BGP Tutorial

Philip Smith <pfs@cisco.com>

APRICOT 2003, Taipei

February 2003

APRICOT BGP Tutorials

- **Four Tutorials over Two Days**

Part 1 – Introduction	**Monday morning**
Part 2 – Deployment	**Monday afternoon**
Part 3 – Multihoming	**Tuesday morning**
Part 4 – Troubleshooting	**Tuesday afternoon**

BGP Tutorial
Part 1 – Introduction

Philip Smith <pfs@cisco.com>

APRICOT 2003, Taipei

February 2003

Presentation Slides

- **Slides are available at**

  ftp://ftp-eng.cisco.com/pfs/seminars/APRICOT02-BGP00.pdf

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

BGP for Internet Service Providers

- Routing Basics
- BGP Basics
- BGP Attributes
- BGP Path Selection
- BGP Policy
- BGP Capabilities
- Scaling BGP

Routing Basics

Terminology and Concepts

Routing Concepts

- IPv4
- Routing
- Forwarding
- Some definitions
- Policy options
- Routing Protocols

IPv4

- **Internet uses IPv4**
  - **addresses are 32 bits long**
  - **range from 1.0.0.0 to 223.255.255.255**
  - **0.0.0.0 to 0.255.255.255 and 224.0.0.0 to 255.255.255.255 have "special" uses**
- **IPv4 address has a network portion and a host portion**

IPv4 address format

- **Address and subnet mask**

  written as

  12.34.56.78 255.255.255.0 *or*

  12.34.56.78/24

  mask represents the number of network bits in the 32 bit address

  the remaining bits are the host bits

What does a router do?

A day in a life of a router

Cisco.com

find path

forward packet, forward packet, forward packet, forward packet...

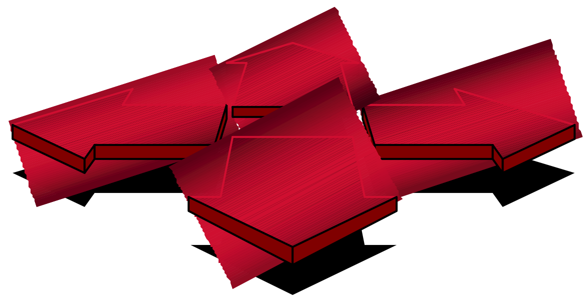
find alternate path

forward packet, forward packet, forward packet, forward packet…

repeat until powered off

Routing versus Forwarding

- **Routing = building maps and giving directions**
- **Forwarding = moving packets between interfaces according to the "directions"**



IP Routing – finding the path

- **Path derived from information received from a routing protocol**
- **Several alternative paths may exist**

  best next hop stored in forwarding table

- **Decisions are updated periodically or as topology changes (event driven)**
- **Decisions are based on:**

  topology, policies and metrics (hop count, filtering, delay, bandwidth, etc.)

IP route lookup

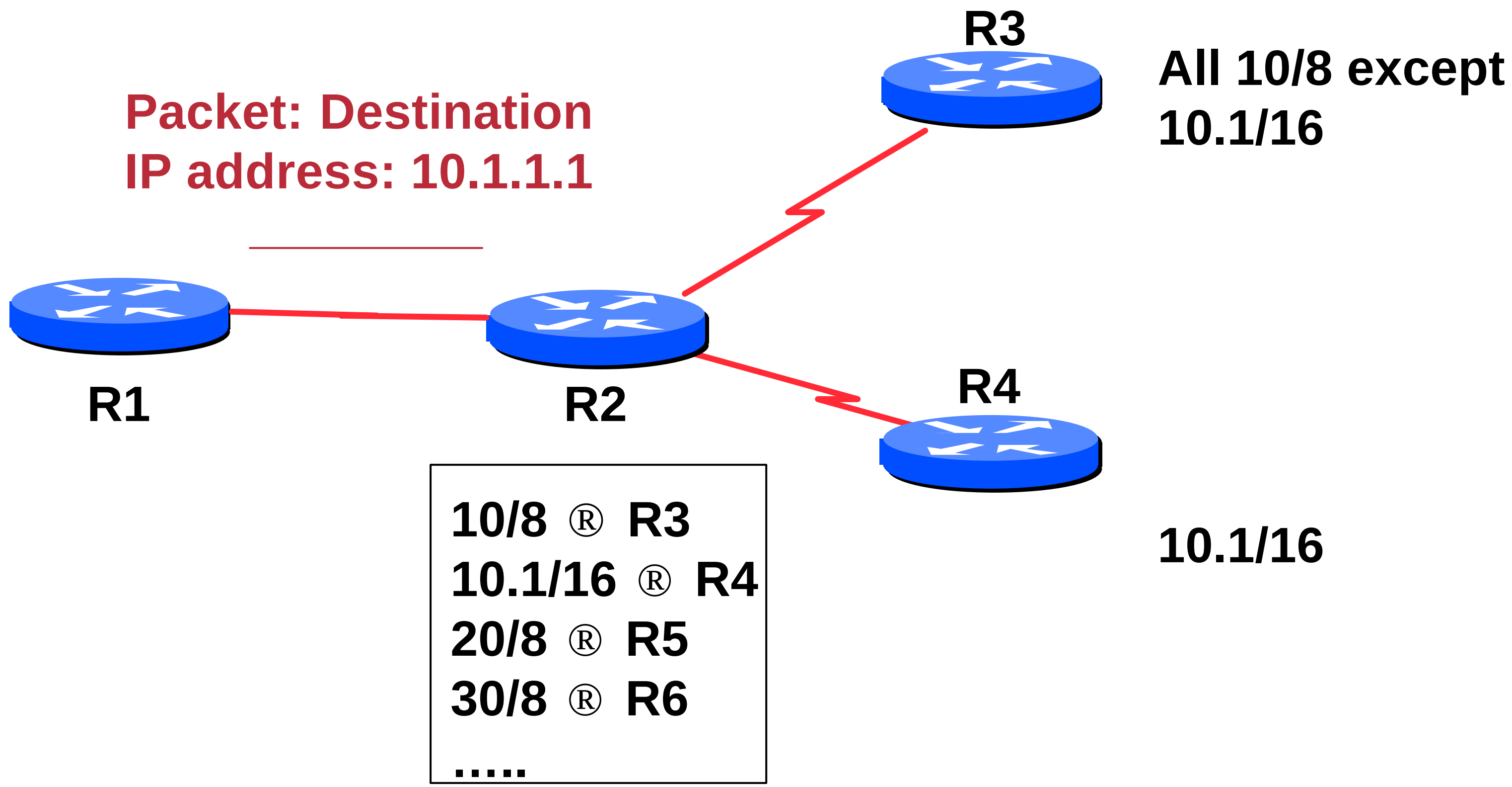
- Based on destination IP packet
- "longest match" routing

