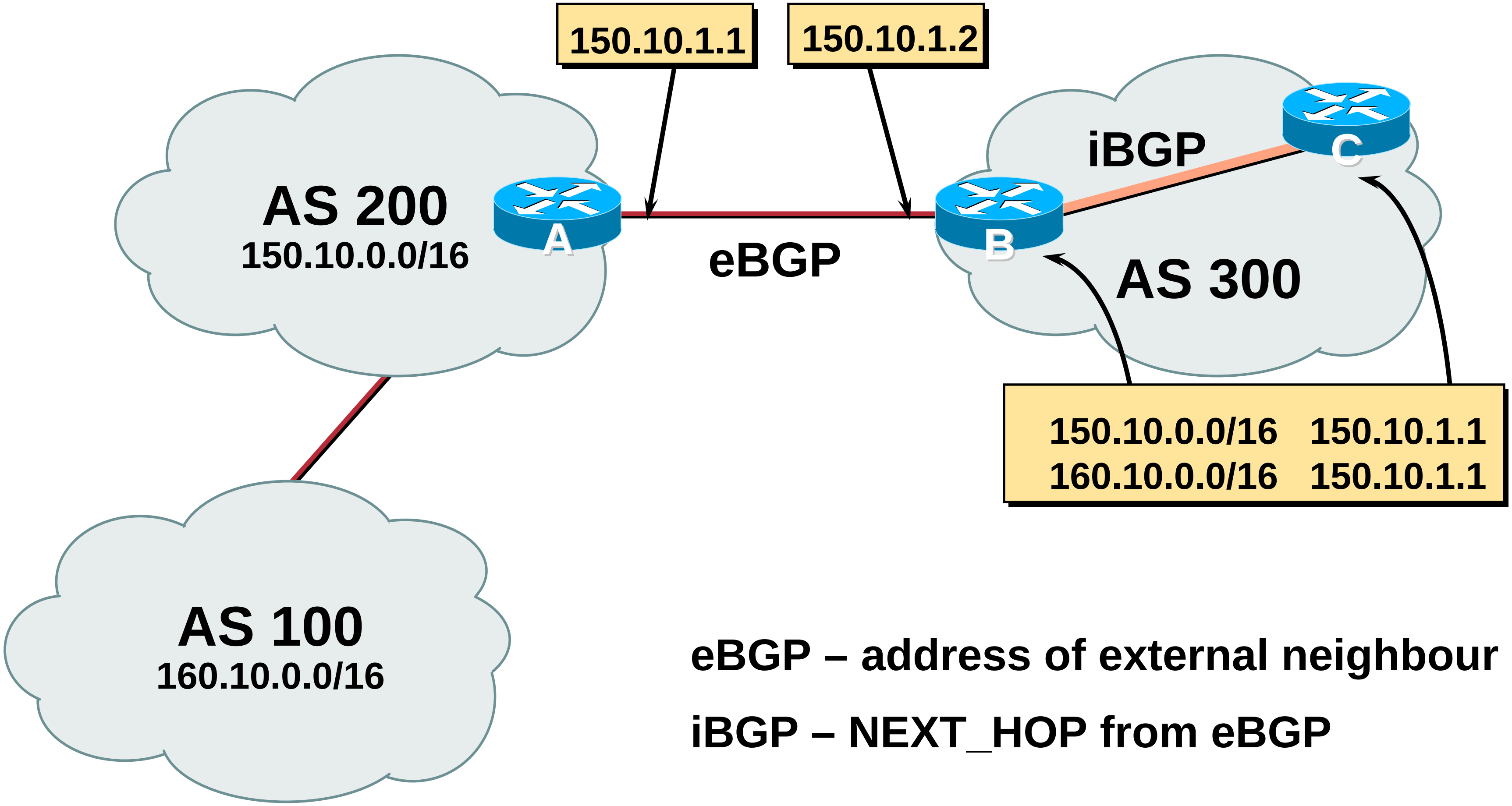
Cisco.com

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



Next Hop

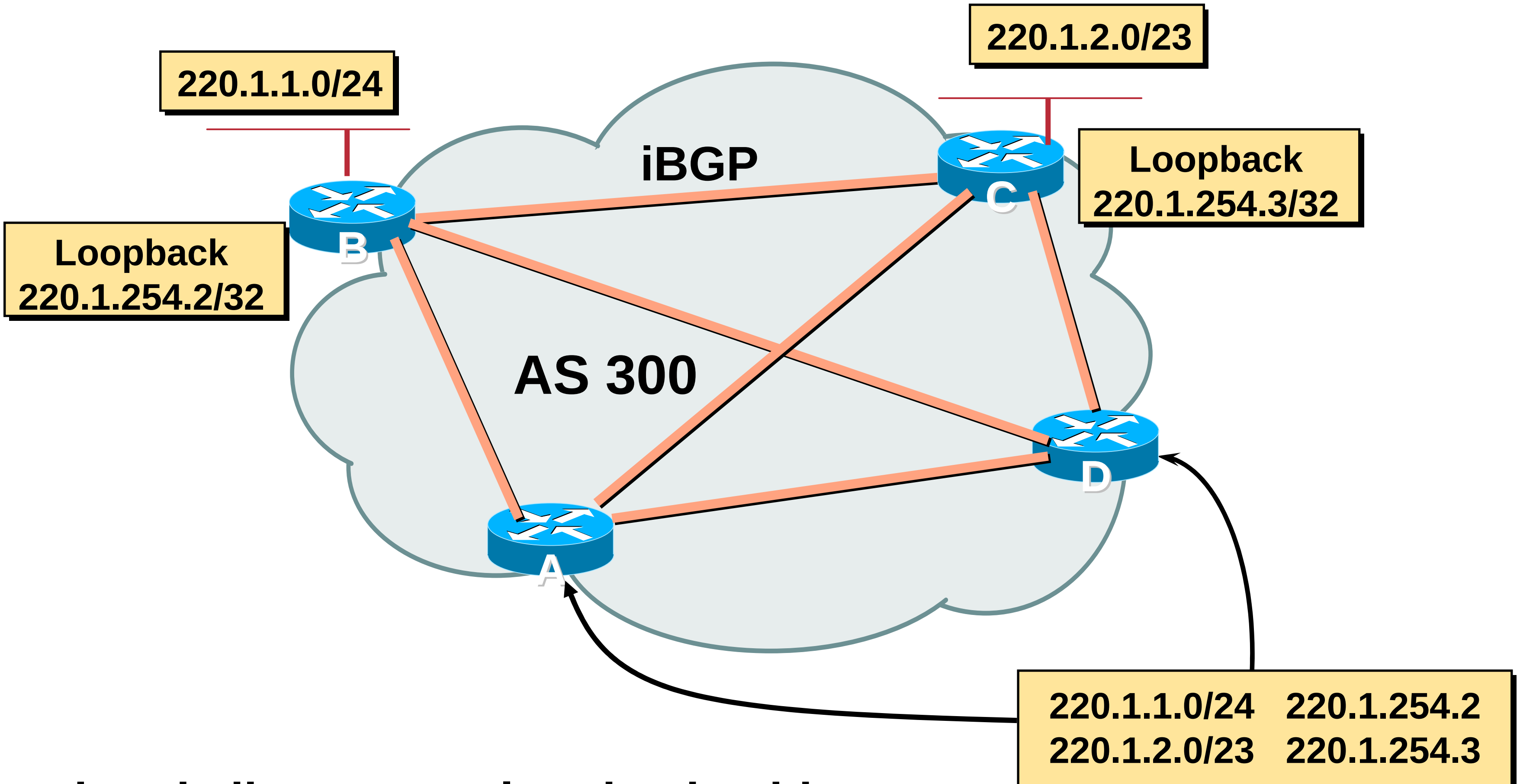
Cisco.com



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)

Cisco.com

- 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

Cisco.com

- 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

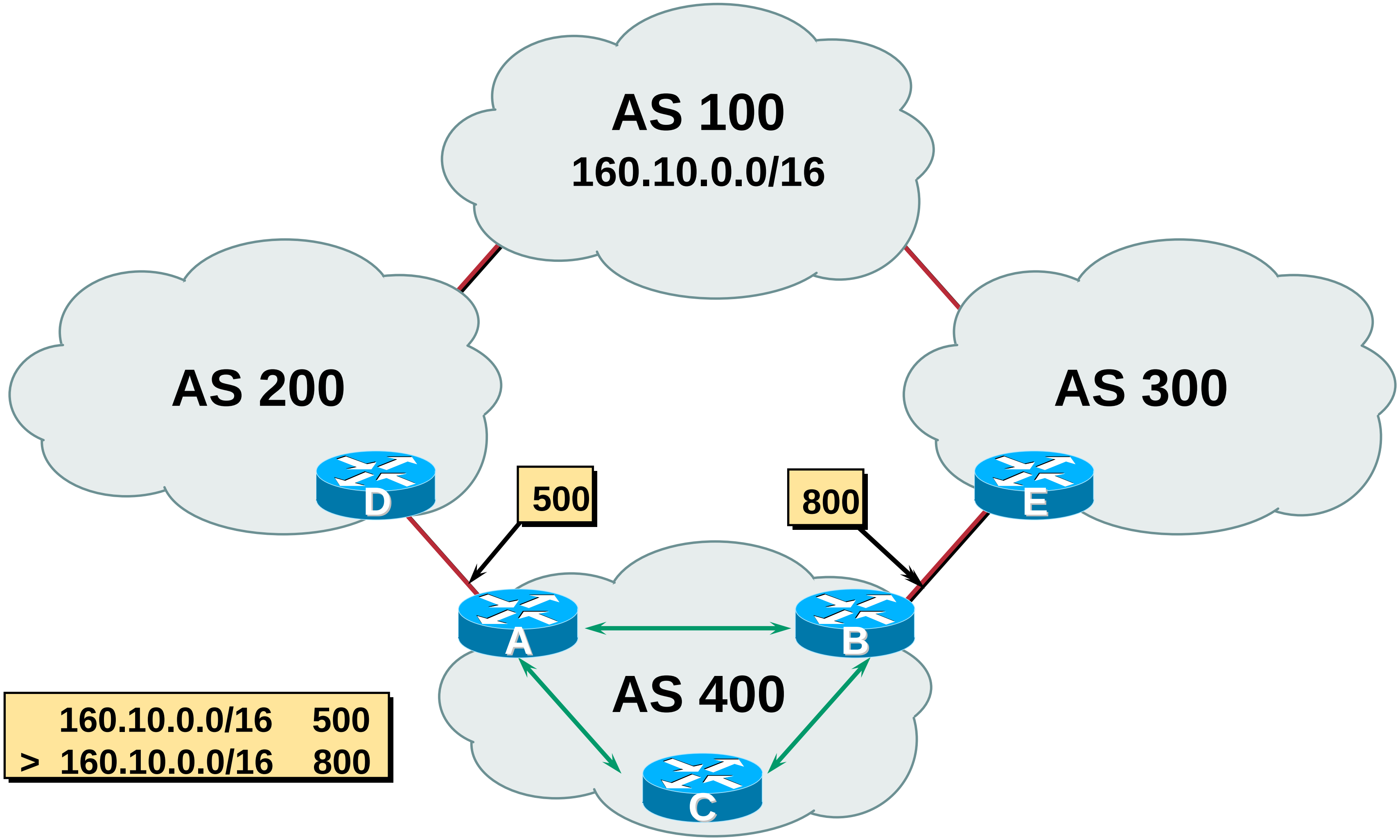
  incomplete – redistributed from another routing protocol

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

Cisco.com



	160.10.0.0/16	500
>	160.10.0.0/16	800

Local Preference

- **Local to an AS – non-transitive**

  Default local preference is 100 (IOS)

- **Used to influence BGP path selection**

  determines best path for *outbound* traffic

- **Path with highest local preference wins**

Local Preference

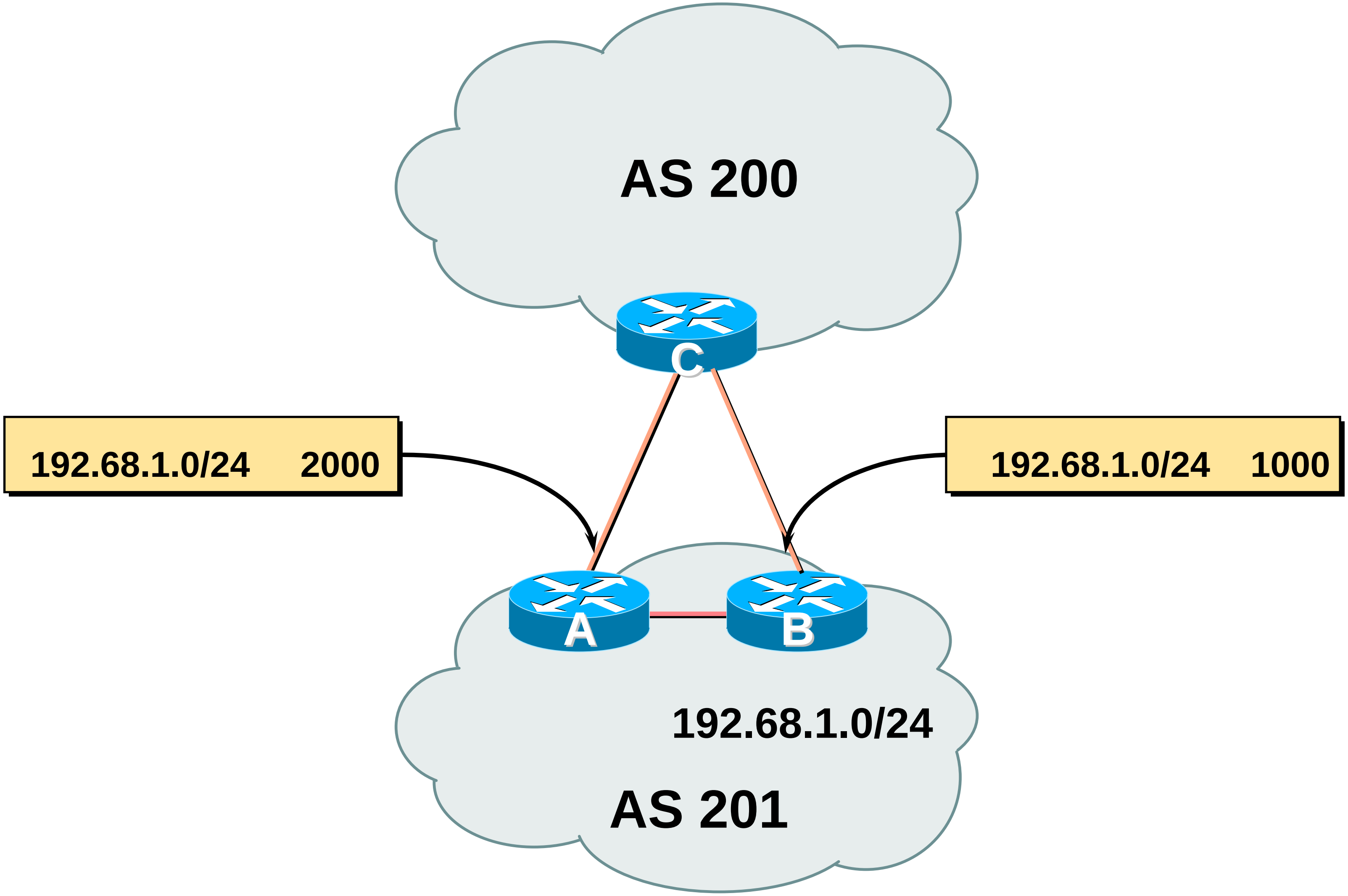
Cisco.com

- **Configuration of Router B:**

```
router bgp 400
 neighbor 220.5.1.1 remote-as 300
 neighbor 220.5.1.1 route-map local-pref in
!
route-map local-pref permit 10
 match ip address prefix-list MATCH
 set local-preference 800
!
ip prefix-list MATCH permit 160.10.0.0/16
```

Multi-Exit Discriminator (MED)

Cisco.com



Multi-Exit Discriminator

- **Inter-AS – non-transitive**
- **Used to convey the relative preference of entry points**

  determines best path for *inbound* traffic

- **Comparable if paths are from same AS**
- **IGP metric can be conveyed as MED**

  set metric-type internal in route-map

Multi-Exit Discriminator

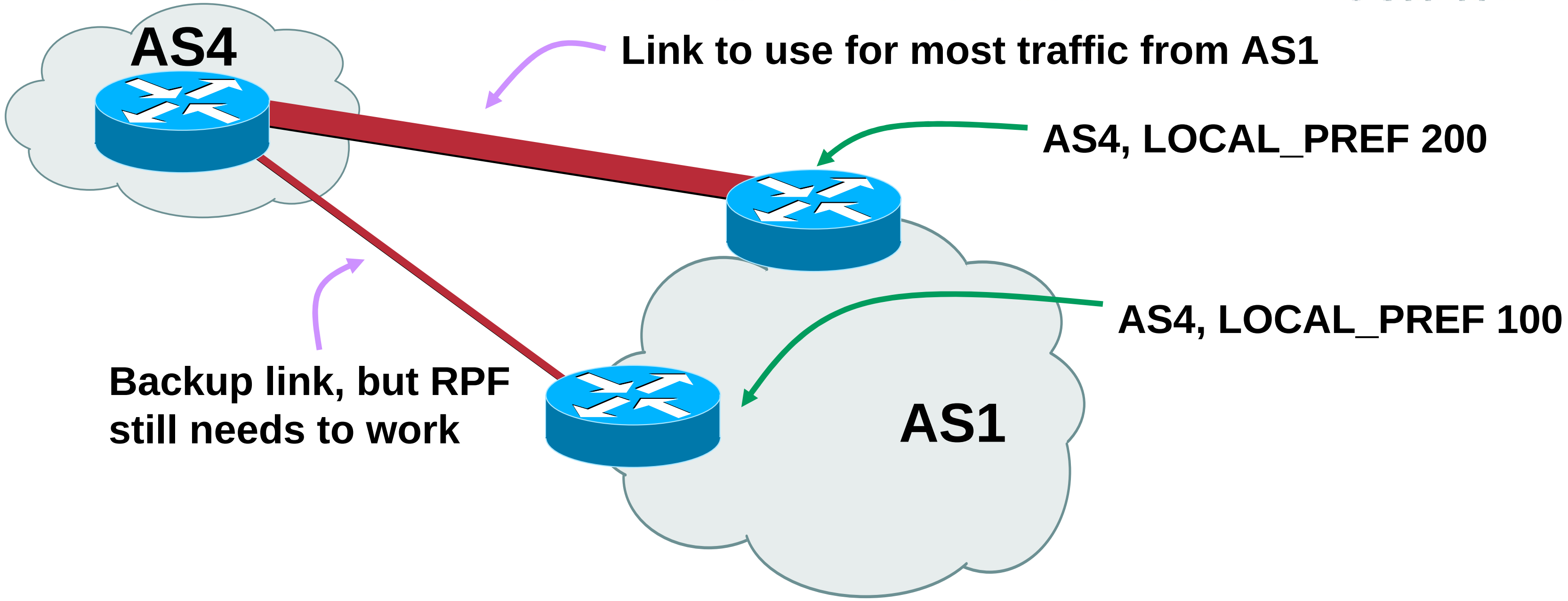
Cisco.com

- **Configuration of Router B:**

```
router bgp 400
 neighbor 220.5.1.1 remote-as 200
 neighbor 220.5.1.1 route-map set-med out
!
route-map set-med permit 10
 match ip address prefix-list MATCH
 set metric 1000
!
ip prefix-list MATCH permit 192.68.1.0/24
```

Weight – Used to Deploy RPF

Cisco.com



- **Local to router on which it's configured**
  - **Not really an attribute**
- **route-map: *set weight***
- **Highest weight wins over all valid paths**
- **Weight customer eBGP on edge routers to allow RPF to work correctly**

Community

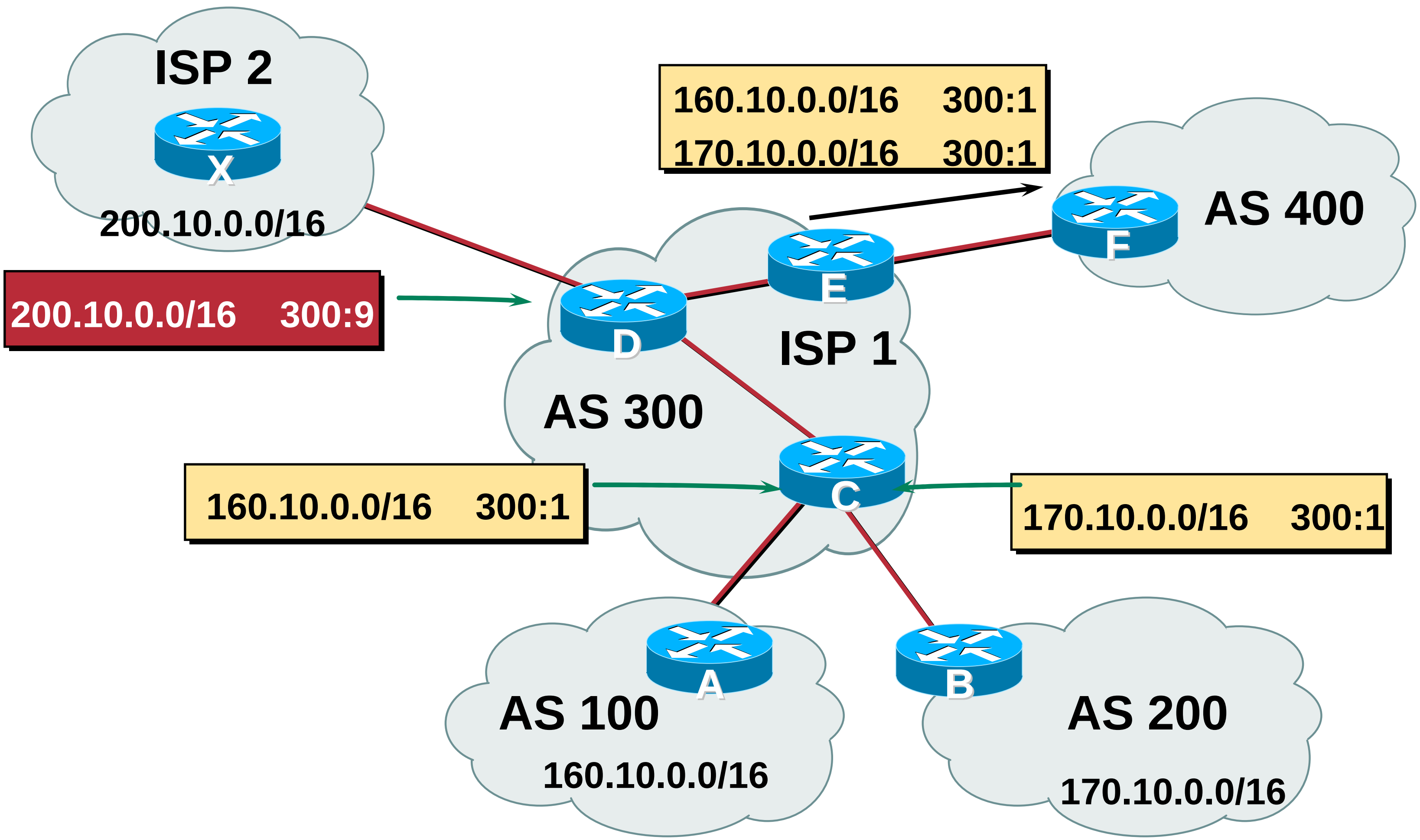
- **Communities described in RFC1997**
- **32 bit integer**

  Represented as two 16 bit integers (RFC1998)
- **Used to group destinations**

  Each destination could be member of multiple communities
- **Community attribute carried across AS's**
- **Very useful in applying policies**

Community

Cisco.com



Well-Known Communities

Cisco.com

- **no-export**

  do not advertise to eBGP peers

- **no-advertise**

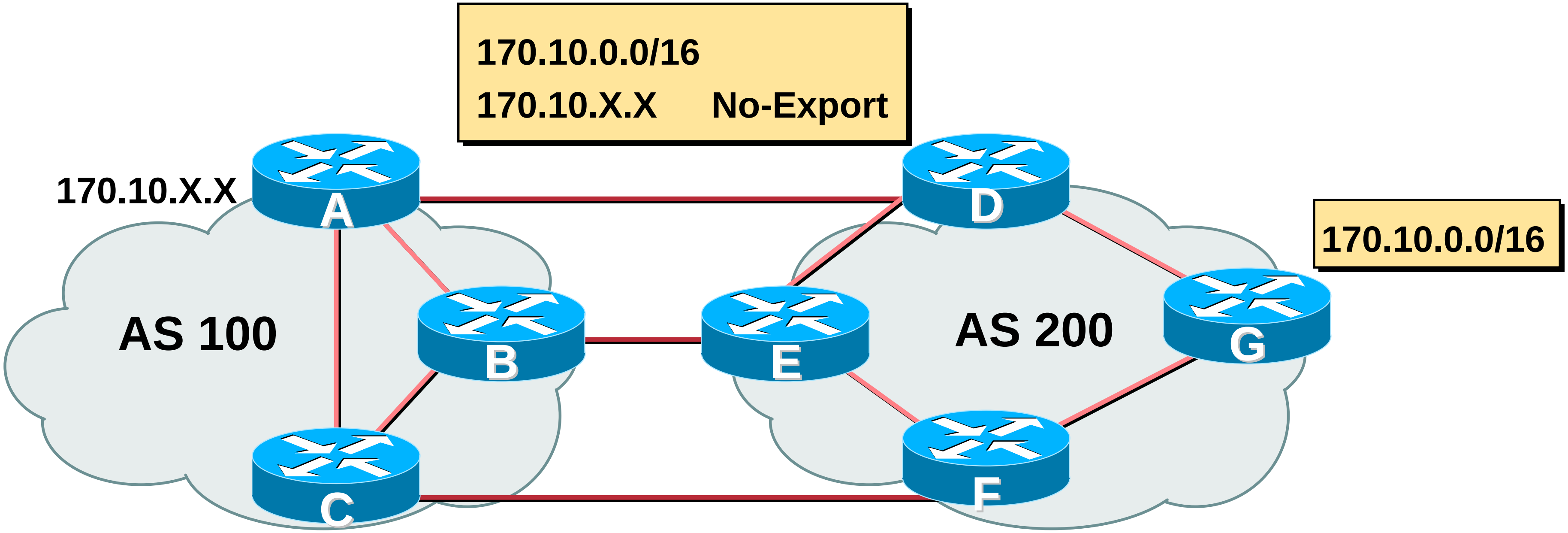
  do not advertise to any peer

- **local-AS**

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

No-Export Community

Cisco.com



- 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

BGP Path Selection Algorithm

Why Is This the Best Path?