  more specific prefix preferred over less specific prefix

  example: packet with destination of 10.1.1.1/32 is sent to the router announcing 10.1/16 rather than the router announcing 10/8.

IP route lookup

- Based on destination IP packet

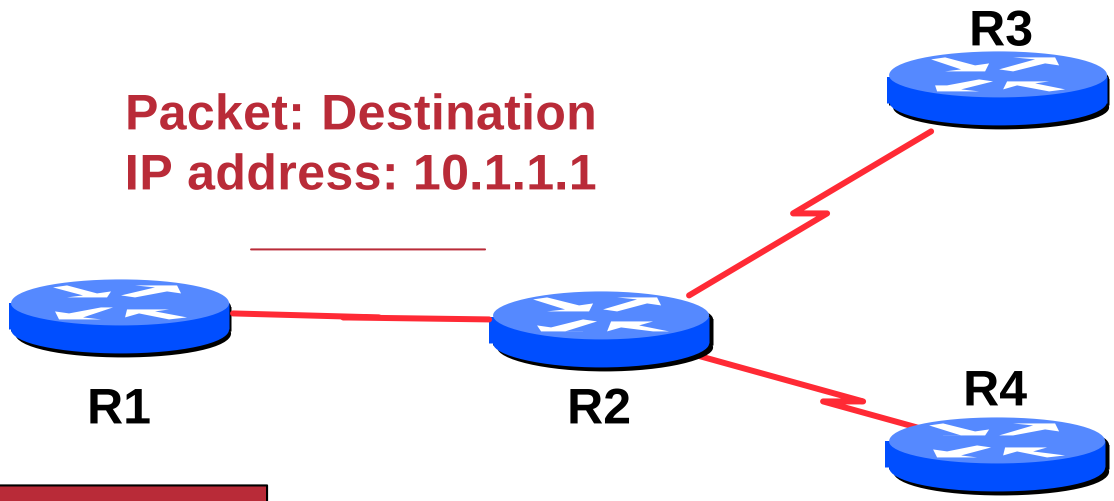


R2's IP routing table

IP route lookup: Longest match routing

- **Based on destination IP packet**

Packet: Destination
IP address: 10.1.1.1



All 10/8 except 10.1/16

10.1/16

10/8 ® R3
10.1/16 ® R4
20/8 ® R5
30/8 ® R6
…..

→ **10.1.1.1 && FF.0.0.0**
vs.
10.0.0.0 && FF.0.0.0

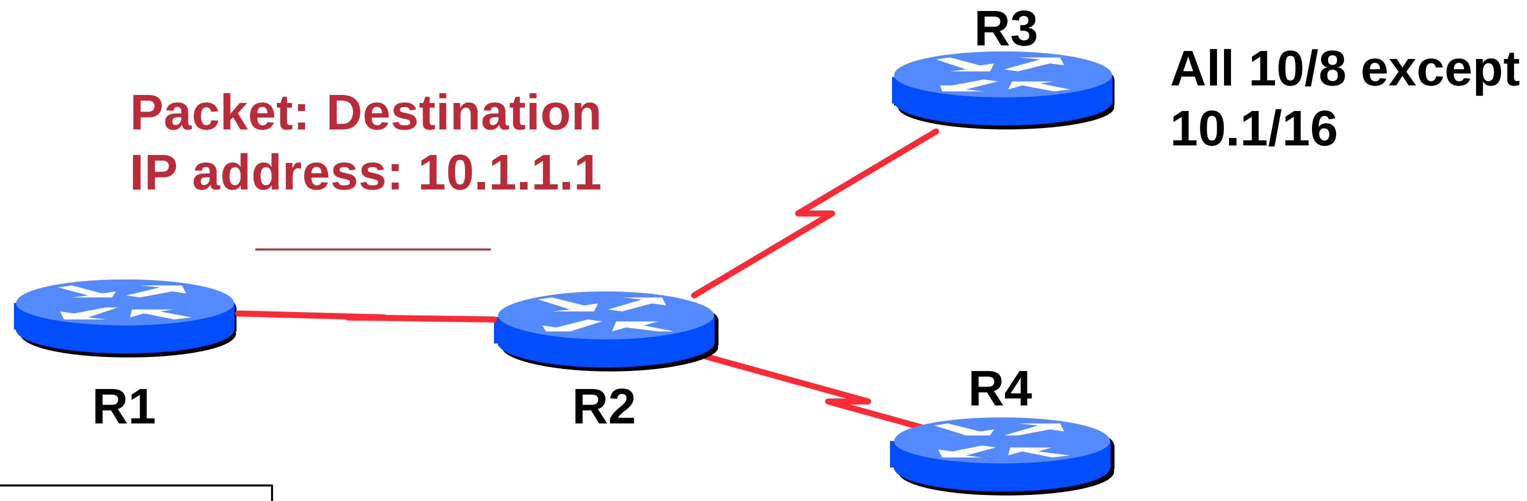
Match!