BGP Path Selection Algorithm

Cisco.com

- 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 (continued)

Cisco.com

- 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 (continued)

Cisco.com

- 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

Applying Policy with BGP

Control!

Applying Policy with BGP

Cisco.com

- **Applying Policy**
  - **Decisions based on AS path, community or the prefix**
  - **Rejecting/accepting selected routes**
  - **Set attributes to influence path selection**
- **Tools:**
  - **Prefix-list (filter prefixes)**
  - **Filter-list (filter ASes)**
  - **Route-maps and communities**

Policy Control
Prefix List

Cisco.com

- **Filter routes based on prefix**
- **Inbound and Outbound**

```
router bgp 200
 neighbor 220.200.1.1 remote-as 210
 neighbor 220.200.1.1 prefix-list PEER-IN in
 neighbor 220.200.1.1 prefix-list PEER-OUT out
!
ip prefix-list PEER-IN deny 218.10.0.0/16
ip prefix-list PEER-IN permit 0.0.0.0/0 le 32
ip prefix-list PEER-OUT permit 215.7.0.0/16
```

Policy Control
Filter List

- **Filter routes based on AS path**
- **Inbound and Outbound**

```
router bgp 100
 neighbor 220.200.1.1 remote-as 210
 neighbor 220.200.1.1 filter-list 5 out
 neighbor 220.200.1.1 filter-list 6 in
!
ip as-path access-list 5 permit ^200$
ip as-path access-list 6 permit ^150$
```

Policy Control
Regular Expressions

Cisco.com

- **Like Unix regular expressions**

.	Match one character
*	Match any number of preceding expression
+	Match at least one of preceding expression
^	Beginning of line
$	End of line
_	Beginning, end, white-space, brace
|	Or
()	brackets to contain expression

Policy Control
Regular Expressions

- **Simple Examples**

.*	**Match anything**
.+	**Match at least one character**
^$	**Match routes local to this AS**
_1800$	**Originated by 1800**
^1800_	**Received from 1800**
_1800_	**Via 1800**
_790_1800_	**Passing through 1800 then 790**
_(1800_)+	**Match at least one of 1800 in sequence**
_\(65350\)_	**Via 65350 (confederation AS)**

Policy Control
Route Maps

Cisco.com

- **A route-map is like a "programme" for IOS**
- **Has "line" numbers, like programmes**
- **Each line is a separate condition/action**
- **Concept is basically:**

  if *match* then do *expression* and *exit*

  else

  if *match* then do *expression* and *exit*

  else *etc*

Policy Control Route Maps

- Example using prefix-lists

```
router bgp 100
 neighbor 1.1.1.1 route-map infilter in
!
route-map infilter permit 10
 match ip address prefix-list HIGH-PREF
 set local-preference 120
!
route-map infilter permit 20
 match ip address prefix-list LOW-PREF
 set local-preference 80
!
route-map infilter permit 30
!
ip prefix-list HIGH-PREF permit 10.0.0.0/8
ip prefix-list LOW-PREF permit 20.0.0.0/8
```

Policy Control
Route Maps

Cisco.com

- **Example using filter lists**

```
router bgp 100
 neighbor 220.200.1.2 route-map filter-on-as-path in
!
route-map filter-on-as-path permit 10
 match as-path 1
 set local-preference 80
!
route-map filter-on-as-path permit 20
 match as-path 2
 set local-preference 200
!
route-map filter-on-as-path permit 30
!
ip as-path access-list 1 permit _150$
ip as-path access-list 2 permit _210_
```

Policy Control
Route Maps

Cisco.com

- **Example configuration of AS-PATH prepend**

```
router bgp 300
 network 215.7.0.0
 neighbor 2.2.2.2 remote-as 100
 neighbor 2.2.2.2 route-map SETPATH out
!
route-map SETPATH permit 10
 set as-path prepend 300 300
```

- **Use your own AS number when prepending**

  Otherwise BGP loop detection may cause disconnects

Policy Control
Setting Communities

Cisco.com

- Example Configuration

```
router bgp 100
 neighbor 220.200.1.1 remote-as 200
 neighbor 220.200.1.1 send-community
 neighbor 220.200.1.1 route-map set-community out
!
route-map set-community permit 10
 match ip address prefix-list NO-ANNOUNCE
 set community no-export
!
route-map set-community permit 20
!
ip prefix-list NO-ANNOUNCE permit 172.168.0.0/16 ge 17
```

BGP Capabilities

Extending BGP

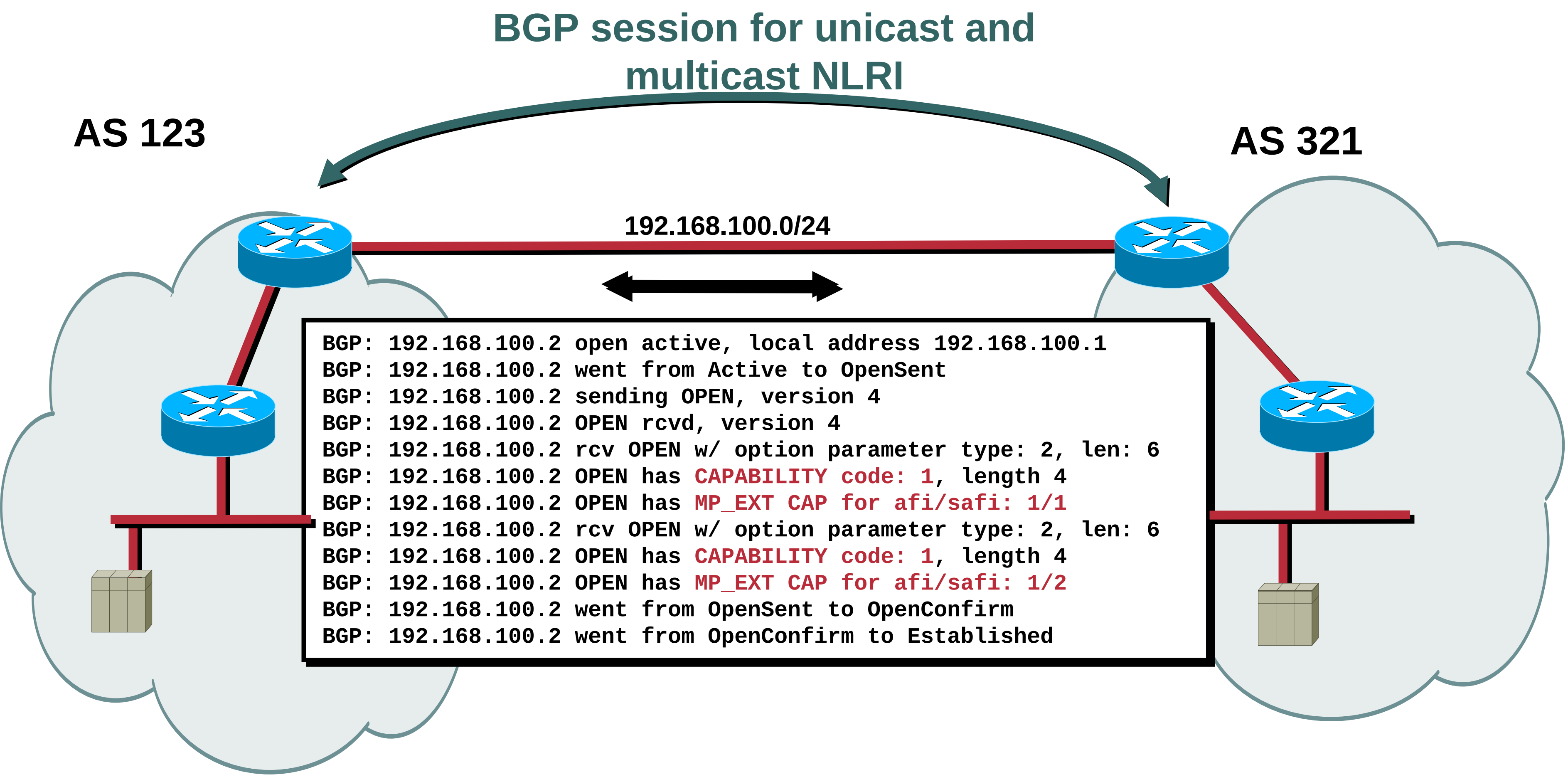
BGP Capabilities

Cisco.com

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

```
 0   Reserved                                     [RFC2842]
 1   Multiprotocol Extensions for BGP-4           [RFC2858]
 2   Route Refresh Capability for BGP-4           [RFC2918]
 3   Cooperative Route Filtering Capability       []
 4   Multiple routes to a destination capability [RFC3107]
64   Graceful Restart Capability                  []
```

BGP Capabilities Negotiation




```
BGP: 192.168.100.2 open active, local address 192.168.100.1
BGP: 192.168.100.2 went from Active to OpenSent
BGP: 192.168.100.2 sending OPEN, version 4
BGP: 192.168.100.2 OPEN rcvd, version 4
BGP: 192.168.100.2 rcv OPEN w/ option parameter type: 2, len: 6
BGP: 192.168.100.2 OPEN has CAPABILITY code: 1, length 4
BGP: 192.168.100.2 OPEN has MP_EXT CAP for afi/safi: 1/1
BGP: 192.168.100.2 rcv OPEN w/ option parameter type: 2, len: 6
BGP: 192.168.100.2 OPEN has CAPABILITY code: 1, length 4
BGP: 192.168.100.2 OPEN has MP_EXT CAP for afi/safi: 1/2
BGP: 192.168.100.2 went from OpenSent to OpenConfirm
BGP: 192.168.100.2 went from OpenConfirm to Established
```

BGP for Internet Service Providers

- BGP Basics (quick recap)
- Scaling BGP
- Deploying BGP in an ISP network
- Multihoming Examples
- Using Communities

BGP Scaling Techniques

BGP Scaling Techniques

Cisco.com

- How to scale iBGP mesh beyond a few peers?
- How to implement new policy without causing flaps and route churning?
- How to reduce the overhead on the routers?
- How to keep the network stable, scalable, as well as simple?

BGP Scaling Techniques

- Route Refresh
- Peer groups
- Route flap damping
- Route Reflectors & Confederations

Route Refresh

Route Refresh

Cisco.com

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:
  - Consumes CPU
  - Severely disrupts connectivity for all networks

Solution:

- Route Refresh

Route Refresh Capability

Cisco.com

- Facilitates non-disruptive policy changes
- No configuration is needed
- No additional memory is used
- Requires peering routers to support “route refresh capability” – RFC2918
- clear ip bgp x.x.x.x in tells peer to resend full BGP announcement
- clear ip bgp x.x.x.x out resends full BGP announcement to peer

Dynamic Reconfiguration

Cisco.com

- Use Route Refresh capability if supported
  - find out from “show ip bgp neighbor”
  - Non-disruptive, “Good For the Internet”
- Otherwise use Soft Reconfiguration IOS feature
- Only hard-reset a BGP peering as a last resort
  - Consider the impact to be equivalent to a router reboot

Soft Reconfiguration

- Router normally stores prefixes which have been received from peer after policy application
    - Enabling soft-reconfiguration means router also stores prefixes/attributes received prior to any policy application
- New policies can be activated without tearing down and restarting the peering session
- Configured on a per-neighbour basis
- Uses more memory to keep prefixes whose attributes have been changed or have not been accepted

Configuring Soft Reconfiguration

Cisco.com

```
router bgp 100
 neighbor 1.1.1.1 remote-as 101
 neighbor 1.1.1.1 route-map infilter in
 neighbor 1.1.1.1 soft-reconfiguration inbound
```

! *Outbound does not need to be configured* !

Then when we change the policy, we issue an exec command

```
clear ip bgp 1.1.1.1 soft [in | out]
```

Peer Groups

Peer Groups

Cisco.com

Without peer groups

- iBGP neighbours receive same update
- Large iBGP mesh slow to build
- Router CPU wasted on repeat calculations

Solution – peer groups!

- Group peers with same outbound policy
- Updates are generated once per group

Peer Groups – Advantages

- Makes configuration easier
- Makes configuration less prone to error
- Makes configuration more readable
- Lower router CPU load
- iBGP mesh builds more quickly
- Members can have different inbound policy
- Can be used for eBGP neighbours too!

Configuring Peer Group

Cisco.com

```
router bgp 100
 neighbor ibgp-peer peer-group
 neighbor ibgp-peer remote-as 100
 neighbor ibgp-peer update-source loopback 0
 neighbor ibgp-peer send-community
 neighbor ibgp-peer route-map outfilter out
 neighbor 1.1.1.1 peer-group ibgp-peer
 neighbor 2.2.2.2 peer-group ibgp-peer
 neighbor 2.2.2.2 route-map  infilter in
 neighbor 3.3.3.3 peer-group ibgp-peer
```

! *note how 2.2.2.2 has different inbound filter from peer-group* !

Configuring Peer Group

Cisco.com

```
router bgp 100
 neighbor external-peer peer-group
 neighbor external-peer send-community
 neighbor external-peer route-map set-metric out
 neighbor 160.89.1.2 remote-as 200
 neighbor 160.89.1.2 peer-group external-peer
 neighbor 160.89.1.4 remote-as 300
 neighbor 160.89.1.4 peer-group external-peer
 neighbor 160.89.1.6 remote-as 400
 neighbor 160.89.1.6 peer-group external-peer
 neighbor 160.89.1.6 filter-list infilter in
```

Route Flap Damping

Stabilising the Network

Route Flap Damping

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

Route Flap Damping (continued)

Cisco.com

- **Requirements**

  Fast convergence for normal route changes

  History predicts future behaviour

  Suppress oscillating routes

  Advertise stable routes

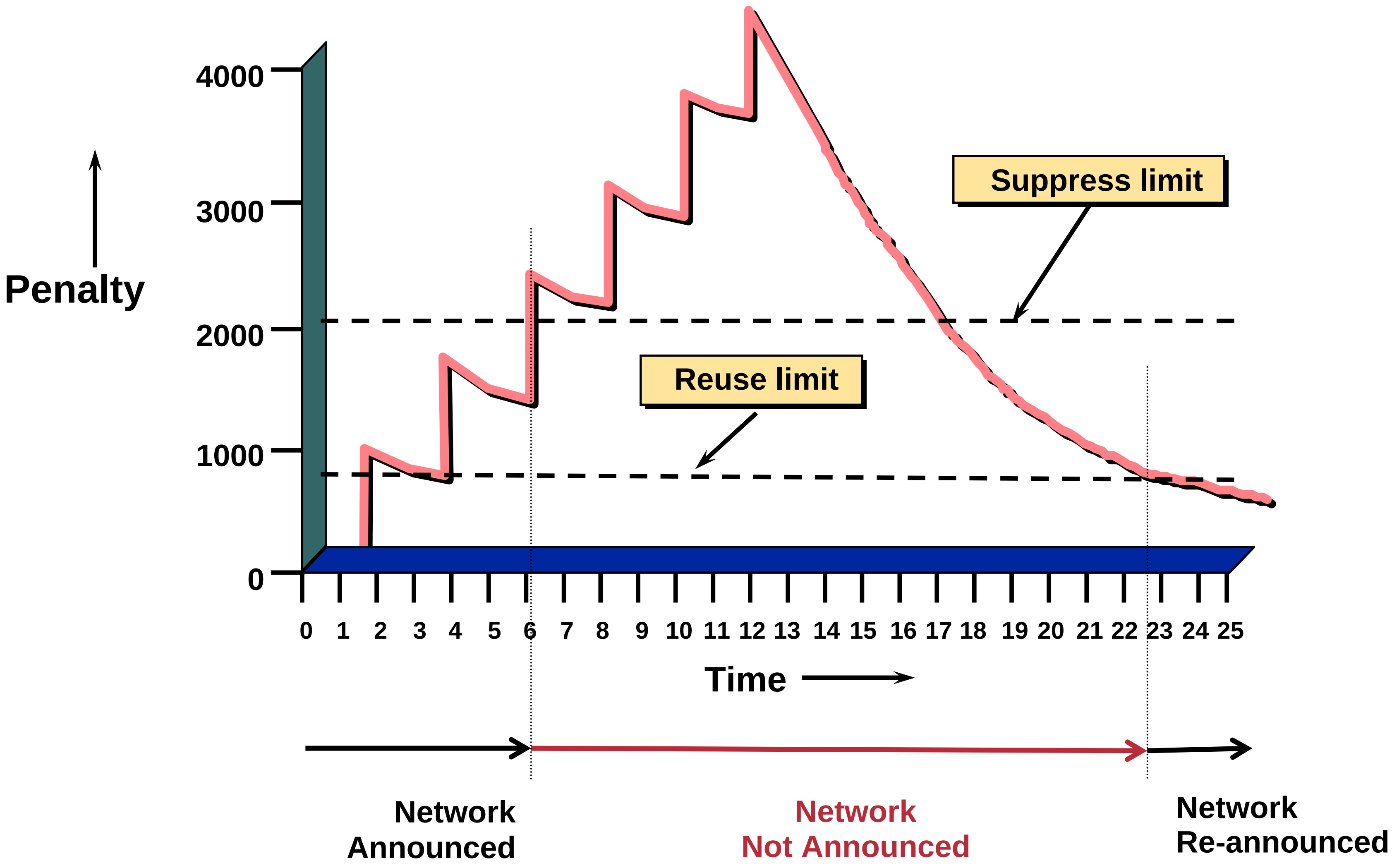
- **Documented in RFC2439**

Operation

Cisco.com

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

Operation


Penalty
4000
3000
2000
1000
0
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
- Controlled by:

  Half-life (default 15 minutes)

  reuse-limit (default 750)

  suppress-limit (default 2000)

  maximum suppress time (default 60 minutes)

Configuration

Cisco.com

Fixed damping

```
router bgp 100
 bgp dampening [<half-life> <reuse-value> <suppress-
  penalty> <maximum suppress time>]
```

Selective and variable damping

```
 bgp dampening [route-map <name>]
```

Variable damping

recommendations for ISPs

http://www.ripe.net/docs/ripe-229.html

Operation

- Care required when setting parameters
- Penalty must be less than reuse-limit at the maximum suppress time
- Maximum suppress time and half life must allow penalty to be larger than suppress limit

Configuration

- **Examples - ✘**

  bgp dampening 30 750 3000 60

  reuse-limit of 750 means maximum possible penalty is 3000 – no prefixes suppressed as penalty cannot exceed suppress-limit

- **Examples - ✓**

  bgp dampening 30 2000 3000 60

  reuse-limit of 2000 means maximum possible penalty is 8000 – suppress limit is easily reached

Maths!

Cisco.com

- **Maximum value of penalty is**

$$\text{max-penalty} = \text{reuse-limit} \times 2^{\left(\frac{\text{max-suppress-time}}{\text{half-life}}\right)}$$

- **Always make sure that suppress-limit is LESS than max-penalty otherwise there will be no flap damping**

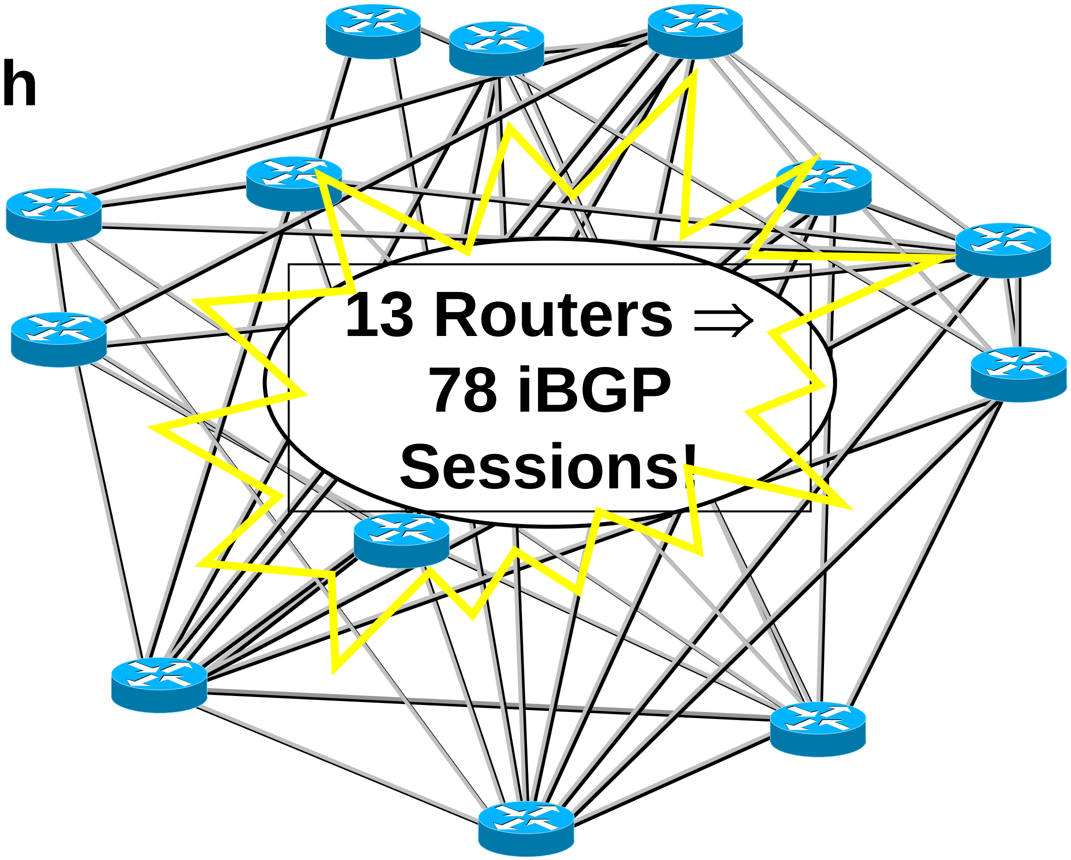
Route Reflectors and Confederations

Scaling iBGP mesh

Cisco.com

Avoid n(n-1)/2 iBGP mesh

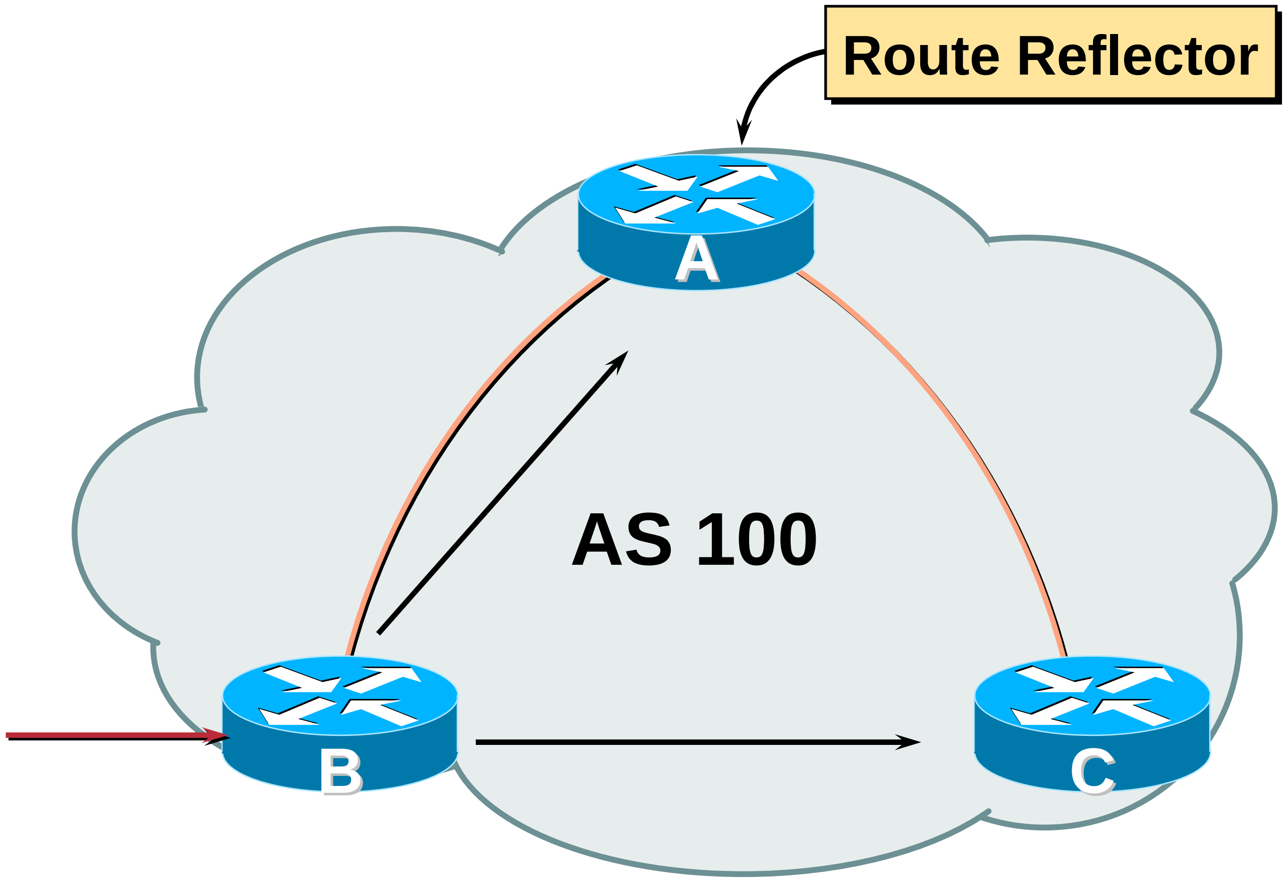
n=1000 ⇒ nearly half a million ibgp sessions!



Two solutions

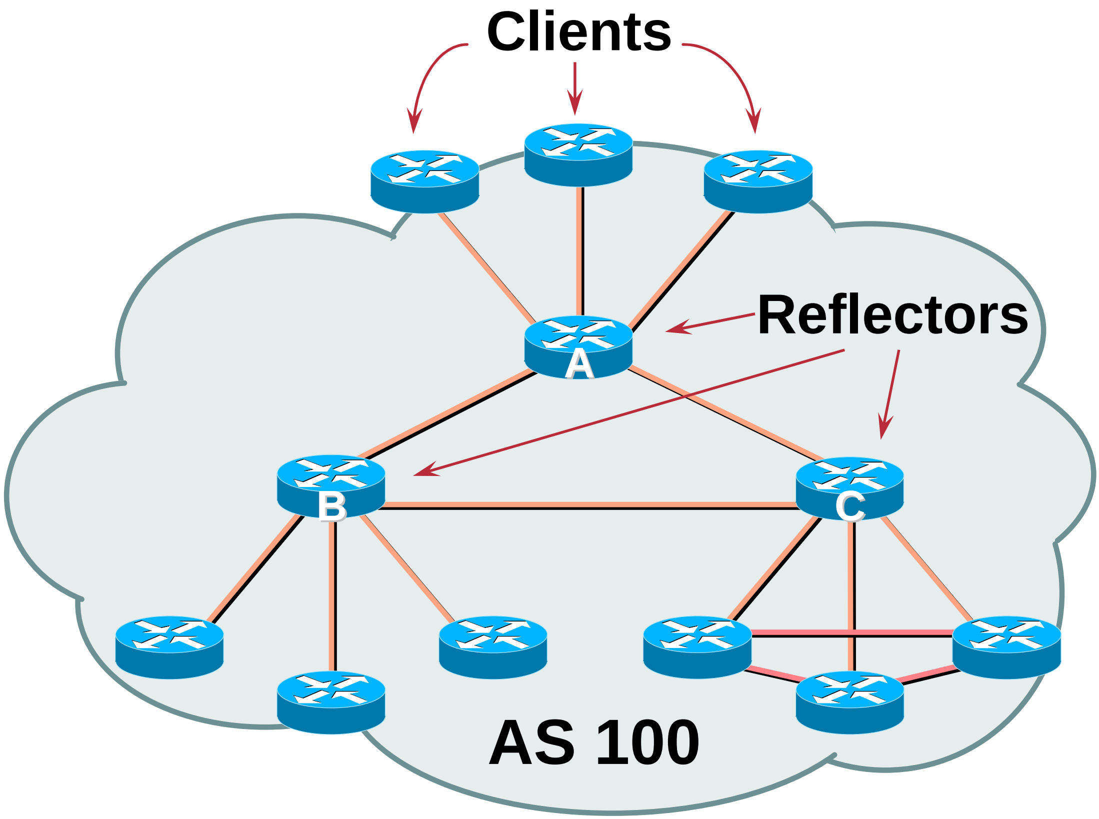
- Route reflector – simpler to deploy and run
- Confederation – more complex, corner case benefits

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

Cisco.com

- **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

  Cluster-id is automatically set from router-id (address of loopback)

  Do NOT use *bgp cluster-id x.x.x.x*

Route Reflectors: Redundancy

Cisco.com

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

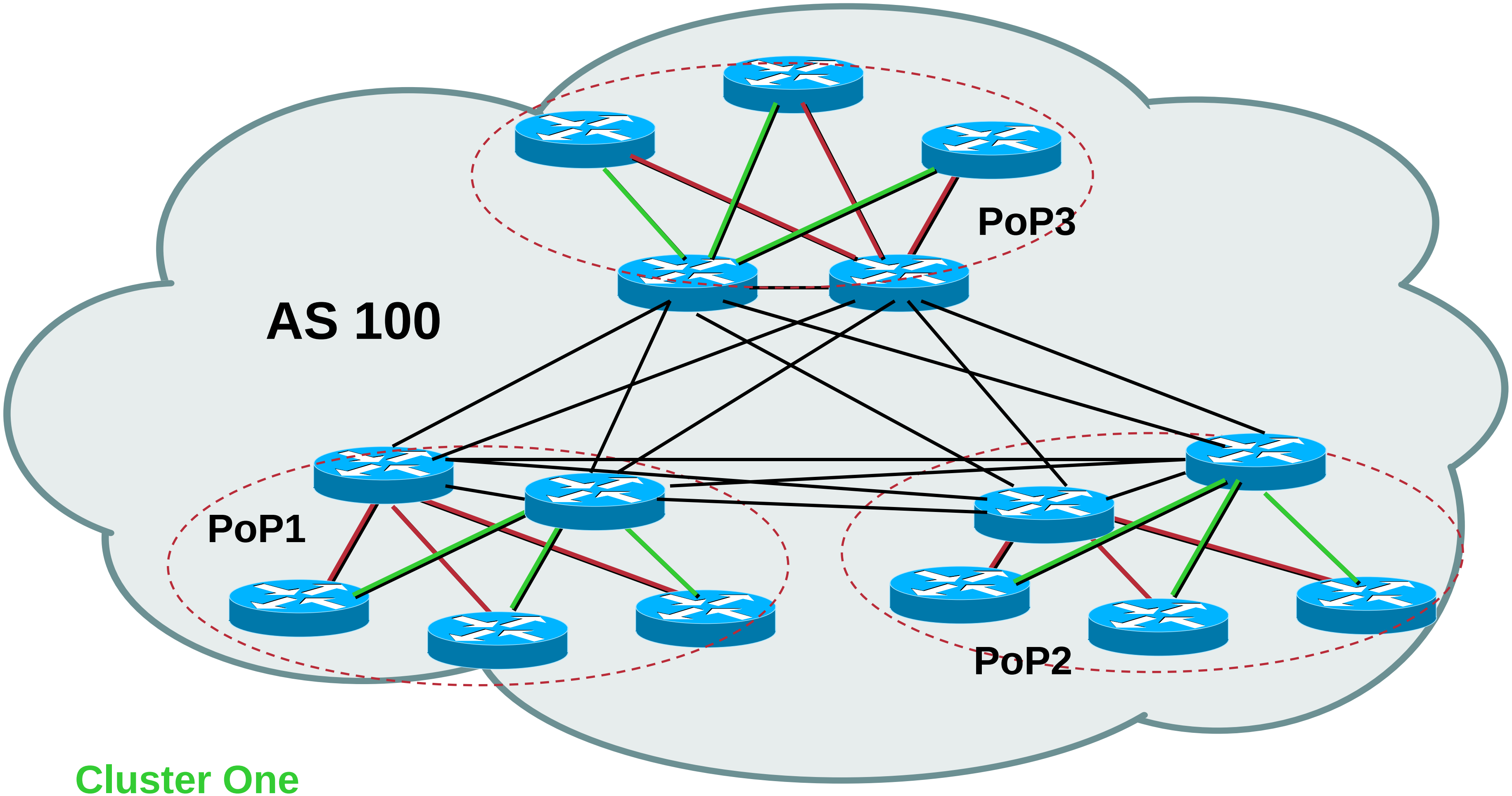
  All RRs in the cluster must have the same cluster-id (otherwise it is a different cluster)

- **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

Cisco.com

- 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

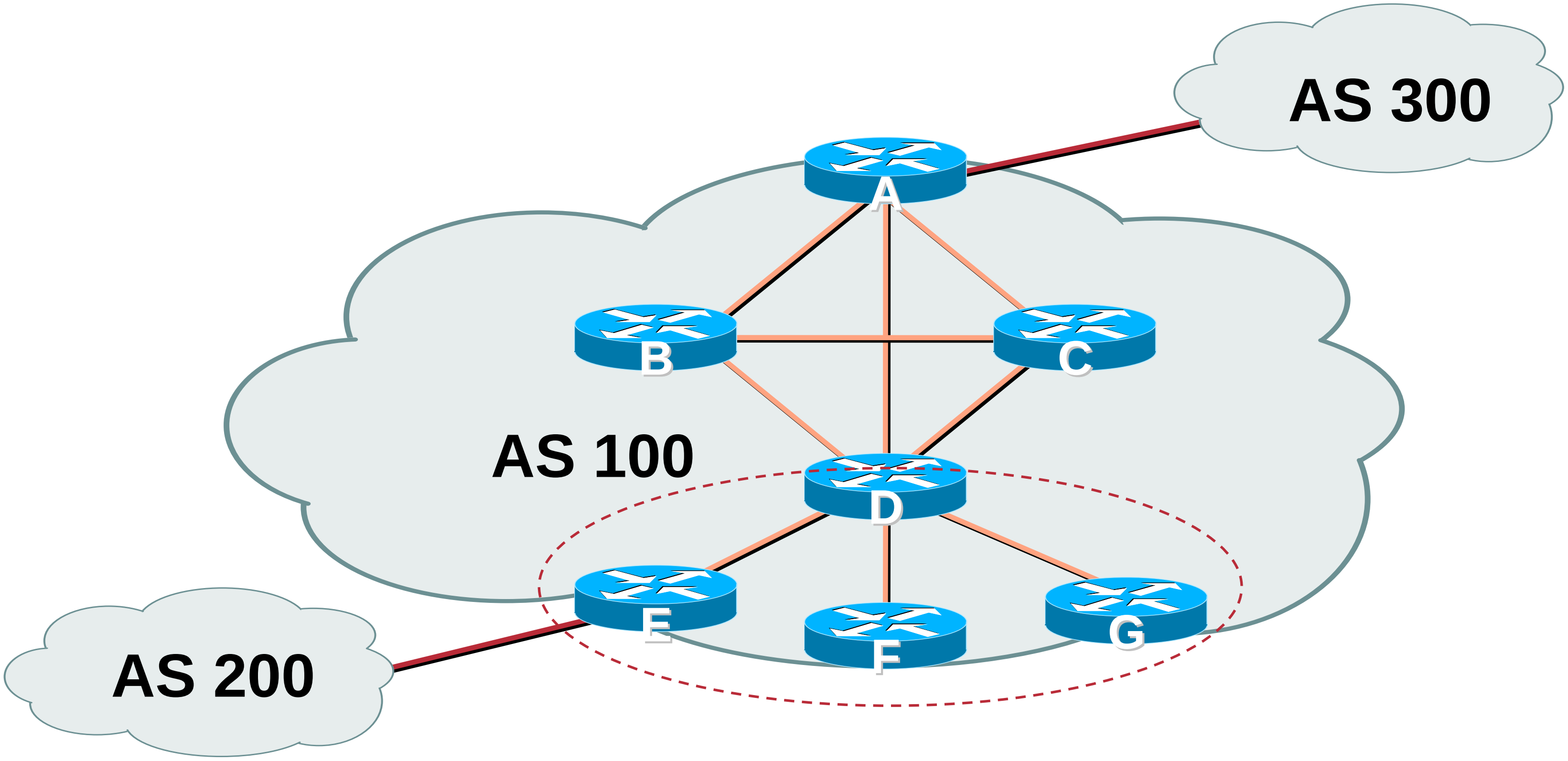
Route Reflectors: Migration

Cisco.com

- **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

Cisco.com



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

Configuring a Route Reflector

Cisco.com

```
router bgp 100
 neighbor 1.1.1.1 remote-as 100
 neighbor 1.1.1.1 route-reflector-client
 neighbor 2.2.2.2 remote-as 100
 neighbor 2.2.2.2 route-reflector-client
 neighbor 3.3.3.3 remote-as 100
 neighbor 3.3.3.3 route-reflector-client
 neighbor 4.4.4.4 remote-as 100
 neighbor 4.4.4.4 route-reflector-client
```

Confederations

Cisco.com

- **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.)

Cisco.com

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

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

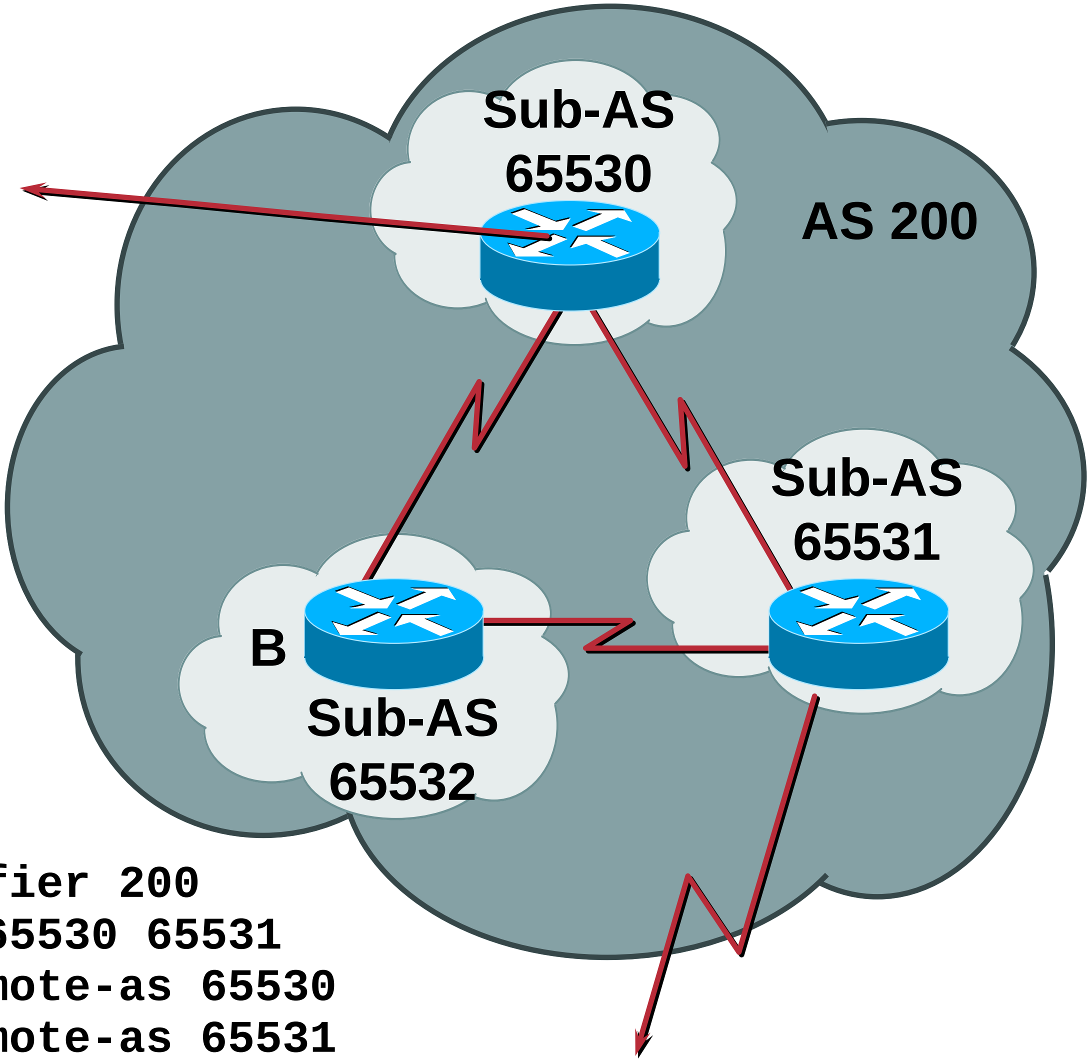
- **iBGP speakers in sub-AS are fully meshed**

  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.)