R2's IP routing table

IP route lookup:
Longest match routing

Cisco.com

- **Based on destination IP packet**



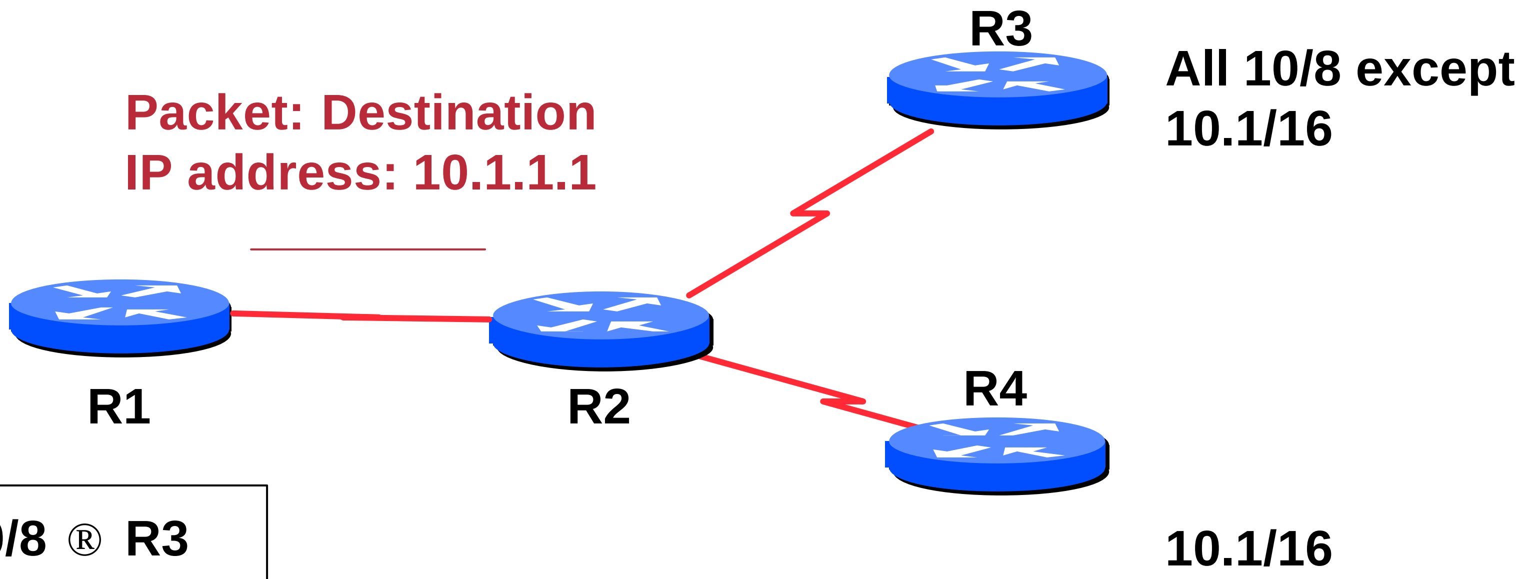
10/8 ® R3	
10.1/16 ® R4	→ 10.1.1.1 && FF.FF.0.0
20/8 ® R5	vs.
30/8 ® R6	10.1.0.0 && FF.FF.0.0
…..	

Match as well!

R2's IP routing table

IP route lookup: Longest match routing

- Based on destination IP packet



10/8 ® R3
10.1/16 ® R4
20/8 ® R5
30/8 ® R6
…..