Cisco.com



- **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
```

Route Propagation Decisions

Cisco.com

- 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

Confederations (cont.)

Cisco.com

- **Example (cont.):**

```
BGP table version is 78, local router ID is 141.153.17.1
Status codes: s suppressed, d damped, h history, * valid, >
best, i - internal
Origin codes: i - IGP, e - EGP, ? - incomplete
Network          Next Hop      Metric LocPrf Weight Path
*> 10.0.0.0      141.153.14.3    0     100       0    (65531) 1 i
*> 141.153.0.0 141.153.30.2    0     100       0    (65530) i
*> 144.10.0.0  141.153.12.1    0     100       0    (65530) i
*> 199.10.10.0 141.153.29.2    0     100       0    (65530) 1 i
```

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	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 local-as feature to do a similar thing

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

BGP Scaling Techniques

Cisco.com

- **These 4 techniques should be core requirements in all ISP networks**

  Route Refresh

  Peer groups

  Route flap damping

  Route reflectors

BGP for Internet Service Providers

- BGP Basics (quick recap)
- Scaling BGP
- Deploying BGP in an ISP network
- Multihoming Examples
- Using Communities

Deploying BGP in an ISP Network

Current Practices

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

Cisco.com

- 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 versus OSPF/ISIS Configuration Example

Cisco.com

```
router bgp 34567
 neighbor core-ibgp peer-group
 neighbor core-ibgp remote-as 34567
 neighbor core-ibgp update-source Loopback0
 neighbor core-ibgp send-community
 neighbor core-ibgp-partial peer-group
 neighbor core-ibgp-partial remote-as 34567
 neighbor core-ibgp-partial update-source Loopback0
 neighbor core-ibgp-partial send-community
 neighbor core-ibgp-partial prefix-list network-ibgp out
 neighbor 222.1.9.10 peer-group core-ibgp
 neighbor 222.1.9.13 peer-group core-ibgp-partial
 neighbor 222.1.9.14 peer-group core-ibgp-partial
```

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**

Aggregation

Quality or Quantity?

Aggregation

Cisco.com

- ISPs receive address block from Regional Registry or upstream provider
- Aggregation means announcing the address block only, not subprefixes

  Subprefixes should only be announced in special cases – see later.

- Aggregate should be generated internally

  Not on the network borders!

Configuring Aggregation

- **ISP has 221.10.0.0/19 address block**
- **To put into BGP as an aggregate:**

```
router bgp 100
 network 221.10.0.0 mask 255.255.224.0
ip route 221.10.0.0 255.255.224.0 null0
```

- **The static route is a "pull up" route**

  more specific prefixes within this address block ensure connectivity to ISP's customers

  "longest match lookup"

Announcing Aggregate – Cisco IOS

- **Configuration Example**

```
router bgp 100
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.1 remote-as 101
 neighbor 222.222.10.1 prefix-list out-filter out
!
ip route 221.10.0.0 255.255.224.0 null0
!
ip prefix-list out-filter permit 221.10.0.0/19
```

Announcing an Aggregate

Cisco.com

- ISPs who don't and won't aggregate are held in poor regard by community
- Registries' minimum allocation size is now a /20

  no real reason to see subprefixes of allocated blocks in the Internet

  BUT there are currently >62000 /24s!

The Internet Today

Cisco.com

- **Current Internet Routing Table Statistics**

BGP Routing Table Entries	113395
Prefixes after maximum aggregation	73252
Unique prefixes in Internet	52213
Prefixes larger than registry alloc	46045
/24s announced	63389
only 5506 /24s are from 192.0.0.0/8	
ASes in use	13176

Receiving Prefixes

Receiving Prefixes: From Downstreams

Cisco.com

- **ISPs should only accept prefixes which have been assigned or allocated to their downstream customer**
- **For example**

  downstream has 220.50.0.0/20 block

  should only announce this to peers

  peers should only accept this from them

Receiving Prefixes: Cisco IOS

- **Configuration Example on upstream**

```
router bgp 100
 neighbor 222.222.10.1 remote-as 101
 neighbor 222.222.10.1 prefix-list customer in
!
ip prefix-list customer permit 220.50.0.0/20
```

Receiving Prefixes: From Upstreams

Cisco.com

- **Not desirable unless really necessary**

  special circumstances – see later

- **Ask upstream to either:**

  originate a default-route

  -or-

  announce one prefix you can use as default

Receiving Prefixes: From Upstreams

Cisco.com

- **Downstream Router Configuration**

```
router bgp 100
 network 221.10.0.0 mask 255.255.224.0
 neighbor 221.5.7.1 remote-as 101
 neighbor 221.5.7.1 prefix-list infilter in
 neighbor 221.5.7.1 prefix-list outfilter out
!
ip prefix-list infilter permit 0.0.0.0/0
!
ip prefix-list outfilter permit 221.10.0.0/19
```

Receiving Prefixes: From Upstreams

- Upstream Router Configuration

```
router bgp 101
 neighbor 221.5.7.2 remote-as 100
 neighbor 221.5.7.2 default-originate
 neighbor 221.5.7.2 prefix-list cust-in in
 neighbor 221.5.7.2 prefix-list cust-out out
!
ip prefix-list cust-in permit 221.10.0.0/19
!
ip prefix-list cust-out permit 0.0.0.0/0
```

Receiving Prefixes: From Peers and Upstreams

- **If necessary to receive prefixes from any provider, care is required**
  - **don’t accept RFC1918 etc prefixes**
    - **http://www.ietf.org/internet-drafts/draft-manning-dsua-08.txt**
  - **don’t accept your own prefix**
  - **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


```
router bgp 100
 network 221.10.0.0 mask 255.255.224.0
 neighbor 221.5.7.1 remote-as 101
 neighbor 221.5.7.1 prefix-list in-filter in
!
ip prefix-list in-filter deny 0.0.0.0/0           ! Block default
ip prefix-list in-filter deny 0.0.0.0/8 le 32
ip prefix-list in-filter deny 10.0.0.0/8 le 32
ip prefix-list in-filter deny 127.0.0.0/8 le 32
ip prefix-list in-filter deny 169.254.0.0/16 le 32
ip prefix-list in-filter deny 172.16.0.0/12 le 32
ip prefix-list in-filter deny 192.0.2.0/24 le 32
ip prefix-list in-filter deny 192.168.0.0/16 le 32
ip prefix-list in-filter deny 221.10.0.0/19 le 32 ! Block local prefix
ip prefix-list in-filter deny 224.0.0.0/3 le 32   ! Block multicast
ip prefix-list in-filter deny 0.0.0.0/0 ge 25     ! Block prefixes >/24
ip prefix-list in-filter permit 0.0.0.0/0 le 32
```

Cisco.com

Prefixes into iBGP

Injecting prefixes into iBGP

- **Use iBGP to carry customer prefixes**

  don't ever use IGP

- **Point static route to customer interface**
- **Use BGP network statement**
- **As long as static route exists (interface active), prefix will be in BGP**

Router Configuration network statement

Cisco.com

- **Example:**

```
interface loopback 0
 ip address 215.17.3.1 255.255.255.255
!
interface Serial 5/0
 ip unnumbered loopback 0
 ip verify unicast reverse-path
!
ip route 215.34.10.0 255.255.252.0 Serial 5/0
!
router bgp 100
 network 215.34.10.0 mask 255.255.252.0
```

Injecting prefixes into iBGP

- **interface flap will result in prefix withdraw and re-announce**

    use "ip route…permanent"

    Static route always exists, even if interface is down → prefix announced in iBGP

- **many ISPs use redistribute static rather than network statement**

    only use this if you understand why

Inserting prefixes into BGP: redistribute static

- **Care required with redistribute!**

  redistribute <routing-protocol> means everything in the <routing-protocol> will be transferred into the current routing protocol

  Does not scale if uncontrolled

  Best avoided if at all possible

  redistribute normally used with “route-maps” and under tight administrative control

Router Configuration: redistribute static

- **Example:**

```
ip route 215.34.10.0 255.255.252.0 Serial 5/0
!
router bgp 100
 redistribute static route-map static-to-bgp
<snip>
!
route-map static-to-bgp permit 10
 match ip address prefix-list ISP-block
 set origin igp
<snip>
!
ip prefix-list ISP-block permit 215.34.10.0/22 le 30
!
```

Injecting prefixes into iBGP

- Route-map ISP-block can be used for many things:
  - setting communities and other attributes
  - setting origin code to IGP, etc
- Be careful with prefix-lists and route-maps
  - absence of either/both could mean all statically routed prefixes go into iBGP

Configuration Tips

iBGP and IGPs

- **Make sure loopback is configured on router**
    - **iBGP between loopbacks, NOT real interfaces**
- **Make sure IGP carries loopback /32 address**
- **Make sure IGP carries DMZ nets**
    - **Use ip-unnumbered where possible**
    - **Or use next-hop-self on iBGP neighbours**
    - **neighbor x.x.x.x next-hop-self**

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 ip unnumbered

  Helps scale network

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

BGP Template – iBGP peers


```
router bgp 100
neighbor internal peer-group
neighbor internal description ibgp peers
neighbor internal remote-as 100
neighbor internal update-source Loopback0
neighbor internal next-hop-self
neighbor internal send-community
neighbor internal version 4
neighbor internal password 7 03085A09
neighbor 1.0.0.1 peer-group internal
neighbor 1.0.0.2 peer-group internal
```

BGP Template – iBGP peers

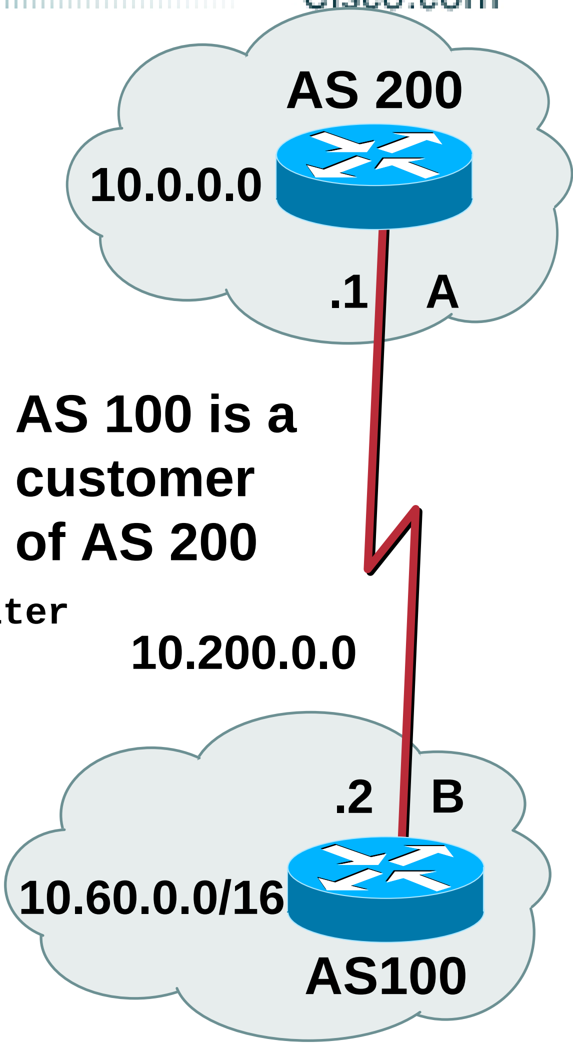
Cisco.com

- Use peer-groups
- iBGP between loopbacks!
- Next-hop-self
  - Keep DMZ and 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

BGP Template – eBGP peers


```
Router B:
router bgp 100
bgp dampening route-map RIPE229-flap
network 10.60.0.0 mask 255.255.0.0
neighbor external peer-group
neighbor external remote-as 200
neighbor external description ISP connection
neighbor external remove-private-AS
neighbor external version 4
neighbor external prefix-list ispout out ! “real” filter
neighbor external filter-list 1 out      ! “accident” filter
neighbor external route-map ispout out
neighbor external prefix-list ispin in
neighbor external filter-list 2 in
neighbor external route-map ispin in
neighbor external password 7 020A0559
neighbor external maximum-prefix 120000 [warning-only]
neighbor 10.200.0.1 peer-group external
!
ip route 10.60.0.0 255.255.0.0 null0 254
```

AS 100 is a customer of AS 200



BGP Template – eBGP peers

Cisco.com

- BGP damping – use RIPE-229 parameters
- Remove private ASes from announcements
  - Common omission today
- Use extensive filters, with "backup"
  - Use as-path filters to backup prefix-lists
  - Use route-maps for policy
- Use password agreed between you and peer on eBGP session
- Use maximum-prefix tracking
  - Router will warn you if there are sudden changes in BGP table size, bringing down eBGP if desired

More BGP “defaults”

Cisco.com

- Log neighbour changes

  bgp log-neighbor-changes

- Enable deterministic MED

  bgp deterministic-med

  Otherwise bestpath could be different every time BGP session is reset

- Make BGP admin distance higher than any IGP

  distance bgp 200 200 200

Customer Aggregation

- **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

BGP Customer Aggregation Guidelines

Cisco.com

- **Define at least three peer groups:**

  cust-default—send default route only

  cust-cust—send customer routes only

  cust-full —send full Internet routes

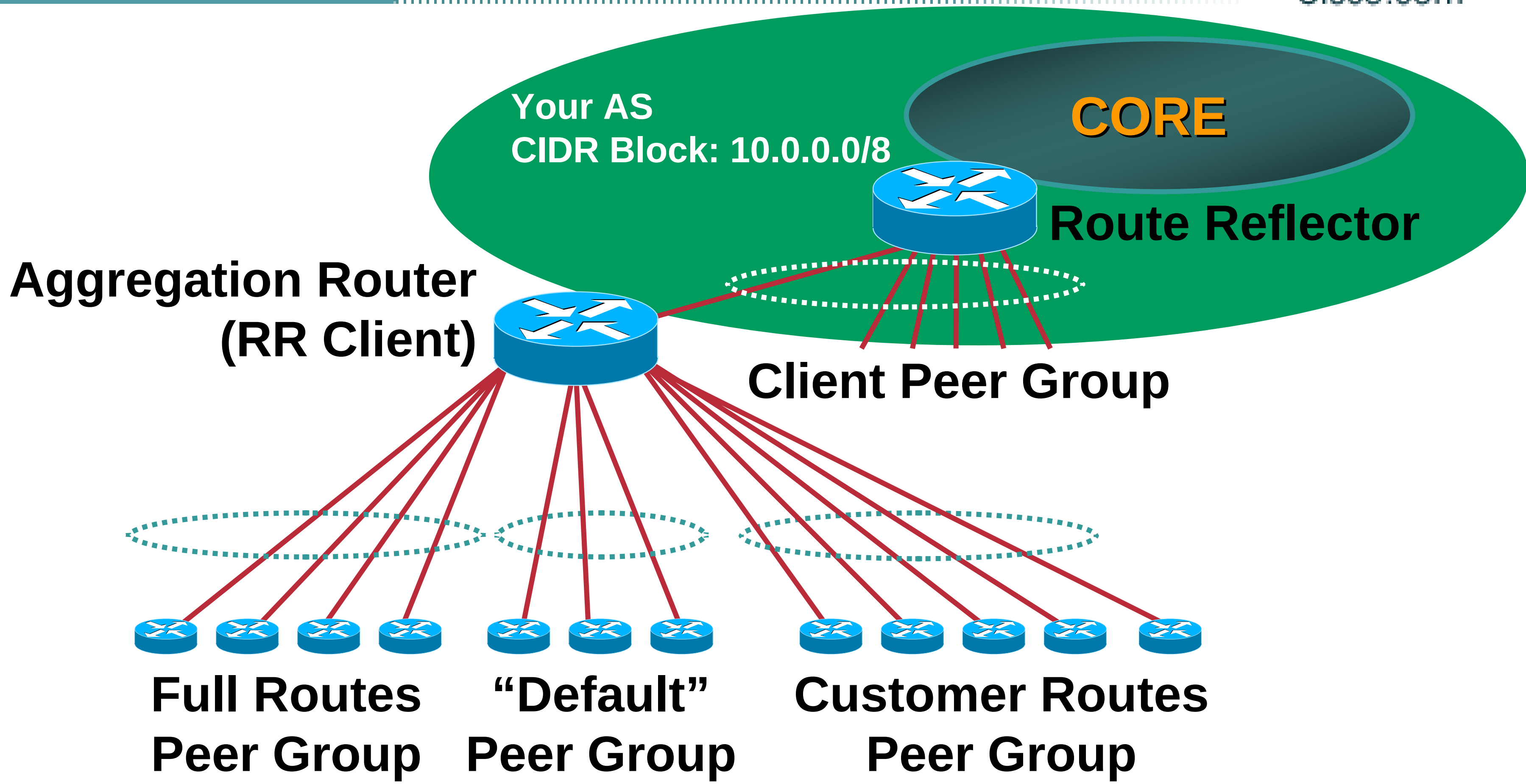
- **Identify routes via communities e.g.**

  100:4100=customers; 100:4500=peers

- **Apply passwords per neighbour**

- **Apply inbound & outbound prefix-list per neighbour**

BGP Customer Aggregation



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

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

- **BGP network statements on aggregation routers set correct community**

Sample core configuration

Cisco.com

- **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)

- **Simple configuration with peer-groups and route-maps**

Acquisitions!

- **Your ISP has just bought another ISP**

  How to merge networks?

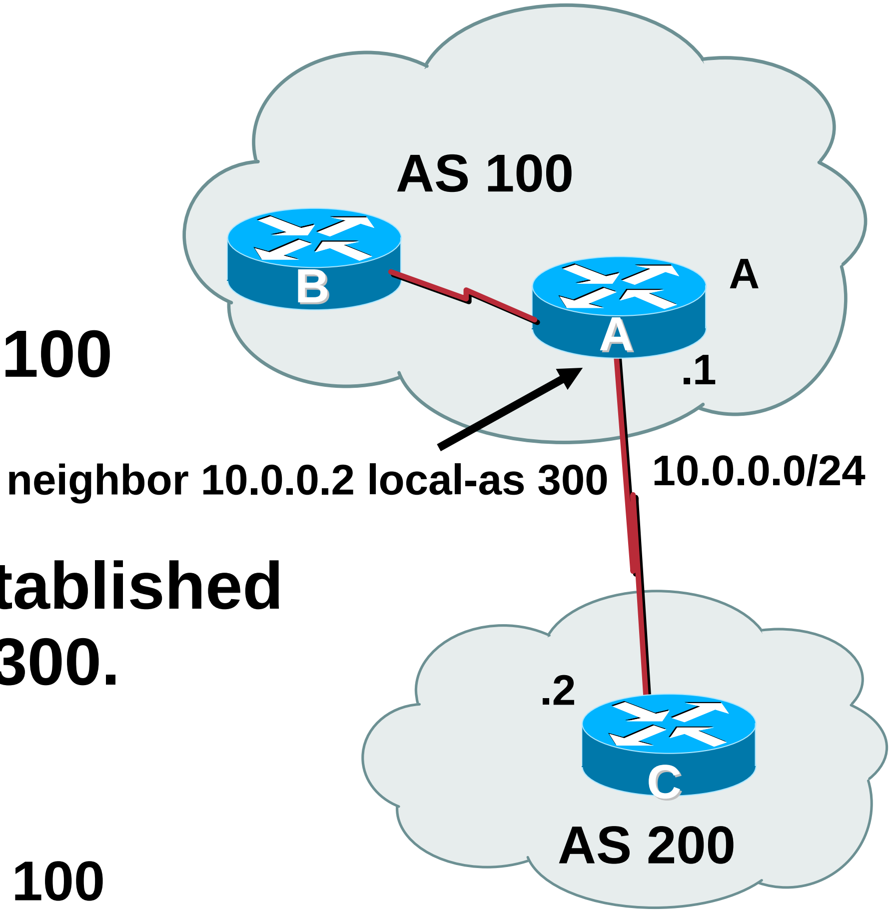
- **Options:**

  use confederations – make their AS a sub-AS (only useful if you are using confederations already)

  use the BGP local-as feature to implement a gradual transition – overrides BGP process ID

  neighbor *x.x.x.x* local-as *as-number*

local-AS – Application



- Router A has a process ID of 100
- The peering with AS200 is established as if router A belonged to AS300.
- AS_PATH

    routes originated in AS100 = 300 100

    routes received from AS200 = 300 200

BGP for Internet Service Providers

- BGP Basics (quick recap)
- Scaling BGP
- Deploying BGP in an ISP network
- Multihoming Examples
- Using Communities

Multihoming

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

Cisco.com

- **An Autonomous System Number is required by BGP**
- **Obtained from upstream ISP or Regional Registry**
- **Necessary when you have links to more than one ISP or exchange point**

Configuring Policy

- Three BASIC Principles

  prefix-lists to filter prefixes

  filter-lists to filter ASNs

  route-maps to apply policy

- Avoids confusion!

Originating Prefixes

Cisco.com

- **Basic Assumptions**

  MUST announce assigned address block to Internet

  MAY also announce subprefixes – reachability is not guaranteed

  RIR minimum allocation is /20

  several ISPs filter RIR blocks on this boundary

  called "Net Police" by some

Part of the “Net Police” prefix list


```
!! APNIC
ip prefix-list FILTER permit 61.0.0.0/8  ge 9 le 20
ip prefix-list FILTER permit 202.0.0.0/7 ge 9 le 20
ip prefix-list FILTER permit 210.0.0.0/7 ge 9 le 20
ip prefix-list FILTER permit 218.0.0.0/7 ge 9 le 20
ip prefix-list FILTER permit 220.0.0.0/8 ge 9 le 20
!! ARIN
ip prefix-list FILTER permit 24.0.0.0/8  ge 9 le 20
ip prefix-list FILTER permit 63.0.0.0/8  ge 9 le 20
ip prefix-list FILTER permit 64.0.0.0/6  ge 9 le 20
ip prefix-list FILTER permit 68.0.0.0/8  ge 9 le 20
ip prefix-list FILTER permit 199.0.0.0/8 ge 9 le 20
ip prefix-list FILTER permit 200.0.0.0/8 ge 9 le 20
ip prefix-list FILTER permit 204.0.0.0/6 ge 9 le 20
ip prefix-list FILTER permit 208.0.0.0/7 ge 9 le 20
ip prefix-list FILTER permit 216.0.0.0/8 ge 9 le 20
!! RIPE NCC
ip prefix-list FILTER permit 62.0.0.0/8  ge 9 le 20
ip prefix-list FILTER permit 80.0.0.0/7  ge 9 le 20
ip prefix-list FILTER permit 193.0.0.0/8 ge 9 le 20
ip prefix-list FILTER permit 194.0.0.0/7 ge 9 le 20
ip prefix-list FILTER permit 212.0.0.0/7 ge 9 le 20
ip prefix-list FILTER permit 217.0.0.0/8 ge 9 le 20
```

“Net Police” prefix list issues

Cisco.com

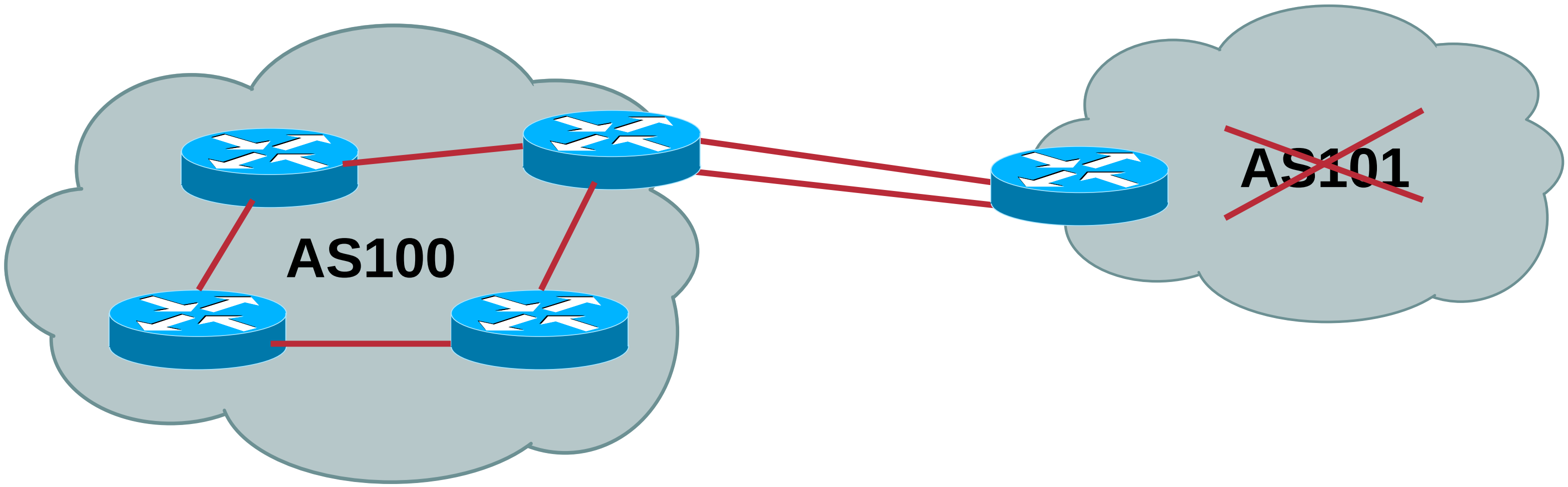
- meant to “punish” ISPs who won’t and don’t aggregate
- impacts legitimate multihoming
- impacts regions where domestic backbone is unavailable or costs $$$ compared with international bandwidth
- hard to maintain – requires updating when RIRs start allocating from new address blocks
- don’t do it unless consequences understood and you are prepared to keep it current

Multihoming Scenarios

When to use BGP, ASNs, and how

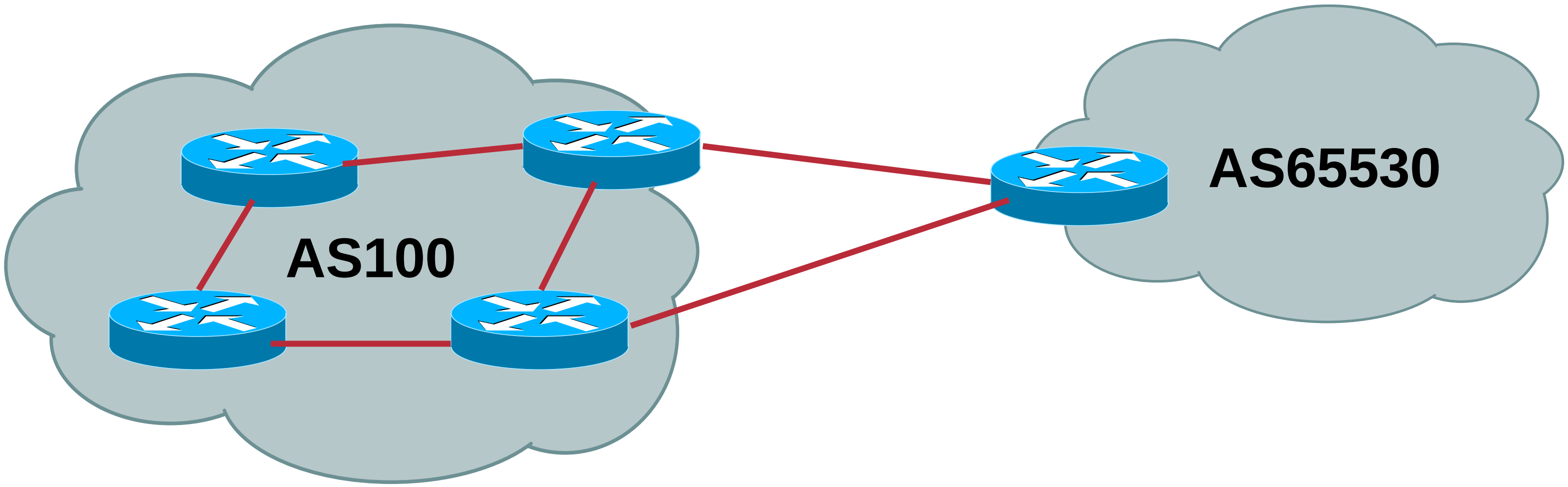
Stub Network

Cisco.com



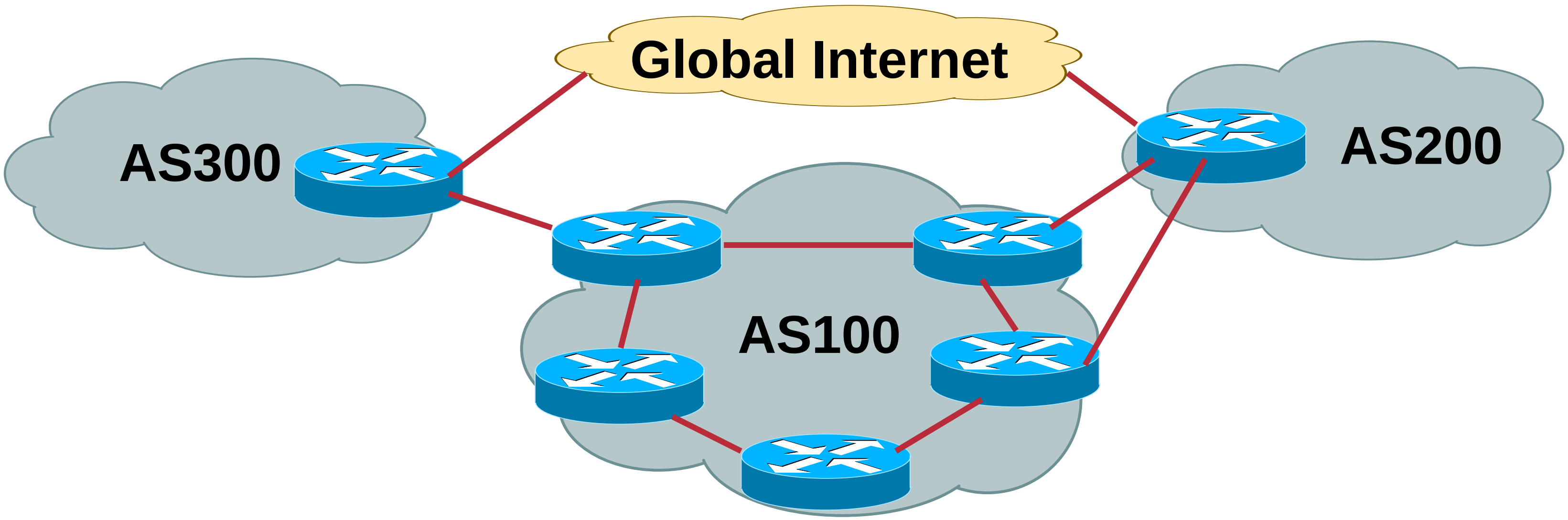
- No need for BGP
- Point static default to upstream ISP
- Upstream ISP advertises stub network
- Policy confined within upstream ISP's policy

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

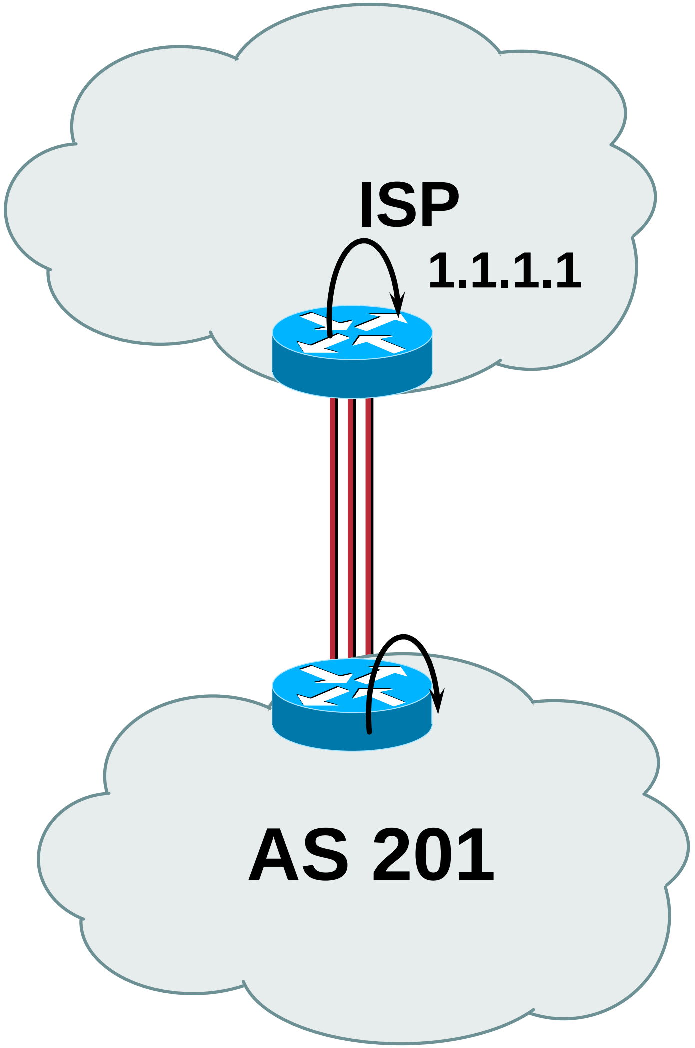


- **Many situations possible**
  - multiple sessions to same ISP
  - secondary for backup only
  - load-share between primary and secondary
  - selectively use different ISPs

Multiple Sessions to an ISP: Example One

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

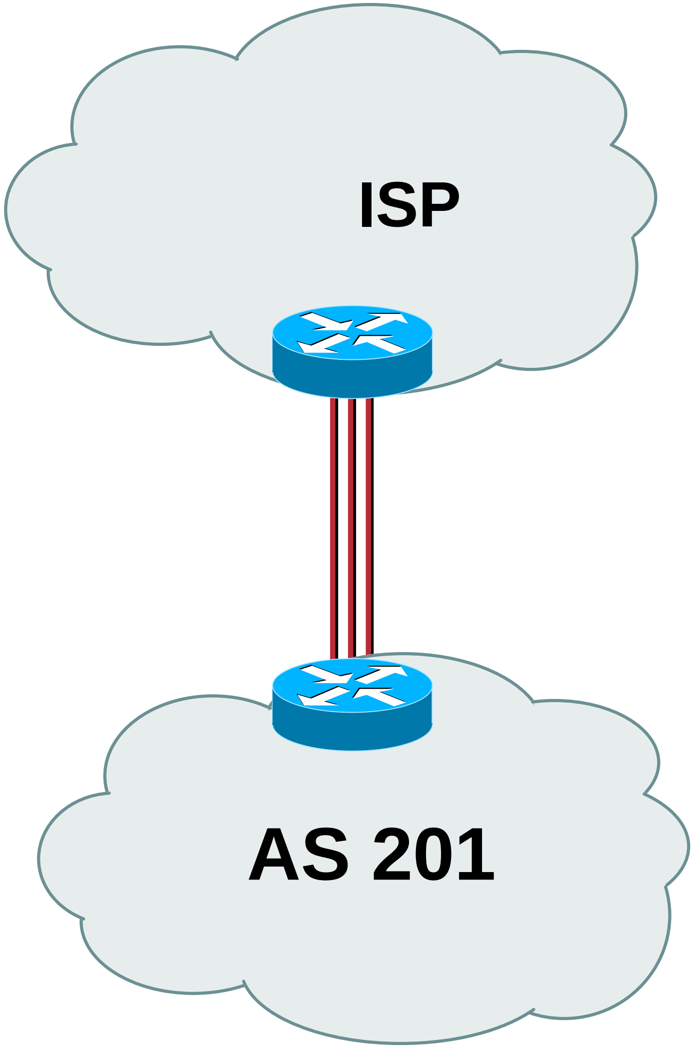
```
router bgp 201
 neighbor 1.1.1.1 remote-as 200
 neighbor 1.1.1.1 ebgp-multihop 3
ip route 1.1.1.1 255.255.255.255 serial 1/0
ip route 1.1.1.1 255.255.255.255 serial 1/1
ip route 1.1.1.1 255.255.255.255 serial 1/2
```



Multiple Sessions to an ISP: Example Two

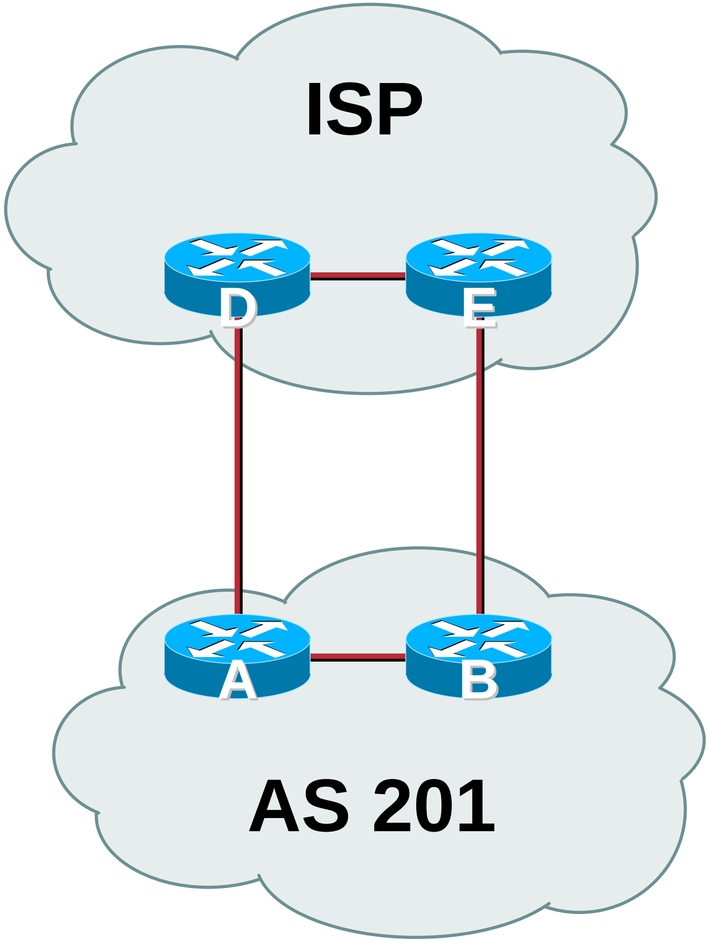
- BGP multi-path
- Three BGP sessions required
- limit of 6 parallel paths

```
router bgp 201
 neighbor 1.1.2.1 remote-as 200
 neighbor 1.1.2.5 remote-as 200
 neighbor 1.1.2.9 remote-as 200
 maximum-paths 3
```



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
- No magic solution

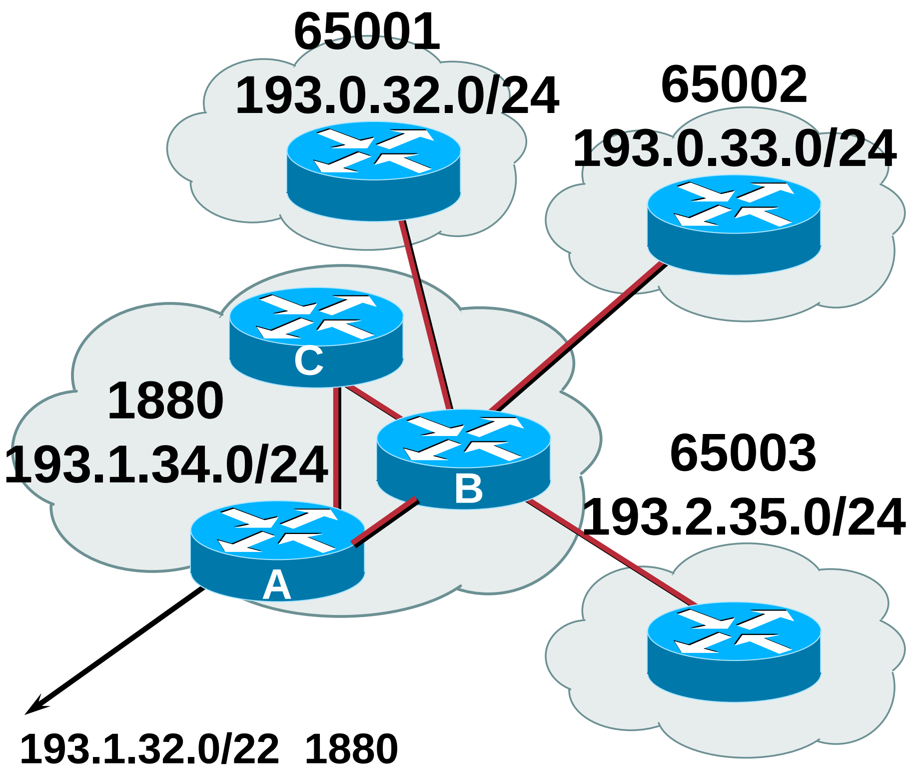


Private-AS – Application

- **Applications**

  ISP with single-homed customers (RFC2270)

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



Private-AS Removal

Cisco.com

- neighbor x.x.x.x remove-private-AS
- Please include in all eBGP configurations
- Rules:

    Available for eBGP neighbours only

    if the update has AS_PATH made up of private-AS numbers, the private-AS will be dropped

    if the AS_PATH includes private and public AS numbers, private AS number will not be removed…it is a configuration error!

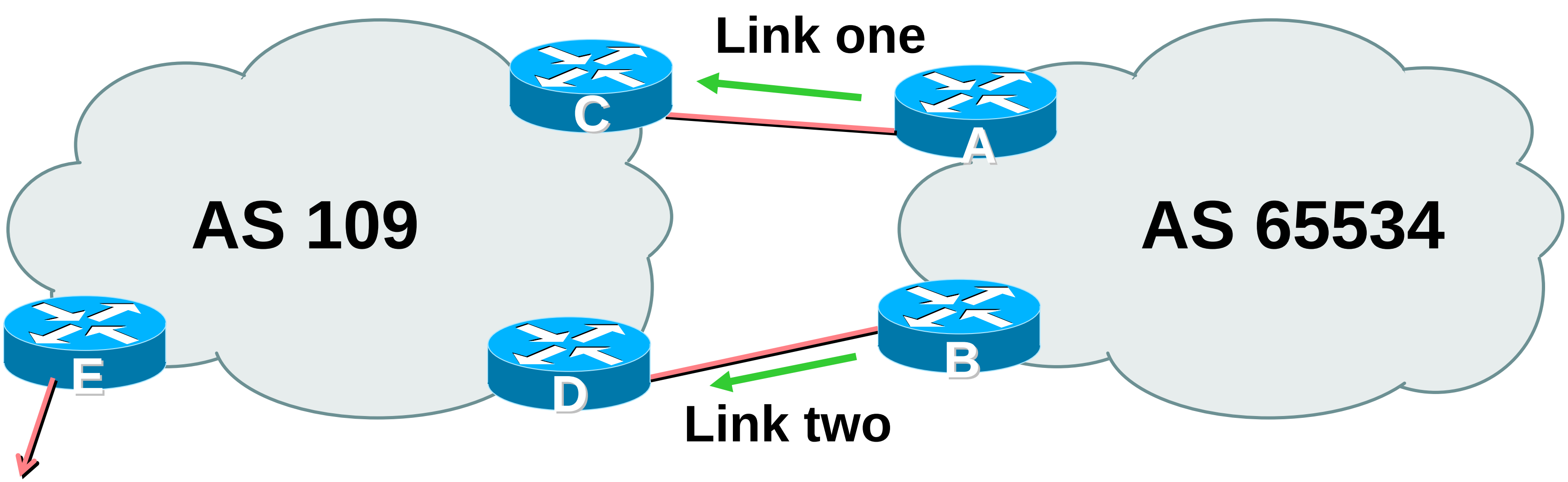
    if AS_PATH contains the AS number of the eBGP neighbor, the private-AS numbers will not be removed

    if used with confederations, it will work as long as the private AS numbers are after the confederation portion of the AS_PATH

Two links to the same ISP

Two links to the same ISP

Cisco.com



- **AS109 removes private AS and any customer subprefixes from Internet announcement**

Loadsharing to the same ISP

Cisco.com

- **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

Two links to the same ISP

Cisco.com

- **Router A Configuration**

```
router bgp 65534
 network 221.10.0.0 mask 255.255.224.0
 network 221.10.0.0 mask 255.255.240.0
 neighbor 222.222.10.2 remote-as 109
 neighbor 222.222.10.2 prefix-list routerC out
 neighbor 222.222.10.2 prefix-list default in
!
ip prefix-list default permit 0.0.0.0/0
ip prefix-list routerC permit 221.10.0.0/20
ip prefix-list routerC permit 221.10.0.0/19
!
ip route 221.10.0.0 255.255.240.0 null0
ip route 221.10.0.0 255.255.224.0 null0
```

Router B configuration is similar but with the other /20

Two links to the same ISP

- **Router C Configuration**

```
router bgp 109
 neighbor 222.222.10.1 remote-as 65534
 neighbor 222.222.10.1 default-originate
 neighbor 222.222.10.1 prefix-list Customer in
 neighbor 222.222.10.1 prefix-list default out
!
ip prefix-list Customer permit 221.10.0.0/19 le 20
ip prefix-list default permit 0.0.0.0/0
```

- **Router C only allows in /19 and /20 prefixes from customer block**
- **Router D configuration is identical**

Loadsharing to the same ISP

- **Loadsharing configuration is only on customer router**
- **Upstream ISP has to**

  remove customer subprefixes from external announcements

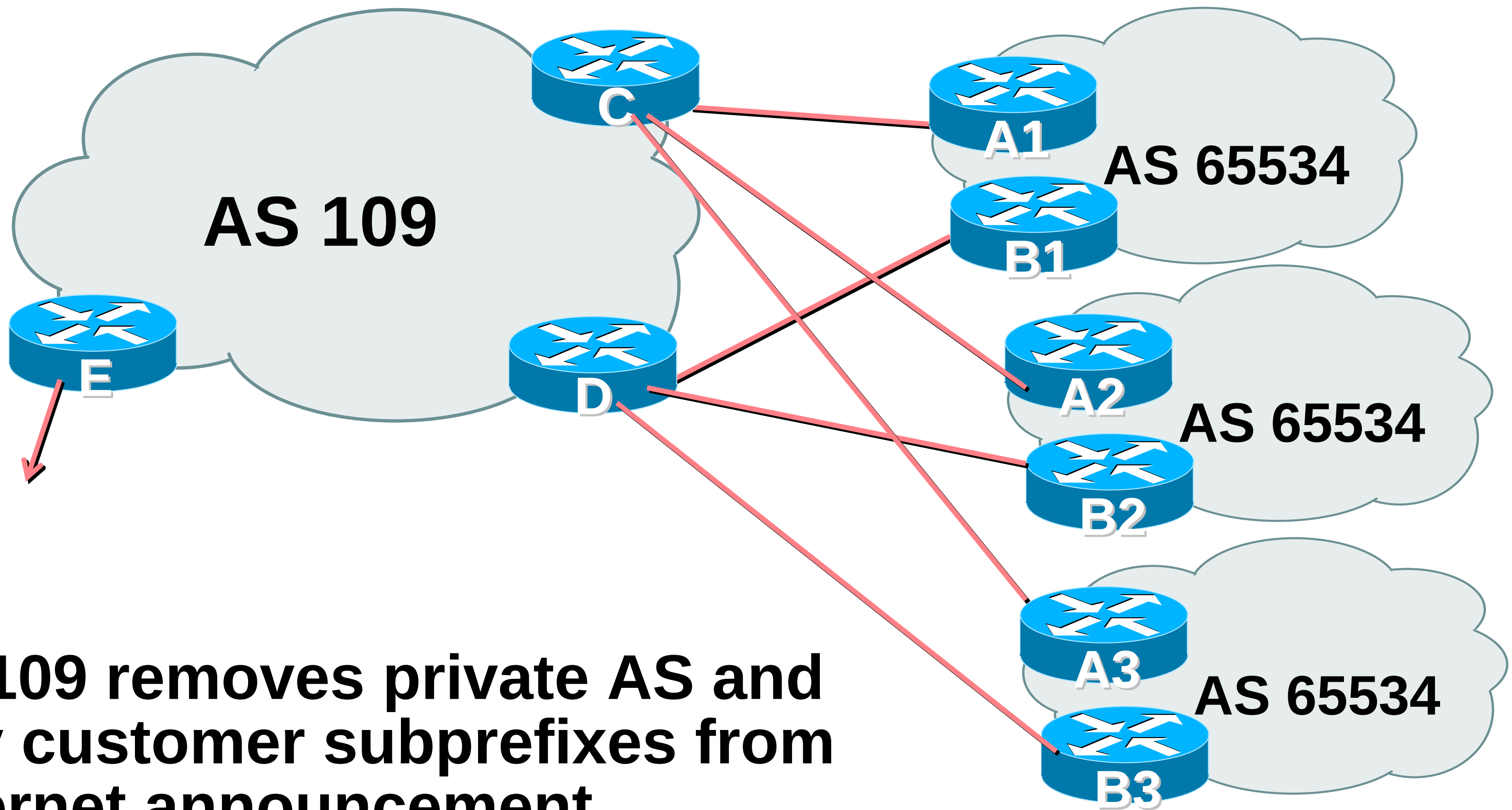
  remove private AS from external announcements
- **Could also use BGP communities**

Two links to the same ISP

Multiple Dualhomed Customers

(RFC2270)

Multiple Dualhomed Customers (RFC2270)



- **AS109 removes private AS and any customer subprefixes from Internet announcement**

Multiple Dualhomed Customers

- **Customer announcements as per previous example**
- **Use the *same* private AS for each customer**
  - **documented in RFC2270**
  - **address space is not overlapping**
  - **each customer hears default only**
- **Router A*n* and B*n* configuration same as Router A and B previously**

Two links to the same ISP

Cisco.com

- **Router A1 Configuration**

```
router bgp 65534
 network 221.10.0.0 mask 255.255.224.0
 network 221.10.0.0 mask 255.255.240.0
 neighbor 222.222.10.2 remote-as 109
 neighbor 222.222.10.2 prefix-list routerC out
 neighbor 222.222.10.2 prefix-list default in
!
ip prefix-list default permit 0.0.0.0/0
ip prefix-list routerC permit 221.10.0.0/20
ip prefix-list routerC permit 221.10.0.0/19
!
ip route 221.10.0.0 255.255.240.0 null0
ip route 221.10.0.0 255.255.224.0 null0
```

Router B1 configuration is similar but for the other /20

Multiple Dualhomed Customers

Cisco.com

- Router C Configuration