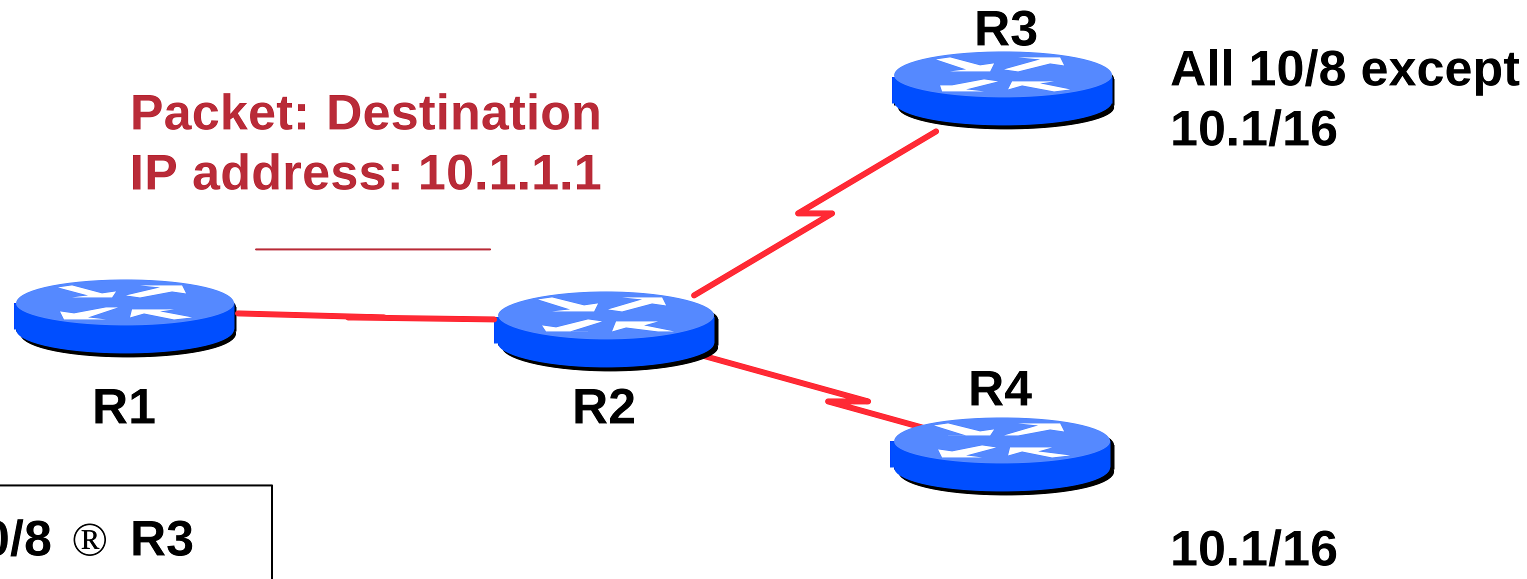
R2's IP routing table

10.1.1.1 && FF.0.0.0
vs.
20.0.0.0 && FF.0.0.0

Does not match!

IP route lookup: Longest match routing

- Based on destination IP packet



10/8 ® R3
10.1/16 ® R4
20/8 ® R5
30/8 ® R6
…..

10.1.1.1 && FF.0.0.0
vs.
30.0.0.0 && FF.0.0.0

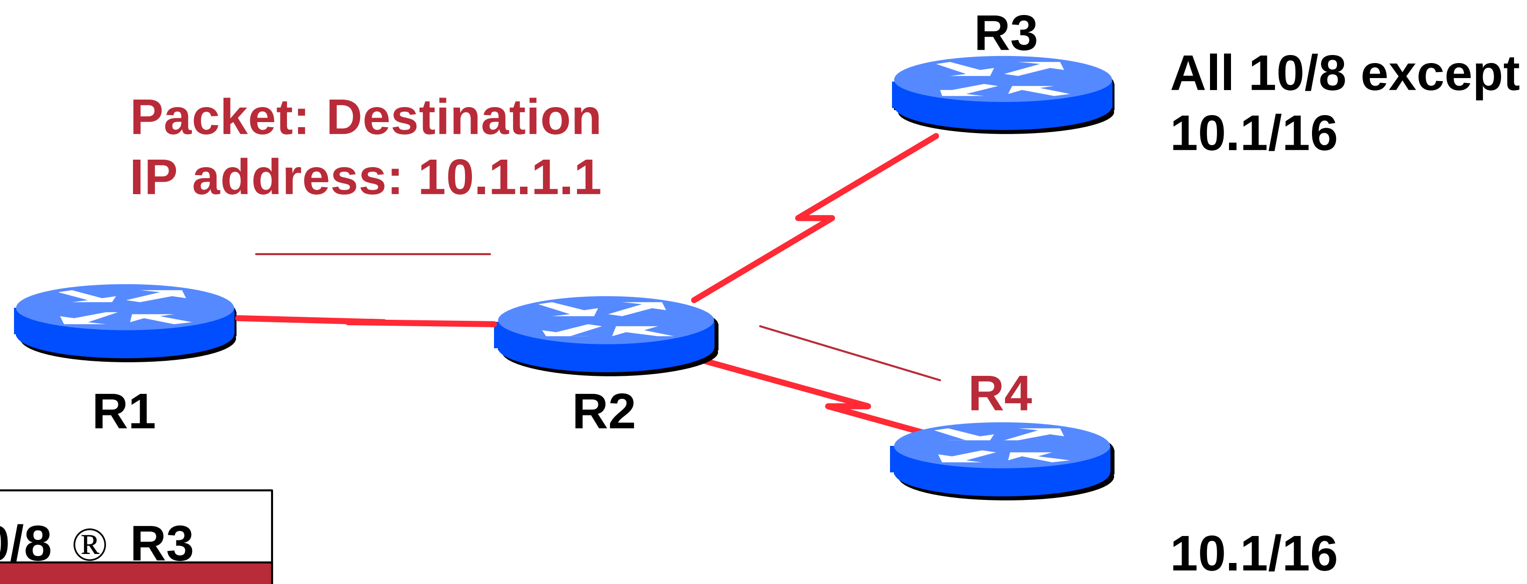
Does not match!

R2's IP routing table

IP route lookup: Longest match routing

Cisco.com

- **Based on destination IP packet**



10/8 ® R3
10.1/16 ® R4
20/8 ® R5
30/8 ® R6
…..

Longest match, 16 bit netmask

R2's IP routing table

IP Forwarding

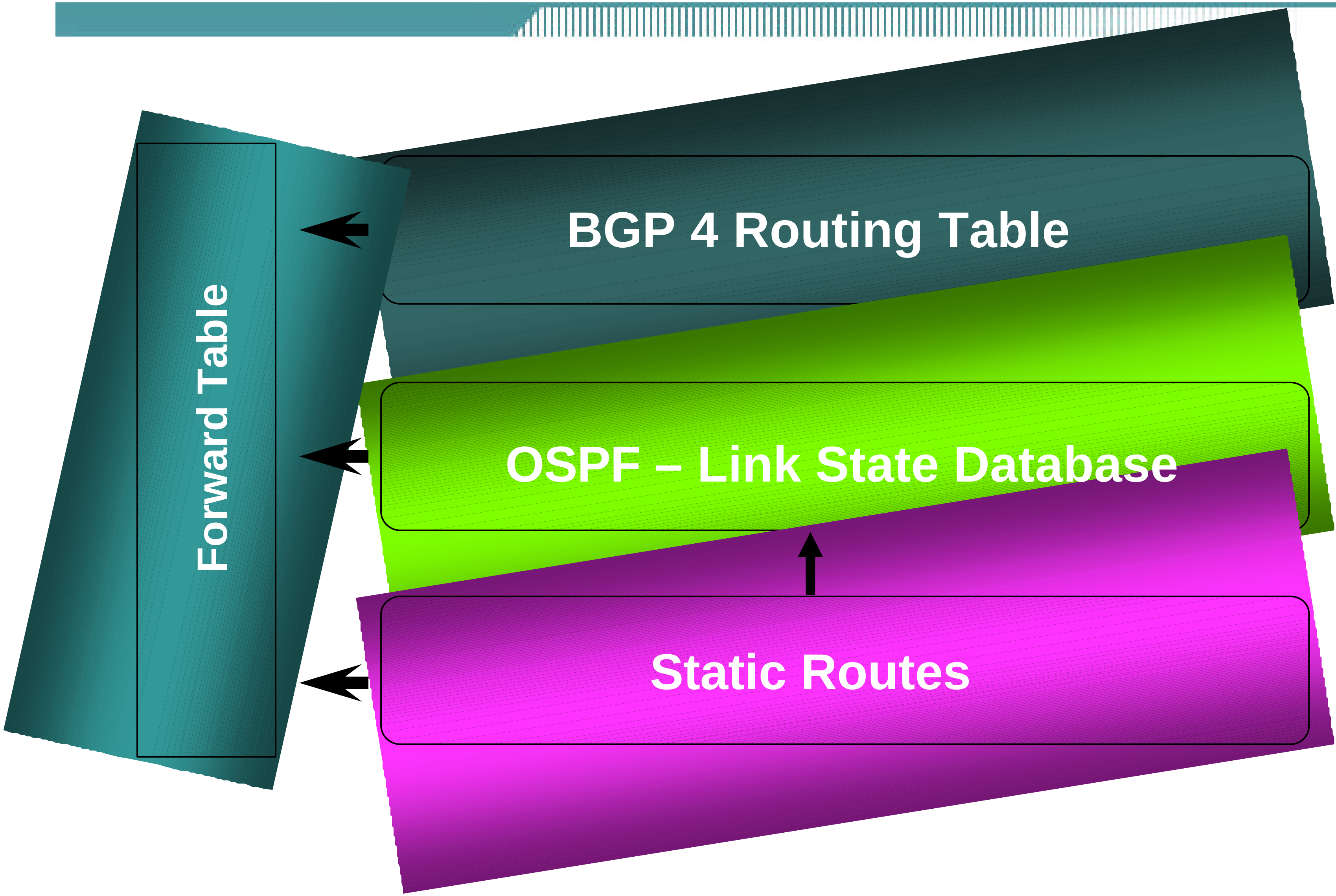
- **Router makes decision on which interface a packet is sent to**
- **Forwarding table populated by routing process**
- **Forwarding decisions:**

  destination address

  class of service (fair queuing, precedence, others)

  local requirements (packet filtering)
- **Can be aided by special hardware**

Routing Tables Feed the Forwarding Table



Explicit versus Default routing

Cisco.com

- **Default:**

  simple, cheap (cycles, memory, bandwidth)

  low granularity (metric games)

- **Explicit (default free zone)**

  high overhead, complex, high cost, high granularity

- **Hybrid**

  minimise overhead

  provide useful granularity

  requires some filtering knowledge

Egress Traffic

- **How packets leave your network**
- **Egress traffic depends on:**

  route availability (what others send you)

  route acceptance (what you accept from others)

  policy and tuning (what you do with routes from others)

  Peering and transit agreements

Ingress Traffic

- **How packets get to your network and your customers' networks**
- **Ingress traffic depends on:**