```
router bgp 109
 neighbor bgp-customers peer-group
 neighbor bgp-customers remote-as 65534
 neighbor bgp-customers default-originate
 neighbor bgp-customers prefix-list default out
 neighbor 222.222.10.1 peer-group bgp-customers
 neighbor 222.222.10.1 description Customer One
 neighbor 222.222.10.1 prefix-list Customer1 in
 neighbor 222.222.10.9 peer-group bgp-customers
 neighbor 222.222.10.9 description Customer Two
 neighbor 222.222.10.9 prefix-list Customer2 in
```

Multiple Dualhomed Customers

Cisco.com

```
 neighbor 222.222.10.17 peer-group bgp-customers
 neighbor 222.222.10.17 description Customer Three
 neighbor 222.222.10.17 prefix-list Customer3 in
!
ip prefix-list Customer1 permit 221.10.0.0/19 le 20
ip prefix-list Customer2 permit 221.16.64.0/19 le 20
ip prefix-list Customer3 permit 221.14.192.0/19 le 20
ip prefix-list default permit 0.0.0.0/0
```

- Router C only allows in /19 and /20 prefixes from customer block
- Router D configuration is almost identical

Multiple Dualhomed Customers

Cisco.com

- **Router E Configuration**

  assumes customer address space is not part of upstream's address block

```
router bgp 109
 neighbor 222.222.10.17 remote-as 110
 neighbor 222.222.10.17 remove-private-AS
 neighbor 222.222.10.17 prefix-list Customers out
!
ip prefix-list Customers permit 221.10.0.0/19
ip prefix-list Customers permit 221.16.64.0/19
ip prefix-list Customers permit 221.14.192.0/19
```

- **Private AS still visible inside AS109**

Multiple Dualhomed Customers

Cisco.com

- If customers' prefixes come from ISP's address block

  do NOT announce them to the Internet

  announce ISP aggregate only

- Router E configuration:

```
router bgp 109
 neighbor 222.222.10.17 remote-as 110
 neighbor 222.222.10.17 prefix-list my-aggregate out
!
ip prefix-list my-aggregate permit 221.8.0.0/13
```

Two links to different ISPs

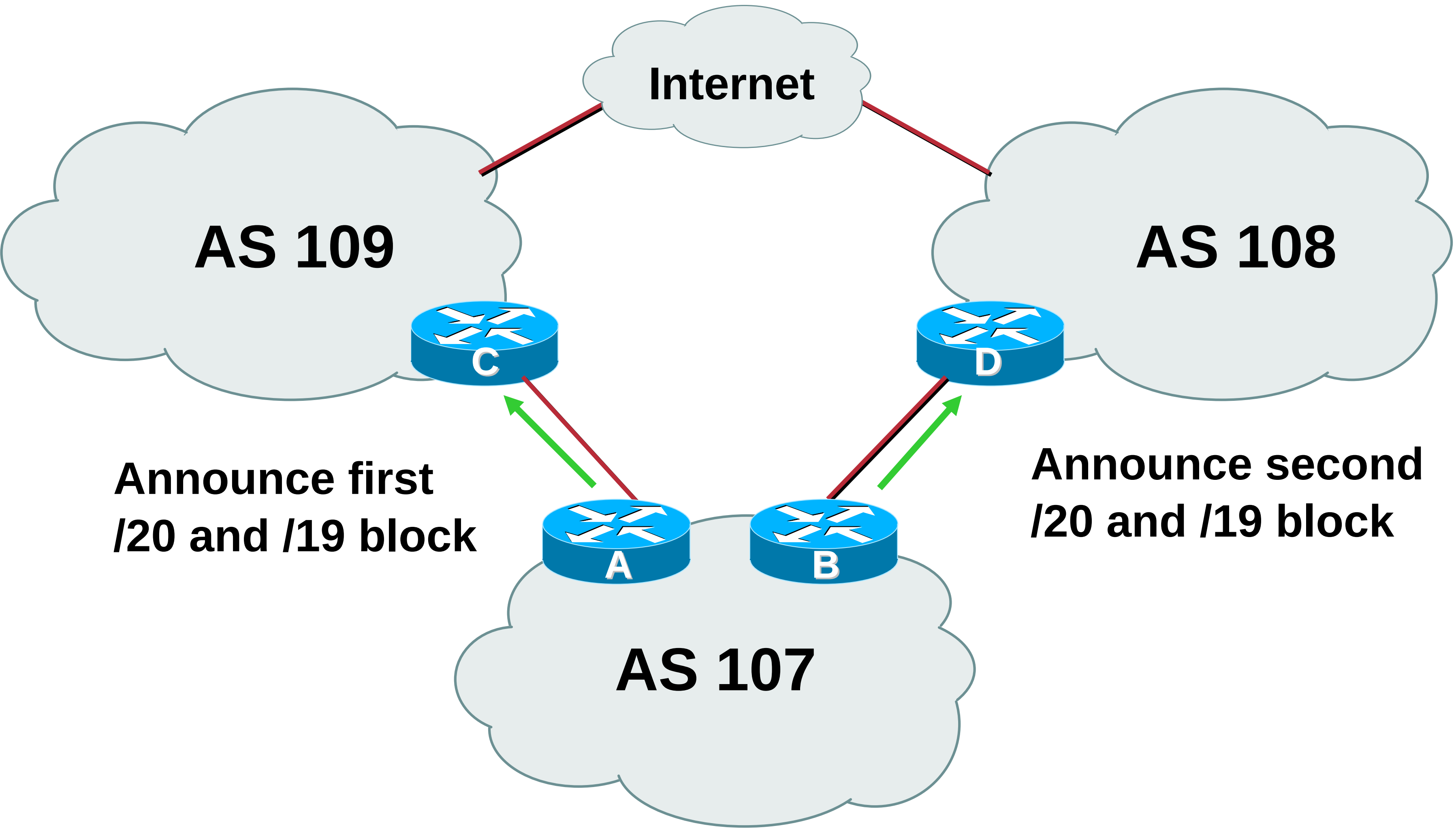
Two links to different ISPs

- **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



Two links to different ISPs

- Router A Configuration

```
router bgp 107
 network 221.10.0.0 mask 255.255.224.0
 network 221.10.0.0 mask 255.255.240.0
 neighbor 222.222.10.1 remote-as 109
 neighbor 222.222.10.1 prefix-list firstblock out
 neighbor 222.222.10.1 prefix-list default in
!
ip prefix-list default permit 0.0.0.0/0
!
ip prefix-list firstblock permit 221.10.0.0/20
ip prefix-list firstblock permit 221.10.0.0/19
```

Two links to different ISPs

- **Router B Configuration**

```
router bgp 107
 network 221.10.0.0 mask 255.255.224.0
 network 221.10.16.0 mask 255.255.240.0
 neighbor 220.1.5.1 remote-as 108
 neighbor 220.1.5.1 prefix-list secondblock out
 neighbor 220.1.5.1 prefix-list default in
!
ip prefix-list default permit 0.0.0.0/0
!
ip prefix-list secondblock permit 221.10.16.0/20
ip prefix-list secondblock permit 221.10.0.0/19
```

Two links to different ISPs

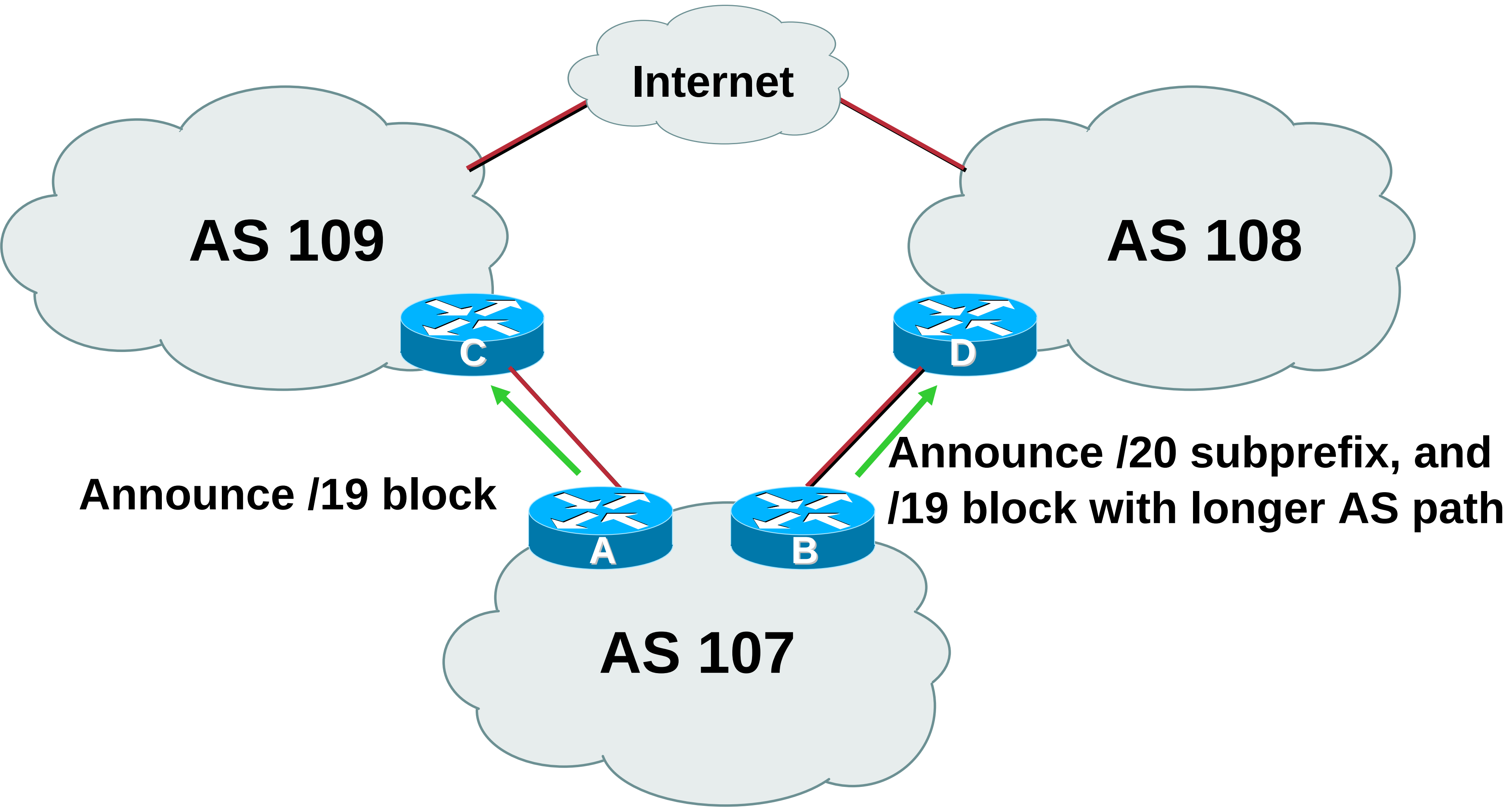
More Controlled Loadsharing

Loadsharing with different ISPs

Cisco.com

- **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



Loadsharing with different ISPs

- **Router A Configuration**

```
router bgp 107
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.1 remote-as 109
 neighbor 222.222.10.1 prefix-list default in
 neighbor 222.222.10.1 prefix-list aggregate out
!
ip prefix-list aggregate permit 221.10.0.0/19
```

Loadsharing with different ISPs

- **Router B Configuration**

```
router bgp 107
 network 221.10.0.0 mask 255.255.224.0
 network 221.10.16.0 mask 255.255.240.0
 neighbor 220.1.5.1 remote-as 108
 neighbor 220.1.5.1 prefix-list default in
 neighbor 220.1.5.1 prefix-list subblocks out
 neighbor 220.1.5.1 route-map routerD out
!
..next slide..
```

Loadsharing with different ISPs

```
route-map routerD permit 10
 match ip address prefix-list aggregate
 set as-path prepend 107 107
route-map routerD permit 20
!
ip prefix-list subblocks permit 221.10.0.0/19 le 20
ip prefix-list aggregate permit 221.10.0.0/19
```

Service Provider Multihoming

Service Provider Multihoming

- **Previous examples dealt with loadsharing inbound traffic**

  What about outbound?

- **ISPs strive to balance traffic flows in both directions**

  Balance link utilisation

  Try and keep most traffic flows symmetric

Service Provider Multihoming

- **Balancing outbound traffic requires inbound routing information**
  - Common solution is "full routing table"
  - Rarely necessary – don't need the "routing mallet" to try solve loadsharing problems
  - Keep It Simple (KISS) is often easier (and $$$ cheaper) than carrying n-copies of the full routing table

Service Provider Multihoming

Cisco.com

- Examples
  - One upstream, one local peer
  - One upstream, local exchange point
  - Two upstreams, one local peer
  - Tier-1 and regional upstreams, with local peers
  - IDC Multihoming
- All examples require BGP and a public ASN

Service Provider Multihoming

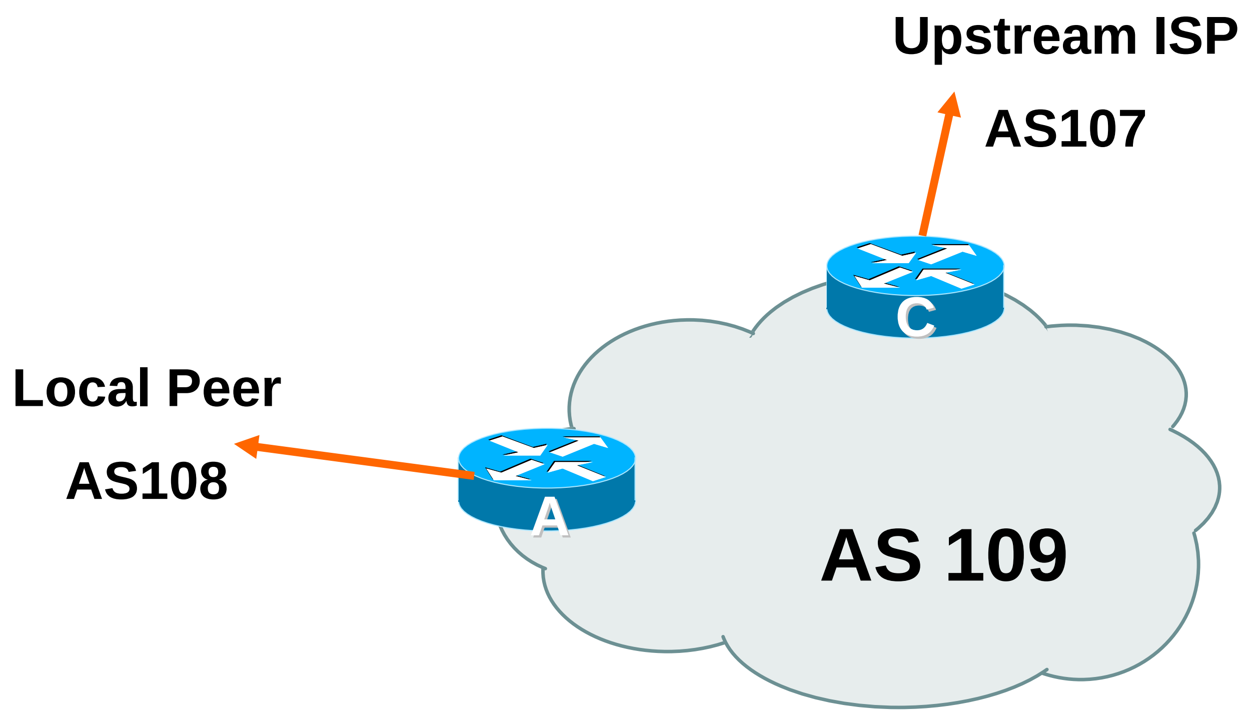
One Upstream, One local peer

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



One Upstream, One Local Peer

- Router A Configuration

```
router bgp 109
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.2 remote-as 108
 neighbor 222.222.10.2 prefix-list my-block out
 neighbor 222.222.10.2 prefix-list AS108-peer in
!
ip prefix-list AS108-peer permit 222.5.16.0/19
ip prefix-list AS108-peer permit 221.240.0.0/20
ip prefix-list my-block permit 221.10.0.0/19
!
ip route 221.10.0.0 255.255.224.0 null0
```

One Upstream, One Local Peer

- **Router A – Alternative Configuration**

```
router bgp 109
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.2 remote-as 108
 neighbor 222.222.10.2 prefix-list my-block out
 neighbor 222.222.10.2 filter-list 10 in
!
ip as-path access-list 10 permit ^(108_)+$
!
ip prefix-list my-block permit 221.10.0.0/19
!
ip route 221.10.0.0 255.255.224.0 null0
```

One Upstream, One Local Peer

- Router C Configuration

```
router bgp 109
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.1 remote-as 107
 neighbor 222.222.10.1 prefix-list default in
 neighbor 222.222.10.1 prefix-list my-block out
!
ip prefix-list my-block permit 221.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
ip route 221.10.0.0 255.255.224.0 null0
```

One Upstream, One Local Peer

- Two configurations possible for Router A
  - Filter-lists assume peer knows what they are doing
  - Prefix-list higher maintenance, but safer
  - Some ISPs do both – filter-list is “accident filter”
- Local traffic goes to and from local peer, everything else goes to upstream

Service Provider Multihoming

One Upstream, Local Exchange Point

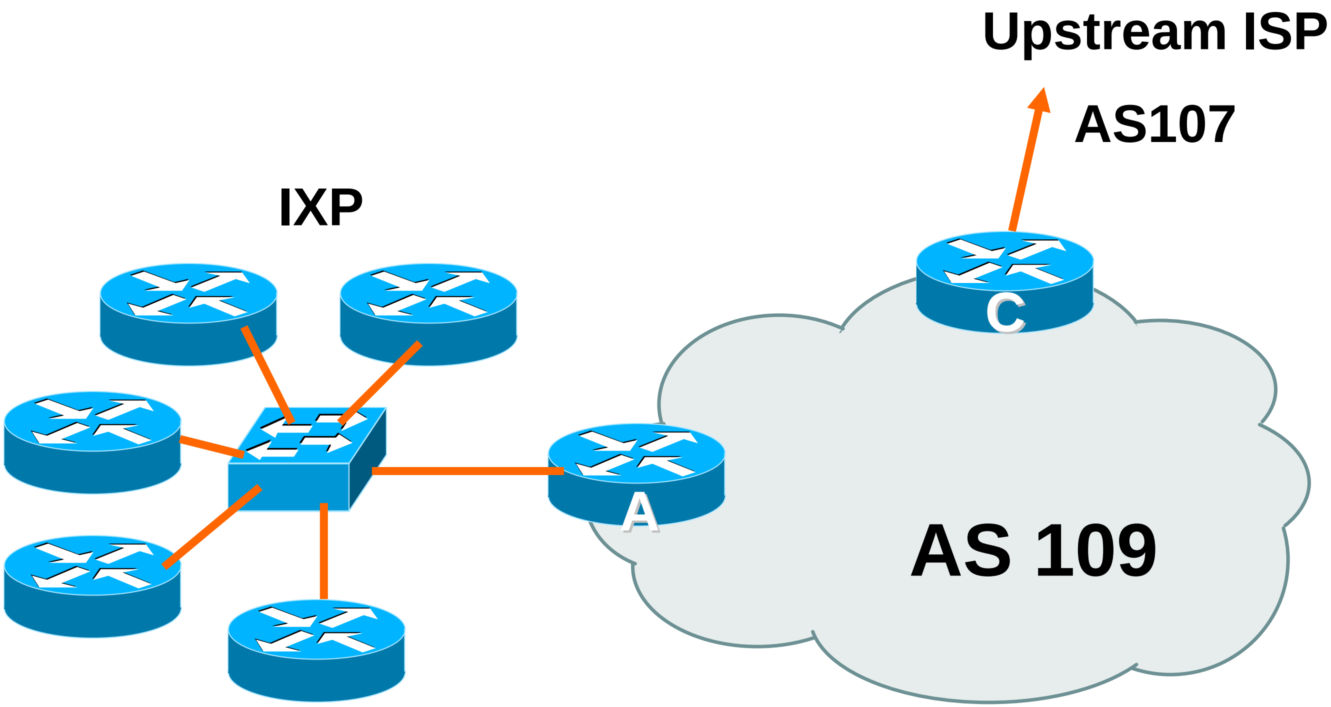
One Upstream, Local Exchange Point

Cisco.com

- **Announce /19 aggregate to every neighbouring AS**
- **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 IXP peers**

One Upstream, Local Exchange Point

Cisco.com



One Upstream, Local Exchange Point

Cisco.com

- **Router A Configuration**

```
interface fastethernet 0/0
 description Exchange Point LAN
 ip address 220.5.10.1 mask 255.255.255.224
 ip verify unicast reverse-path
 no ip directed-broadcast
 no ip proxy-arp
 no ip redirects
!
router bgp 109
 network 221.10.0.0 mask 255.255.224.0
 neighbor ixp-peers peer-group
 neighbor ixp-peers soft-reconfiguration in
 neighbor ixp-peers prefix-list my-block out
..next slide
```

One Upstream, Local Exchange Point

Cisco.com

```
neighbor 220.5.10.2 remote-as 100
neighbor 222.5.10.2 peer-group ixp-peers
neighbor 222.5.10.2 prefix-list peer100 in
neighbor 220.5.10.3 remote-as 101
neighbor 222.5.10.3 peer-group ixp-peers
neighbor 222.5.10.3 prefix-list peer101 in
neighbor 220.5.10.4 remote-as 102
neighbor 222.5.10.4 peer-group ixp-peers
neighbor 222.5.10.4 prefix-list peer102 in
neighbor 220.5.10.5 remote-as 103
neighbor 222.5.10.5 peer-group ixp-peers
neighbor 222.5.10.5 prefix-list peer103 in
```

..next slide

One Upstream, Local Exchange Point

Cisco.com

```
ip route 221.10.0.0 255.255.224.0 null0
!
ip route 221.10.0.0 255.255.224.0 null0
!
ip prefix-list my-block permit 221.10.0.0/19
ip prefix-list peer100 permit 222.0.0.0/19
ip prefix-list peer101 permit 222.30.0.0/19
ip prefix-list peer102 permit 222.12.0.0/19
ip prefix-list peer103 permit 222.18.128.0/19
!
```

One Upstream, Local Exchange Point

- Router C Configuration