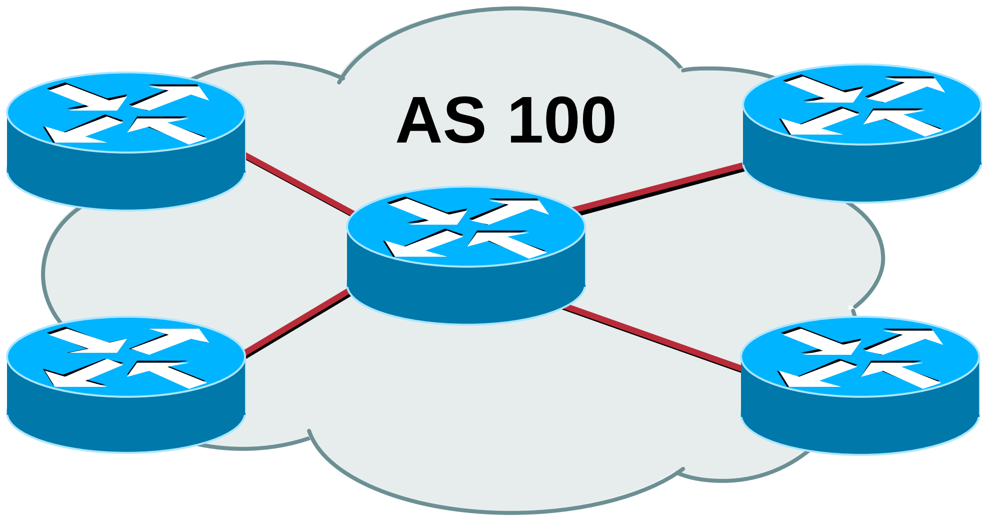
  what information you send and to whom

  based on your addressing and AS's

  based on others' policy (what they accept from you and what they do with it)

Autonomous System (AS)

Cisco.com



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

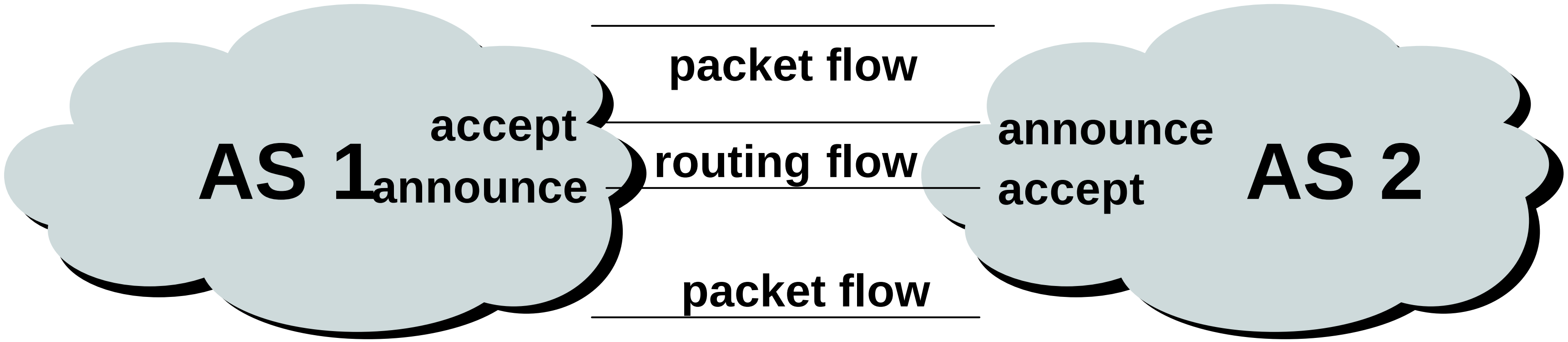
Definition of terms

Cisco.com

- **Neighbours** – AS's which directly exchange routing information
- **Announce** – send routing information to a neighbour
- **Accept** – receive and use routing information sent by a neighbour
- **Originate** – insert routing information into external announcements (usually as a result of the IGP)
- **Peers** – routers in neighbouring AS's or within one AS which exchange routing and policy information

Routing flow and packet flow

Cisco.com



For networks in AS1 and AS2 to communicate:

AS1 must announce to AS2

AS2 must accept from AS1

AS2 must announce to AS1

AS1 must accept from AS2

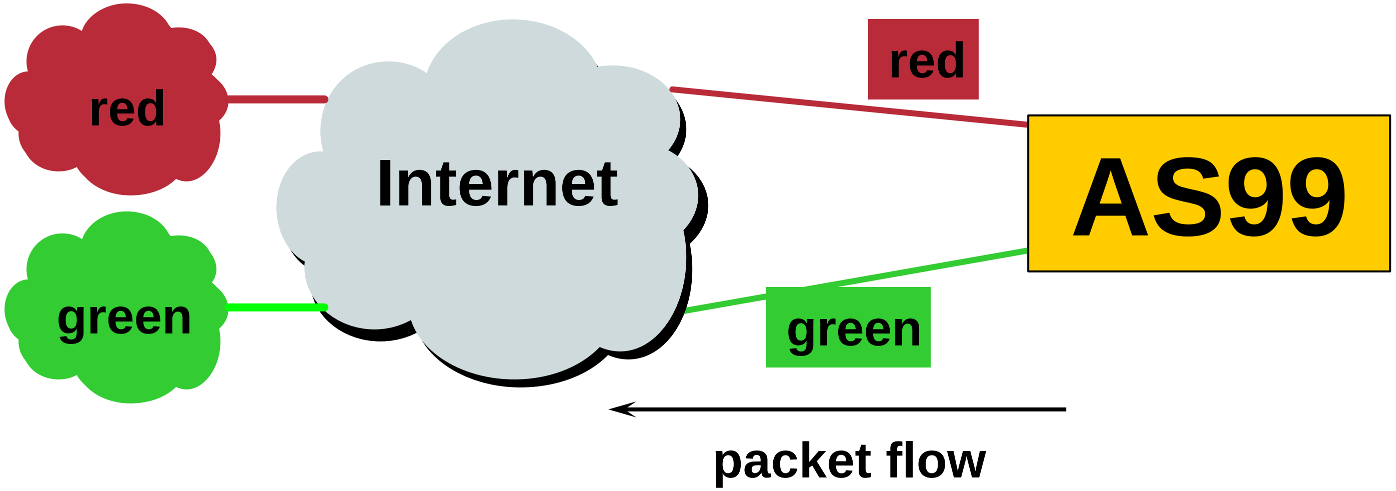
Routing flow and Traffic flow

- **Traffic flow is always in the opposite direction of the flow of routing information**

  filtering outgoing routing information inhibits traffic flowing in

  filtering incoming routing information inhibits traffic flowing out

Routing policy limitations



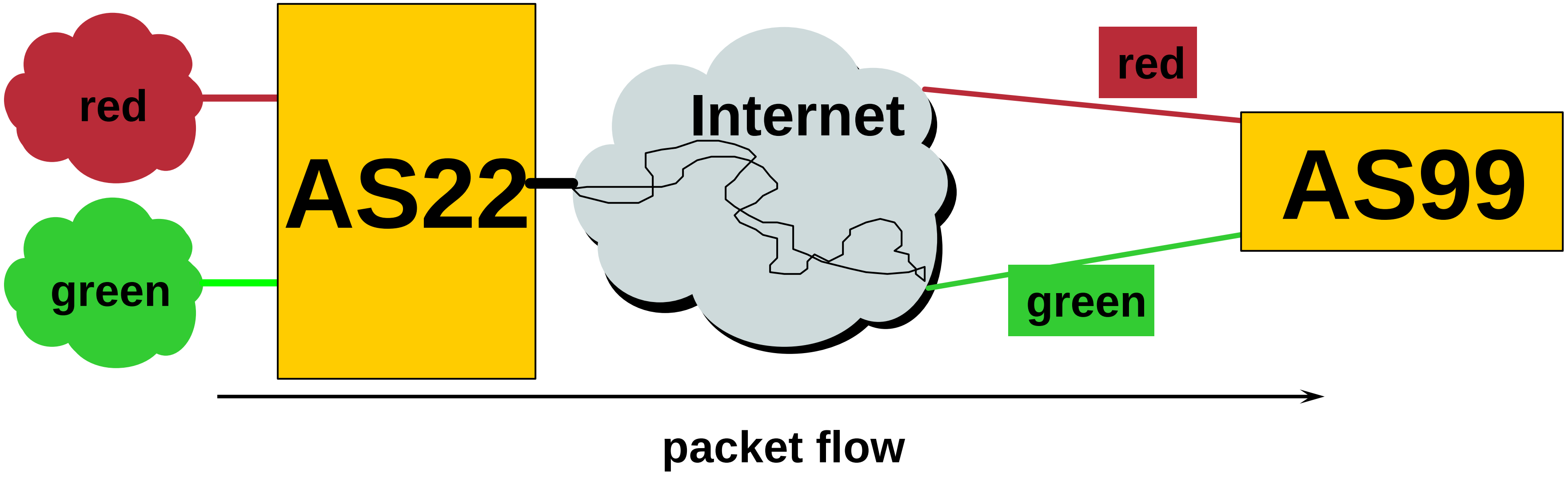
AS99 uses red link for traffic going to the red AS and green link for traffic going to the green AS

To implement this policy for AS99:

- **accept routes originating in the red AS on the red link**
- **accept all other routes on the green link**