```
router bgp 109
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.1 remote-as 107
 neighbor 222.222.10.1 prefix-list default in
 neighbor 222.222.10.1 prefix-list my-block out
!
ip prefix-list my-block permit 221.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
ip route 221.10.0.0 255.255.224.0 null0
```

One Upstream, Local Exchange Point

- **Note Router A configuration**
  - **Prefix-list higher maintenance, but safer**
  - **uRPF on the FastEthernet interface**
- **IXP traffic goes to and from local IXP, everything else goes to upstream**

Service Provider Multihoming

Two Upstreams, One local peer

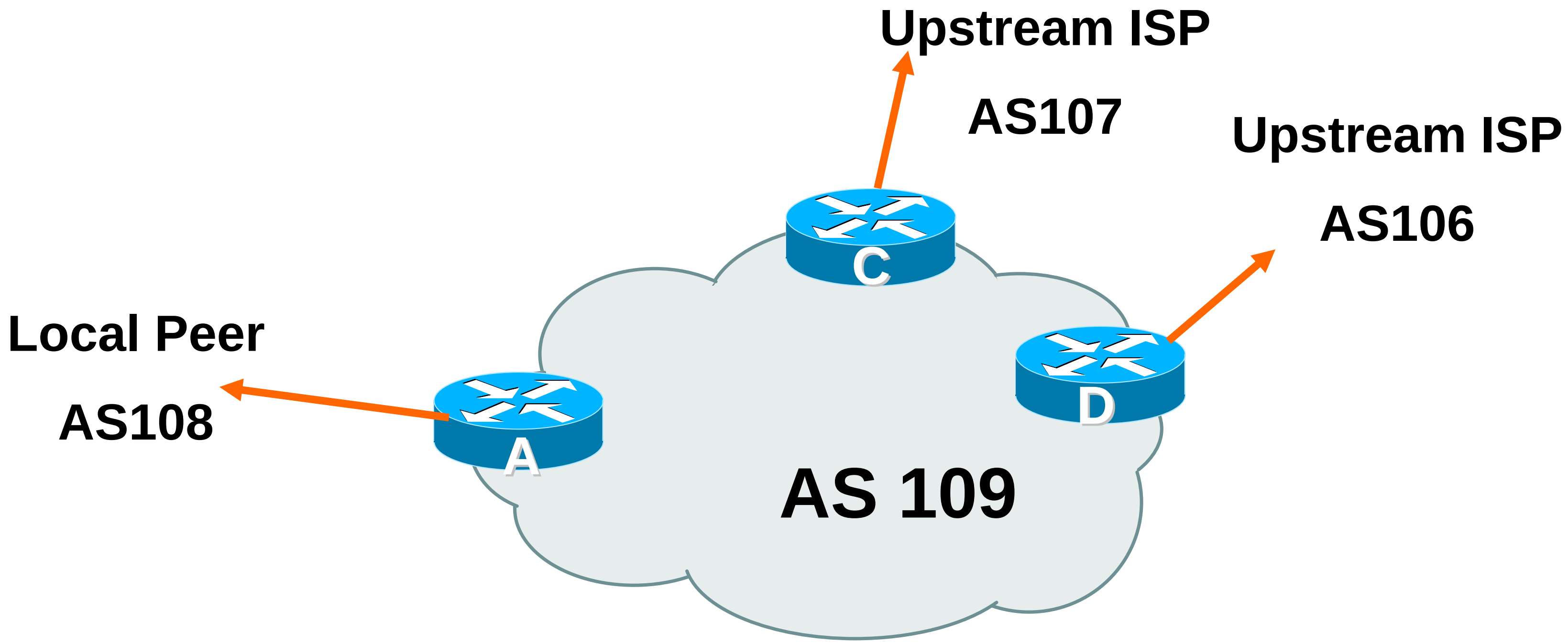
Two Upstreams, One Local Peer

- **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



Two Upstreams, One Local Peer

- **Router A**

  Same routing configuration as in example with one upstream and one local peer

  Same hardware configuration

Two Upstreams, One Local Peer

- Router C Configuration

```
router bgp 109
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.1 remote-as 107
 neighbor 222.222.10.1 prefix-list default in
 neighbor 222.222.10.1 prefix-list my-block out
!
ip prefix-list my-block permit 221.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
ip route 221.10.0.0 255.255.224.0 null0
```

Two Upstreams, One Local Peer

Cisco.com

- Router D Configuration

```
router bgp 109
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.5 remote-as 106
 neighbor 222.222.10.5 prefix-list default in
 neighbor 222.222.10.5 prefix-list my-block out
!
ip prefix-list my-block permit 221.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
ip route 221.10.0.0 255.255.224.0 null0
```

Two Upstreams, One Local Peer

- **This is the simple configuration for Router C and D**
- **Traffic out to the two upstreams will take nearest exit**

  Inexpensive routers required

  This is not useful in practice especially for international links

  Loadsharing needs to be better

Two Upstreams, One Local Peer

Cisco.com

- **Better configuration options:**
  - **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

```
router bgp 109
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.1 remote-as 107
 neighbor 222.222.10.1 prefix-list rfc1918-deny in
 neighbor 222.222.10.1 prefix-list my-block out
 neighbor 222.222.10.1 route-map AS107-loadshare in
!
ip prefix-list my-block permit 221.10.0.0/19
! See earlier in presentation for RFC1918 list
..next slide
```

Two Upstreams, One Local Peer: Full Routes


```
ip route 221.10.0.0 255.255.224.0 null0
!
ip as-path access-list 10 permit ^(107_)+$
ip as-path access-list 10 permit ^(107_)+_[0-9]+$
!
route-map AS107-loadshare permit 10
 match ip as-path 10
 set local-preference 120
route-map AS107-loadshare permit 20
 set local-preference 80
!
```

Two Upstreams, One Local Peer: Full Routes

- **Router D Configuration**

```
router bgp 109
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.5 remote-as 106
 neighbor 222.222.10.5 prefix-list rfc1918-deny in
 neighbor 222.222.10.5 prefix-list my-block out
!
ip prefix-list my-block permit 221.10.0.0/19
! See earlier in presentation for RFC1918 list
```

Two Upstreams, One Local Peer: Full Routes

Cisco.com

- Router C configuration:
  - Accept full routes from AS107
  - Tag prefixes originated by AS107 and AS107's neighbouring ASes with local preference 120
    - Traffic to those ASes will go over AS107 link
  - Remaining prefixes tagged with local preference of 80
    - Traffic to other all other ASes will go over the link to AS106
- Router D configuration same as Router C without the route-map

Two Upstreams, One Local Peer: Full Routes

- Full routes from upstreams

  Needs lots of memory

  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 presentation for examples

Two Upstreams, One Local Peer: Partial Routes

- Router C Configuration

```
router bgp 109
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.1 remote-as 107
 neighbor 222.222.10.1 prefix-list rfc1918-nodef-deny in
 neighbor 222.222.10.1 prefix-list my-block out
 neighbor 222.222.10.1 filter-list 10 in
 neighbor 222.222.10.1 route-map tag-default-low in
!
ip prefix-list my-block permit 221.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
..next slide
```

Two Upstreams, One Local Peer: Partial Routes

Cisco.com

```
! See earlier presentation for RFC1918 list
!
ip route 221.10.0.0 255.255.224.0 null0
!
ip as-path access-list 10 permit ^(107_)+$
ip as-path access-list 10 permit ^(107_)+_[0-9]+$
!
route-map tag-default-low permit 10
 match ip address prefix-list default
 set local-preference 80
route-map tag-default-low permit 20
!
```

Two Upstreams, One Local Peer: Partial Routes

- **Router D Configuration**

```
router bgp 109
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.5 remote-as 106
 neighbor 222.222.10.5 prefix-list default in
 neighbor 222.222.10.5 prefix-list my-block out
!
ip prefix-list my-block permit 221.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
ip route 221.10.0.0 255.255.224.0 null0
```

Two Upstreams, One Local Peer: Partial Routes

- Router C configuration:
    - Accept full routes from AS107
        - (or get them to send less)
    - Filter ASNs so only AS107 and AS107's neighbouring ASes are accepted
    - Allow default, and set it to local preference 80
    - Traffic to those ASes will go over AS107 link
    - Traffic to other all other ASes will go over the link to AS106
    - If AS106 link fails, backup via AS107 – and vice-versa

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 presentation for examples

Two Upstreams, One Local Peer: Partial Routes

- **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**

Two Upstreams, One Local Peer: Partial Routes

Cisco.com

- **Router C Configuration**

```
router ospf 109
 default-information originate metric 30
 passive-interface Serial 0/0
!
router bgp 109
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.1 remote-as 107
 neighbor 222.222.10.1 prefix-list rfc1918-deny in
 neighbor 222.222.10.1 prefix-list my-block out
 neighbor 222.222.10.1 filter-list 10 in
!
..next slide
```

Two Upstreams, One Local Peer: Partial Routes


```
ip prefix-list my-block permit 221.10.0.0/19
! See earlier presentation for RFC1918 list
!
ip route 221.10.0.0 255.255.224.0 null0
ip route 0.0.0.0 0.0.0.0 serial 0/0 254
!
ip as-path access-list 10 permit ^(107_)+$
ip as-path access-list 10 permit ^(107_)+_[0-9]+$
!
```

Two Upstreams, One Local Peer: Partial Routes

- **Router D Configuration**

```
router ospf 109
 default-information originate metric 10
 passive-interface Serial 0/0
!
router bgp 109
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.5 remote-as 106
 neighbor 222.222.10.5 prefix-list deny-all in
 neighbor 222.222.10.5 prefix-list my-block out
!
..next slide
```

Two Upstreams, One Local Peer: Partial Routes


```
ip prefix-list deny-all deny 0.0.0.0/0 le 32
ip prefix-list my-block permit 221.10.0.0/19
! See earlier presentation for RFC1918 list
!
ip route 221.10.0.0 255.255.224.0 null0
ip route 0.0.0.0 0.0.0.0 serial 0/0 254
!
```

Two Upstreams, One Local Peer: Partial Routes

- **Partial routes from upstreams**

  Use OSPF to determine outbound path

  Router D default has metric 10 – primary outbound path

  Router C default has metric 30 – backup outbound path

  Serial interface goes down, static default is removed from routing table, OSPF default withdrawn

Service Provider Multihoming

Two Tier-1 upstreams, two regional upstreams, and local peers

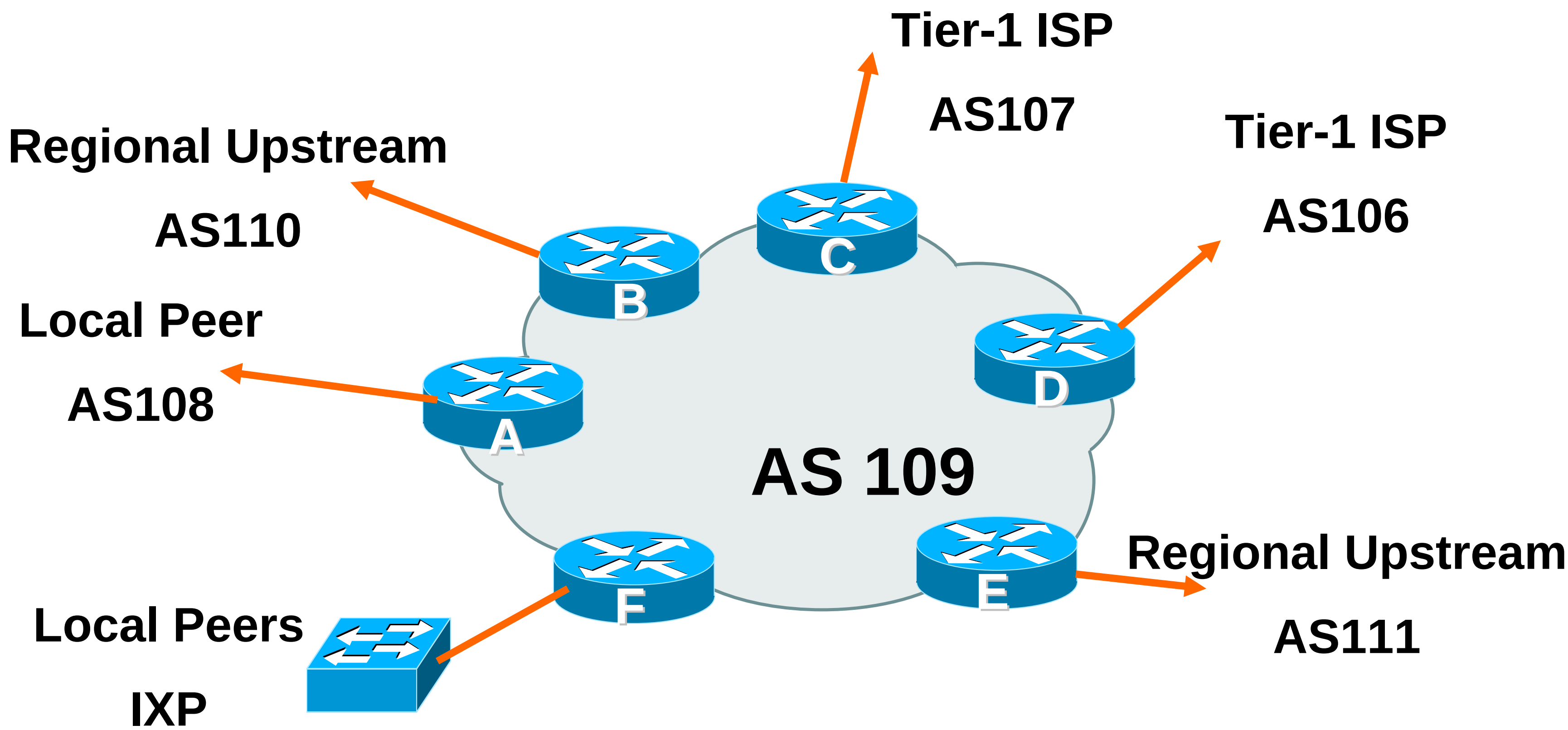
Tier-1 and Regional Upstreams, Local Peers

Cisco.com

- **Announce /19 aggregate on each link**
- **Accept partial/default routes from upstreams**

  For default, use 0.0.0.0/0 or a network which can be used as default
- **Accept all routes from local peer**
- **Accept all partial routes from regional upstreams**
- **This is more complex, but a very typical scenario**

Tier-1 and Regional Upstreams, Local Peers



Tier-1 and Regional Upstreams, Local Peers – Detail

- Router A – local private peer

  Accept all (local) routes

  Local traffic stays local

  Use prefix and/or AS-path filters

  Set >100 local preference on inbound announcements

Tier-1 and Regional Upstreams, Local Peers – Detail

- **Router F – local IXP peering**
  - Accept all (local) routes
  - Local traffic stays local
  - Use prefix and/or AS-path filters
  - Set >100 local preference on inbound announcements

Tier-1 and Regional Upstreams, Local Peers – Detail

- **Router B – regional upstream**

  They provide transit to Internet, but longer AS path than Tier-1 Upstreams

  Accept all regional routes from them

  e.g. ^110_[0-9]+$

  Ask them to send default, or send a network you can use as default

  Set local pref on “default” to 60

  Will provide backup to Internet only when direct Tier-1 links go down

Tier-1 and Regional Upstreams, Local Peers – Detail

- **Router E – regional upstream**

  They provide transit to Internet, but longer AS path than Tier-1 Upstreams

  Accept all regional routes from them

  e.g. ^111_[0-9]+$

  Ask them to send default, or send a network you can use as default

  Set local pref on “default” to 70

  Will provide backup to Internet only when direct Tier-1 links go down

Tier-1 and Regional Upstreams, Local Peers – Detail

Cisco.com

- Router C – first Tier-1 upstream

  Accept all their customer and AS neighbour routes from them

    e.g. ^107_[0-9]+$

  Ask them to send default, or send a network you can use as default

    Set local pref on “default” to 80

  Will provide backup to Internet only when link to second Tier-1 upstream goes down

Tier-1 and Regional Upstreams, Local Peers – Detail

Cisco.com

- Router D – second Tier-1 upstream

  Ask them to send default, or send a network you can use as default

    This has local preference 100 by default

  All traffic without any more specific path will go out this way

Tier-1 and Regional Upstreams, Local Peers – Summary

- Local traffic goes to local peer and IXP
- Regional traffic goes to two regional upstreams
- Everything else is shared between the two Tier-1 upstreams
- To modify loadsharing tweak what is heard from the two regionals and the first Tier-1 upstream

  Best way is through modifying the AS-path filter

Tier-1 and Regional Upstreams, Local Peers

Cisco.com

- **What about outbound announcement strategy?**

  This is to determine incoming traffic flows

  /19 aggregate must be announced to everyone!

  /20 or /21 more specifics can be used to improve or modify loadsharing

  See earlier for hints and ideas

Tier-1 and Regional Upstreams, Local Peers

- **What about unequal circuit capacity?**

  AS-path filters are very useful

- **What if upstream will only give me full routing table or nothing**

  AS-path and prefix filters are very useful

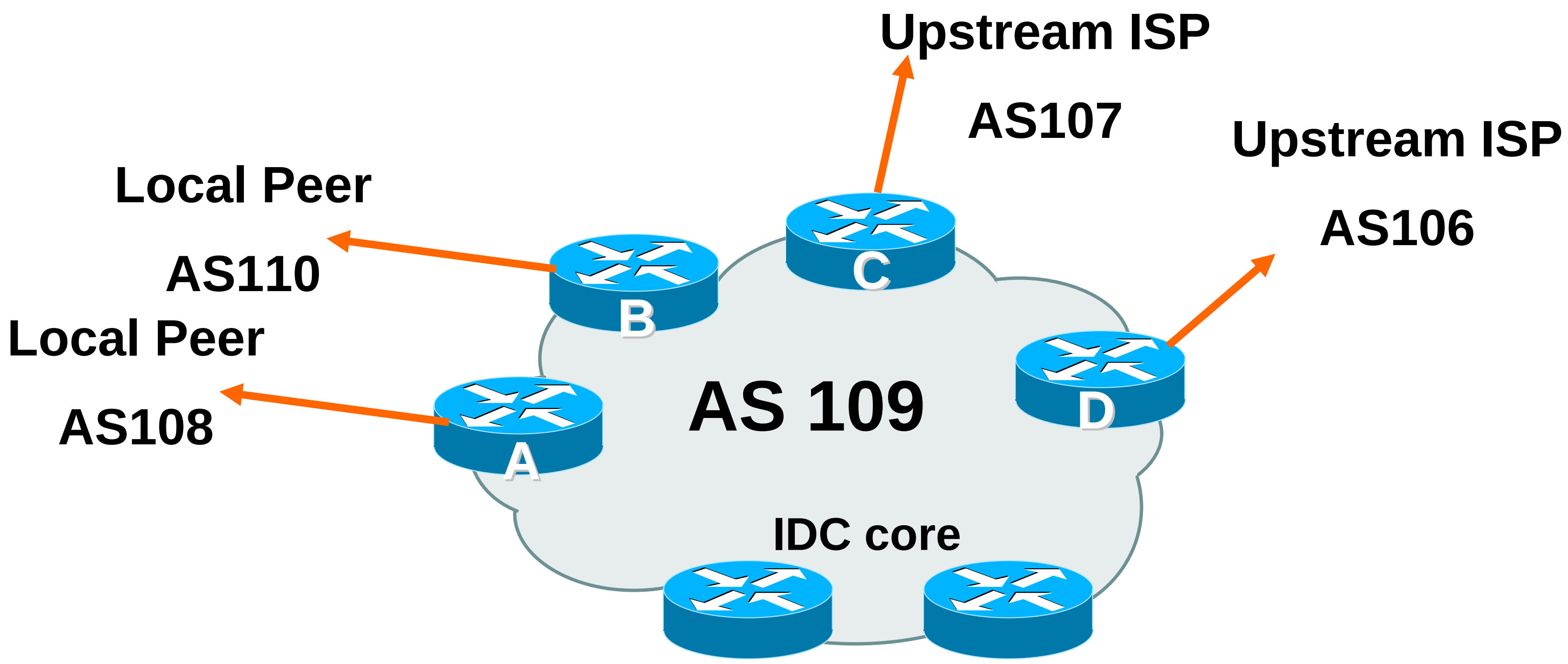
IDC Multihoming

IDC Multihoming

Cisco.com

- IDCs typically are not registry members so don't get their own address block
  - Situation also true for small ISPs and "Enterprise Networks"
- Smaller address blocks being announced
  - Address space comes from both upstreams
  - Should be apportioned according to size of circuit to upstream
- Outbound traffic paths matter

Two Upstreams, Two Local Peers: IDC



Assigned /24 from AS107 and /23 from AS106.

Circuit to AS107 is 2Mbps, circuit to AS106 is 4Mbps

IDC Multihoming

- **Router A and B configuration**

  In: Should accept all routes from AS108 and AS110

  Out: Should announce all address space to AS108 and AS110

  Straightforward

IDC Multihoming

- **Router C configuration**

  In: Accept partial routes from AS107

  e.g. ^107_[0-9]+$

  In: Ask for a route to use as default

  set local preference on default to 80

  Out: Send /24, and send /23 with AS-PATH prepend of one AS

IDC Multihoming

- **Router D configuration**

  In: Ask for a route to use as default

  Leave local preference of default at 100

  Out: Send /23, and send /24 with AS-PATH prepend of one AS

IDC Multihoming
Fine Tuning

- **For local fine tuning, increase circuit capacity**
  - **Local circuits usually are cheap**
  - **Otherwise…**
- **For longer distance fine tuning**
  - **In: Modify as-path filter on Router C**
  - **Out: Modify as-path prepend on Routers C and D**
  - **Outbound traffic flow is usual critical for an IDC so inbound policies need to be carefully thought out**

IDC Multihoming
Other Details

- **Redundancy**
  - **Circuits are terminated on separate routers**
- **Apply thought to address space use**
  - **Request from both upstreams**
  - **Utilise address space evenly across IDC**
    - **Don’t start with /23 then move to /24 – use both blocks at the same time in the same proportion**
    - **Helps with loadsharing – yes, really!**

IDC Multihoming
Other Details

Cisco.com

- **What about failover?**
  - **/24 and /23 from upstreams' blocks announced to the Internet routing table all the time**
  - **No obvious alternative at the moment**
    - **Conditional advertisement can help in steady state, but subprefixes still need to be announced in failover condition**

Service Provider Multihoming

Case Study

Case Study Requirements (1)

- **ISP needs to multihome:**
  - **To AS5400 in Europe**
  - **To AS2516 in Japan**
  - **/19 allocated by APNIC**
  - **AS 17660 assigned by APNIC**
  - **1Mbps circuits to both upstreams**

Case Study Requirements (2)

Cisco.com

- **ISP wants:**

  Symmetric routing and equal link utilisation in and out (as close as possible)

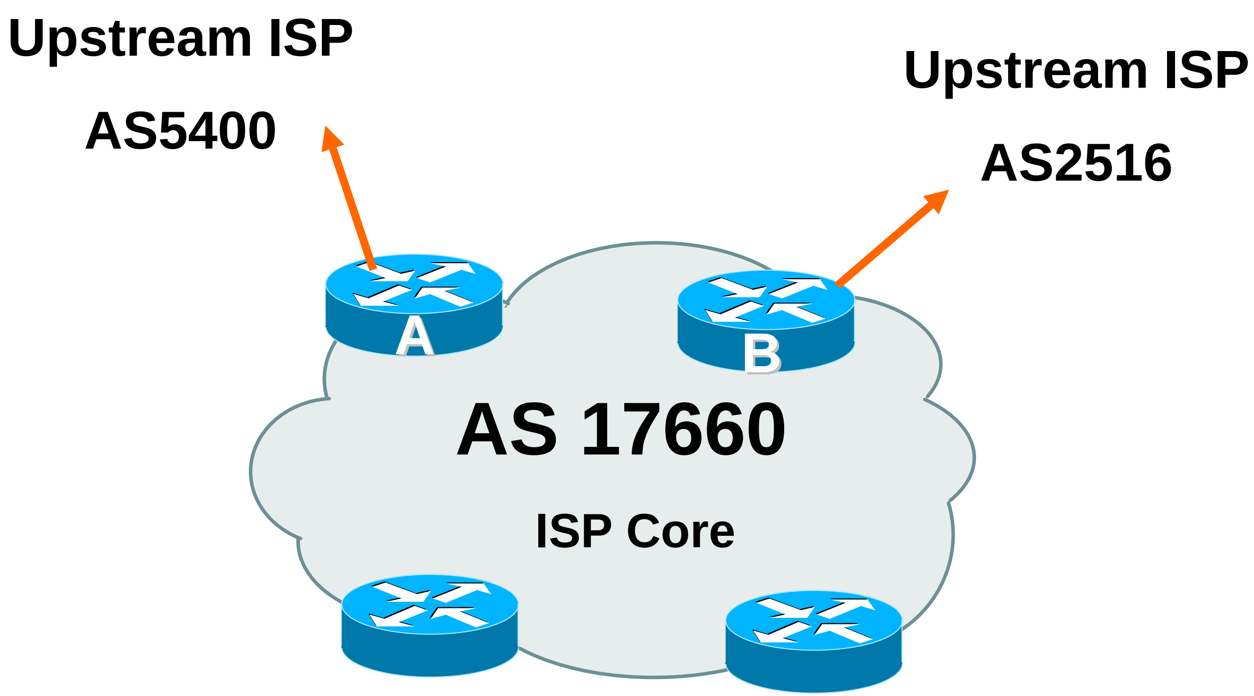
    international circuits are expensive

  Has two border routers with 64Mbytes memory

    Cannot afford to upgrade memory or hardware on border routers or internal routers

- **"Philip, make it work, please"**

Case Study



Allocated /19 from APNIC

Circuit to AS5400 is 1Mbps, circuit to AS2516 is 1Mbps

Case Study

- **Both providers stated that routers with 128Mbytes memory required for AS17660 to multihome**

  Wrong!

  Full routing table is rarely required or desired

- **Solution:**

  Accept default from one upstream

  Accept partial prefixes from the other

Case Study
Inbound Loadsharing

- **First cut: Went to a few US Looking Glasses**
  - Checked the AS path to AS5400
  - Checked the AS path to AS2516
  - AS2516 was one hop “closer”
  - Sent AS-PATH prepend of one AS on AS2516 peering

Case Study
Inbound Loadsharing

- **Refinement**

  Did not need any

  First cut worked, seeing on average 600kbps inbound on each circuit

  Does vary according to time of day, but this is as balanced as it can get, given customer profile

  ☺

Case Study
Outbound Loadsharing

- **First cut:**
  - **Requested default from AS2516**
  - **Requested full routes from AS5400**
- **Then looked at my Routing Report**
  - **Picked the top 5 ASNs and created a filter-list**
    - **If 701, 1, 7018, 1239 or 7046 are in AS-PATH, prefixes are discarded**
    - **Allowed prefixes originated by AS5400 and up to two AS hops away**
  - **Resulted in 32000 prefixes being accepted in AS17660**

Case Study
Outbound Loadsharing

- **Refinement**
  - 32000 prefixes quite a lot, seeing more outbound traffic on the AS5400 path
  - Traffic was very asymmetric
    - out through AS5400, in through AS2516
  - Added the next 3 ASNs from the Top 20 list
    - 209, 2914 and 3549
  - Now seeing 14000 prefixes
  - Traffic is now evenly loadshared outbound
    - Around 200kbps on average
    - Mostly symmetric

Case Study
Configuration Router A

```
router ospf 100
 log-adjacency-changes
 passive-interface default
 no passive-interface Ethernet0/0
 default-information originate metric 20
!
router bgp 17660
 no synchronization
 no bgp fast-external-fallover
 bgp log-neighbor-changes
 bgp deterministic-med
...next slide
```

Case Study
Configuration Router A

```
neighbor 166.49.165.13 remote-as 5400
neighbor 166.49.165.13 description eBGP multihop to AS5400
neighbor 166.49.165.13 ebgp-multihop 5
neighbor 166.49.165.13 update-source Loopback0
neighbor 166.49.165.13 prefix-list in-filter in
neighbor 166.49.165.13 prefix-list out-filter out
neighbor 166.49.165.13 filter-list 1 in
neighbor 166.49.165.13 filter-list 3 out
!
prefix-list in-filter  deny rfc1918etc in
prefix-list out-filter permit 202.144.128.0/19
!
ip route 0.0.0.0 0.0.0.0 serial 0/0 254
...next slide
```

Case Study
Configuration Router A


```
ip as-path access-list 1 deny _701_
ip as-path access-list 1 deny _1_
ip as-path access-list 1 deny _7018_
ip as-path access-list 1 deny _1239_
ip as-path access-list 1 deny _7046_
ip as-path access-list 1 deny _209_
ip as-path access-list 1 deny _2914_
ip as-path access-list 1 deny _3549_
ip as-path access-list 1 permit _5400$
ip as-path access-list 1 permit _5400_[0-9]+$
ip as-path access-list 1 permit _5400_[0-9]+_[0-9]+$
ip as-path access-list 1 deny .*
ip as-path access-list 3 permit ^$
!
```

Case Study Configuration Router B

Cisco.com

```
router ospf 100
 log-adjacency-changes
 passive-interface default
 no passive-interface Ethernet0/0
 default-information originate
!
router bgp 17660
 no synchronization
 no auto-summary
 no bgp fast-external-fallover
...next slide
```

Case Study
Configuration Router B

Cisco.com

```
bgp log-neighbor-changes
bgp deterministic-med
 neighbor 210.132.92.165 remote-as 2516
 neighbor 210.132.92.165 description eBGP peering
 neighbor 210.132.92.165 soft-reconfiguration inbound
 neighbor 210.132.92.165 prefix-list default-route in
 neighbor 210.132.92.165 prefix-list out-filter out
 neighbor 210.132.92.165 route-map as2516-out out
 neighbor 210.132.92.165 maximum-prefix 100
 neighbor 210.132.92.165 filter-list 2 in
 neighbor 210.132.92.165 filter-list 3 out
!
```

...next slide

Case Study
Configuration Router B

Cisco.com

```
!
prefix-list default-route permit 0.0.0.0/0
prefix-list out-filter permit 202.144.128.0/19
!
ip as-path access-list 2 permit _2516$
ip as-path access-list 2 deny .*
ip as-path access-list 3 permit ^$
!
route-map as2516-out permit 10
 set as-path prepend 17660
!
```

Configuration Summary

- **Router A**

  Hears full routing table – throws away most of it

  AS5400 BGP options are all or nothing

  Static default pointing to serial interface – if link goes down, OSPF default removed

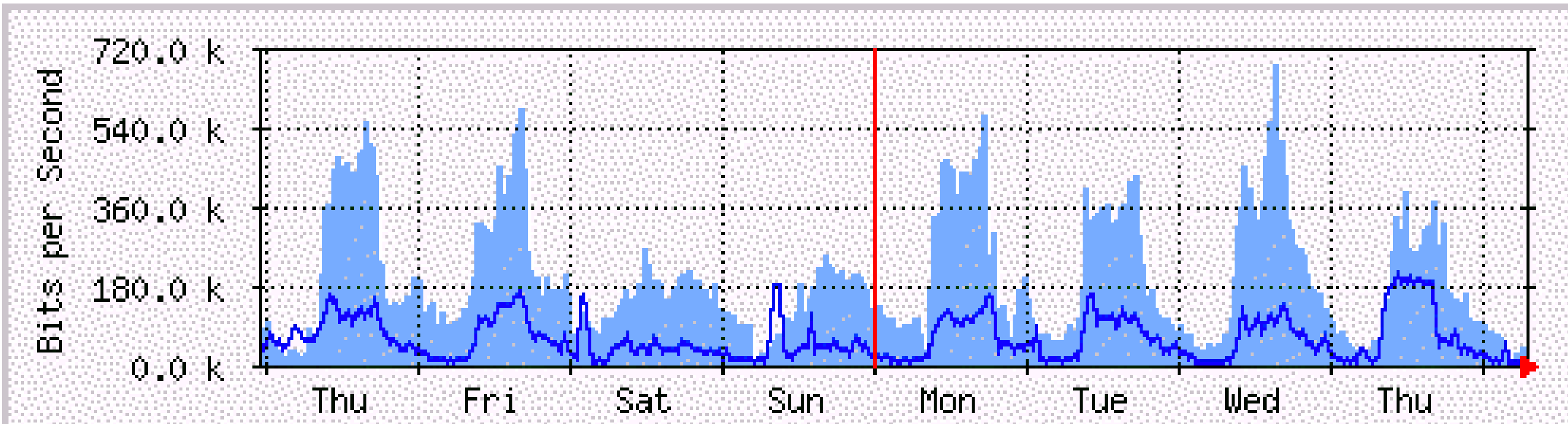
- **Router B**

  Hears default from AS2516

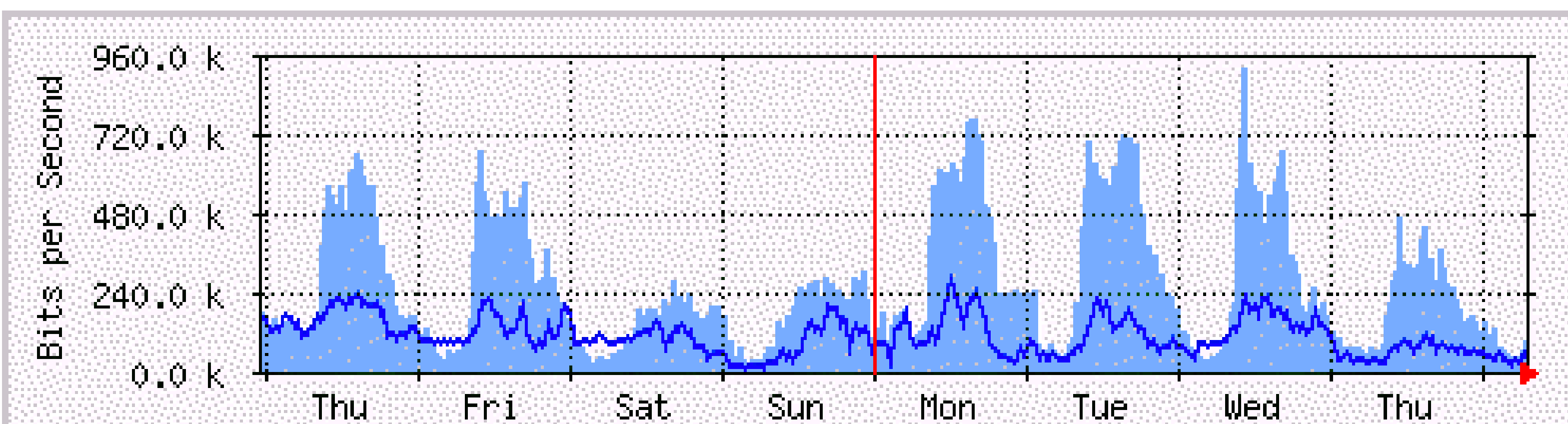
  If default disappears (BGP goes down or link goes down), OSPF default is removed

Case Study
MRTG Graphs

Cisco.com



Router A to AS5400



Router B to AS2516

Case Study Summary

- **Multihoming is not hard, really!**

    Needs a bit of thought, a bit of planning

    Use this case study as an example strategy

    Does not require sophisticated equipment, big memory, fast CPUs…

BGP for Internet Service Providers

- BGP Basics (quick recap)
- Scaling BGP
- Deploying BGP in an ISP network
- Multihoming Examples
- Using Communities

Cisco.com

Communities

RFC1998

- **Informational RFC**
- **Describes how to implement loadsharing and backup on multiple inter-AS links**

  BGP communities used to determine local preference in upstream's network
- **Gives control to the customer**
- **Simplifies upstream's configuration**

  simplifies network operation!

RFC1998

- **Community values defined to have particular meanings:**

ASx:100	set local pref 100	preferred route
ASx:90	set local pref 90	backup route if dualhomed on ASx
ASx:80	set local pref 80	main link is to another ISP with same AS path length
ASx:70	set local pref 70	main link is to another ISP

RFC1998

Cisco.com

- **Sample Customer Router Configuration**

```
router bgp 107
 neighbor x.x.x.x remote-as 109
 neighbor x.x.x.x description Backup ISP
 neighbor x.x.x.x route-map config-community out
 neighbor x.x.x.x send-community
!
ip as-path access-list 20 permit ^$
ip as-path access-list 20 deny .*
!
route-map config-community permit 10
 match as-path 20
 set community 109:90
```

RFC1998

Cisco.com

- **Sample ISP Router Configuration**

```
! Homed to another ISP
ip community-list 70 permit 109:70
! Homed to another ISP with equal ASPATH length
ip community-list 80 permit 109:80
! Customer backup routes
ip community-list 90 permit 109:90
!
route-map set-customer-local-pref permit 10
 match community 70
 set local-preference 70
```

RFC1998

Cisco.com

- **Sample ISP Router Configuration**

```
route-map set-customer-local-pref permit 20
 match community 80
 set local-preference 80
!
route-map set-customer-local-pref permit 30
 match community 90
 set local-preference 90
!
route-map set-customer-local-pref permit 40
 set local-preference 100
```

RFC1998

Cisco.com

- **Supporting RFC1998**

  many ISPs do, more should

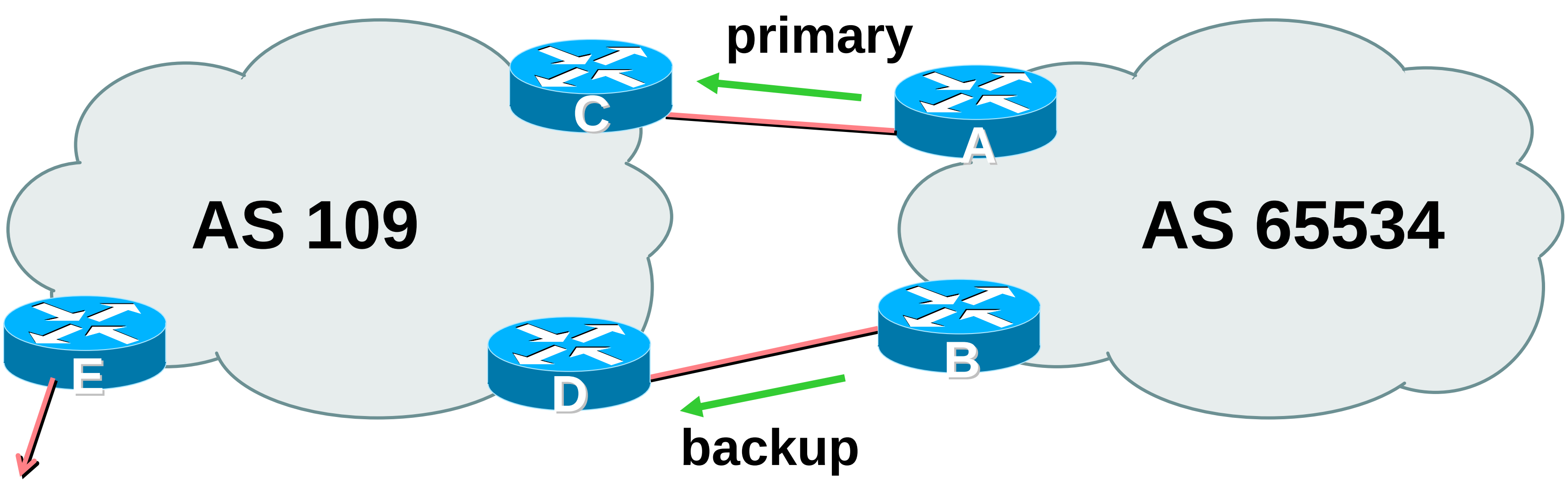
  check AS object in the Internet Routing Registry

  if you do, insert comment in AS object in the IRR

Two links to the same ISP

One link primary, the other link backup only

Two links to the same ISP



- **AS109 "proxy aggregates" for AS 65534**

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

- **Announce /19 aggregate on each link**
  - primary link makes standard announcement
  - backup link sends community
- **When one link fails, the announcement of the /19 aggregate via the other link ensures continued connectivity**

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

- Router A Configuration

```
router bgp 65534
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.2 remote-as 109
 neighbor 222.222.10.2 description RouterC
 neighbor 222.222.10.2 prefix-list aggregate out
 neighbor 222.222.10.2 prefix-list default in
!
ip prefix-list aggregate permit 221.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
```

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

Cisco.com

- **Router B Configuration**

```
router bgp 65534
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.6 remote-as 109
 neighbor 222.222.10.6 description RouterD
 neighbor 222.222.10.6 send-community
 neighbor 222.222.10.6 prefix-list aggregate out
 neighbor 222.222.10.6 route-map routerD-out out
 neighbor 222.222.10.6 prefix-list default in
 neighbor 222.222.10.6 route-map routerD-in in
!
..next slide
```

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


```
ip prefix-list aggregate permit 221.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
route-map routerD-out permit 10
 match ip address prefix-list aggregate
 set community 109:90
route-map routerD-out permit 20
!
route-map routerD-in permit 10
 set local-preference 90
!
```

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

- Router C Configuration (main link)

```
router bgp 109
 neighbor 222.222.10.1 remote-as 65534
 neighbor 222.222.10.1 default-originate
 neighbor 222.222.10.1 prefix-list Customer in
 neighbor 222.222.10.1 prefix-list default out
!
ip prefix-list Customer permit 221.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
```

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

Cisco.com

- Router D Configuration (backup link)

```
router bgp 109
 neighbor 222.222.10.5 remote-as 65534
 neighbor 222.222.10.5 default-originate
 neighbor 222.222.10.5 prefix-list Customer in
 neighbor 222.222.10.5 route-map bgp-cust-in in
 neighbor 222.222.10.5 prefix-list default out
!
ip prefix-list Customer permit 221.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
..next slide
```

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

Cisco.com

```
ip prefix-list Customer permit 221.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
ip community-list 90 permit 109:90
!
<snip>
route-map bgp-cust-in permit 30
 match community 90
 set local-preference 90
route-map bgp-cust-in permit 40
 set local-preference 100
```

Service Providers use of Communities

Some working examples

Some ISP Examples

Cisco.com

- ISPs create communities to give customers bigger routing policy control
- Public policy is usually listed in the IRR
    - Following examples are all in the IRR
- 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


```
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

Cisco.com

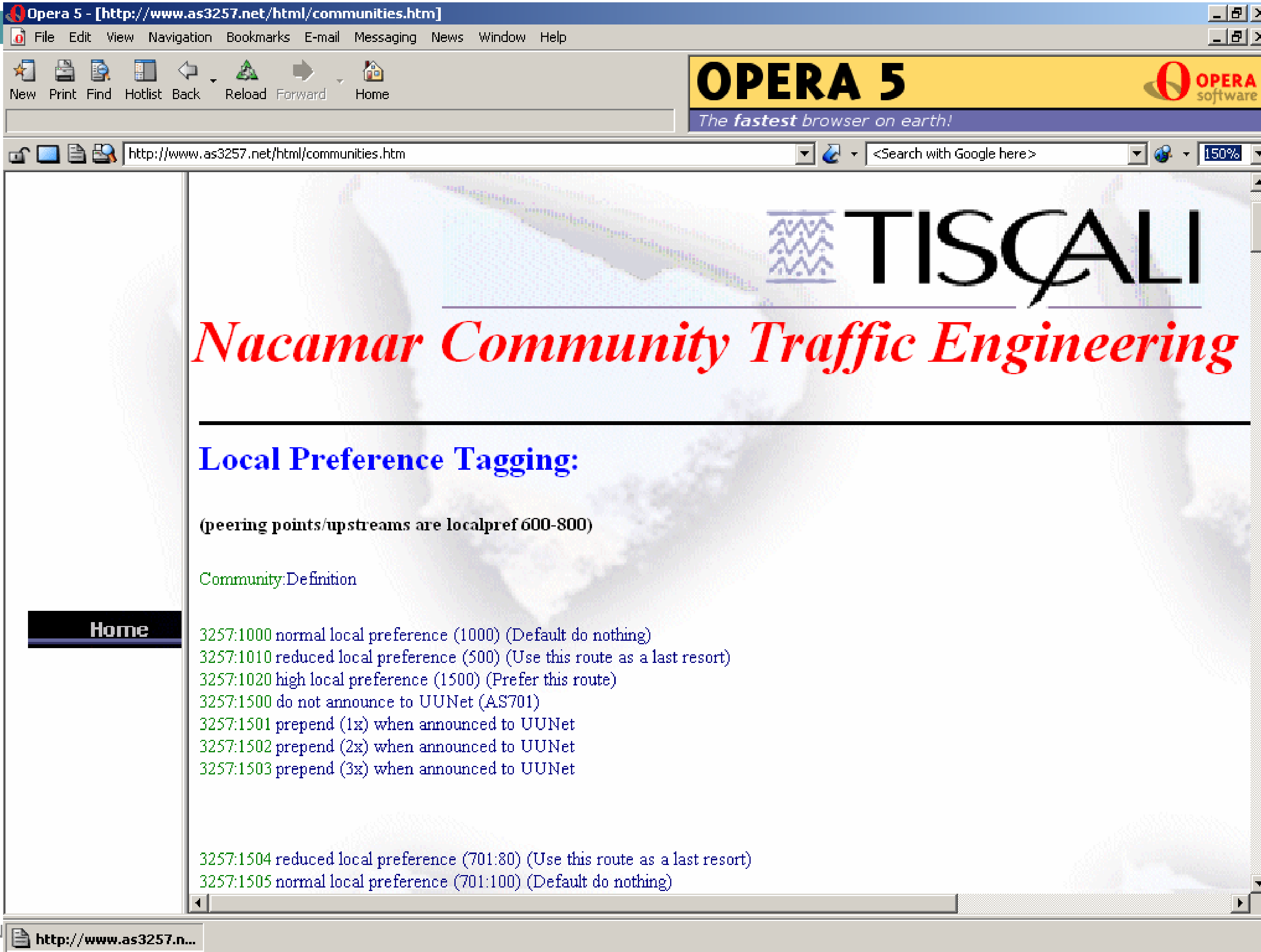
```
aut-num: AS702
as-name: AS702
descr:   UUNET - Commercial IP service provider in Europe
remarks: -------------------------------------------------------------
remarks: UUNET uses the following communities with its customers:
remarks: 702:80  Set Local Pref 80 within AS702
remarks: 702:120 Set Local Pref 120 within AS702
remarks: 702:20  Announce only to UUNET AS'es and UUNET customers
remarks: 702:30  Keep within Europe, don't announce to other UUNET AS's
remarks: 702:1   Prepend AS702 once at edges of UUNET to Peers
remarks: 702:2   Prepend AS702 twice at edges of UUNET to Peers
remarks: 702:3   Prepend AS702 thrice at edges of UUNET to Peers
remarks: Details of UUNET's peering policy and how to get in touch with
remarks: UUNET regarding peering policy matters can be found at:
remarks: http://www.uu.net/peering/
remarks: --------------------------------------------------------------
mnt-by:  UUNET-MNT
changed: eric-apps@eu.uu.net 20010928
source:  RIPE
```

Some ISP Examples
Concert Europe

Cisco.com

```
aut-num:      AS5400
as-name:      CIPCORE
descr:        Concert European Core Network
remarks:      Communities scheme:
remarks:      The following BGP communities can be set by Concert BGP
remarks:      customers to affect announcements to major peerings.
remarks:
remarks:      Community to                          Community to
remarks:      Not announce      To peer:            AS prepend 5400
remarks:
remarks:      5400:1000         European peers      5400:2000
remarks:      5400:1001         Ebone (AS1755)      5400:2001
remarks:      5400:1002         Eunet (AS286)       5400:2002
remarks:      5400:1003         Unisource (AS3300)  5400:2003
<snip>
remarks:      5400:1100         US peers            5400:2100
notify:       peertech@concert.net
mnt-by:       CIP-MNT
source:       RIPE
```

Some ISP Examples Tiscali/Nacamar



ISP Examples

- **Several more…**
- **Tiscali is very detailed**

  Consult their website for more information

  Includes IOS configuration examples

- **Many ISP support communities for multihoming preferences**

BGP for Internet Service Providers

- BGP Basics (quick recap)
- Scaling BGP
- Deploying BGP in an ISP network
- Multihoming Examples
- Using Communities

BGP for Internet Service Providers

End of Tutorial