Routing policy limitations

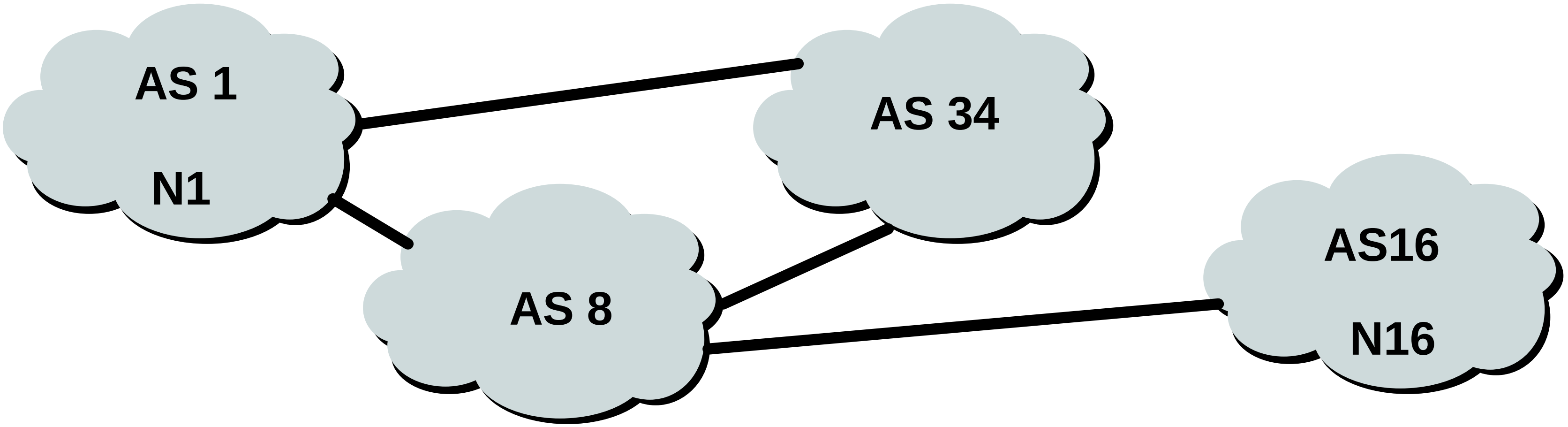
Cisco.com



For packets flowing *toward* AS 99:

Unless AS 22 and all other intermediate AS's co-operate in pushing green traffic to the green link then some reasonable policies can not be implemented.

Routing policy with multiple ASes



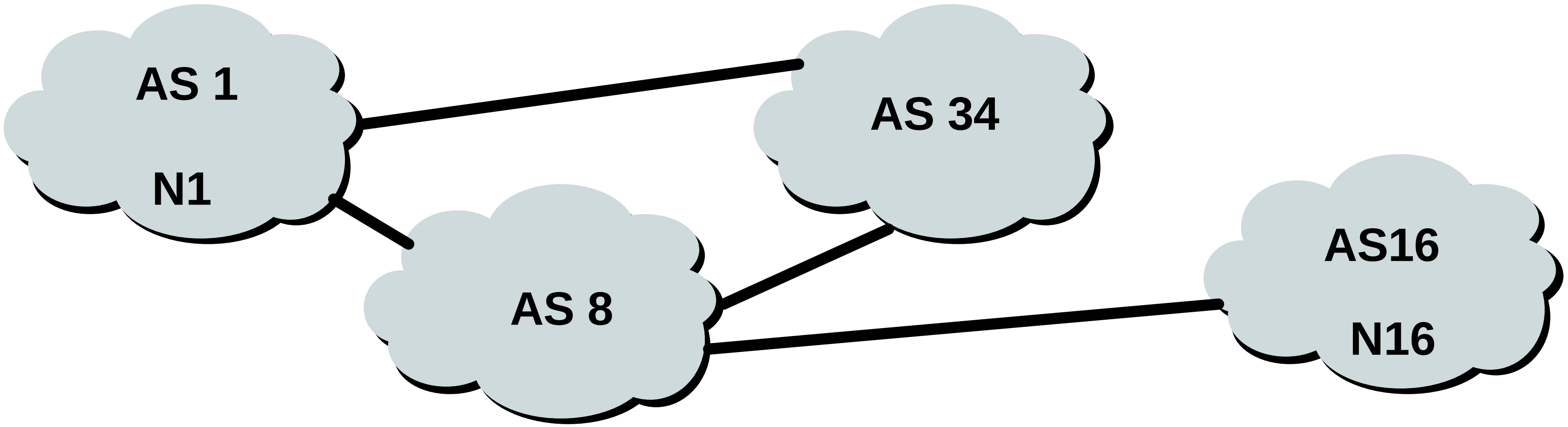
For net N1 in AS1 to send traffic to net N16 in AS16:

- **AS16 must originate and announce N16 to AS8.**
- **AS8 must accept N16 from AS16.**
- **AS8 must announce N16 to AS1 or AS34.**
- **AS1 must accept N16 from AS8 or AS34.**

For two-way packet flow, similar policies must exist for N1.

Routing policy with multiple AS's

Cisco.com



As multiple paths between sites are implemented it is easy to see how policies can become quite complex.

Granularity of routing policy

- **What to announce/accept**
- **Preferences between multiple accepts**

  single route

  routes originated by single AS

  routes originated by a group of AS's

  routes traversing specific path

  routes traversing specific AS

  routes belonging to other groupings (including combinations)

Routing Policy Issues

Cisco.com

- 120000 prefixes (not realistic to set policy on all of them individually)
- 15000 origin AS's (too many)
- routes tied to a specific AS or path may be unstable regardless of connectivity
- groups of AS's are a natural abstraction for filtering purposes

What Is an IGP?

- Interior Gateway Protocol
- Within an Autonomous System
- Carries information about internal infrastructure prefixes
- Examples – OSPF, ISIS, EIGRP…

Why Do We Need an IGP?

- **ISP backbone scaling**

  Hierarchy

  Modular infrastructure construction

  Limiting scope of failure

  Healing of infrastructure faults using dynamic routing with fast convergence

What Is an EGP?

- Exterior Gateway Protocol
- Used to convey routing information between Autonomous Systems
- De-coupled from the IGP
- Current EGP is BGP

Why Do We Need an EGP?

- **Scaling to large network**
  - **Hierarchy**
  - **Limit scope of failure**
- **Define Administrative Boundary**
- **Policy**
  - **Control reachability to prefixes**
  - **Merge separate organizations**
  - **Connect multiple IGPs**

Interior versus Exterior Routing Protocols

- **Interior**
  - automatic neighbour discovery
  - generally trust your IGP routers
  - prefixes go to all IGP routers
  - binds routers in one AS together

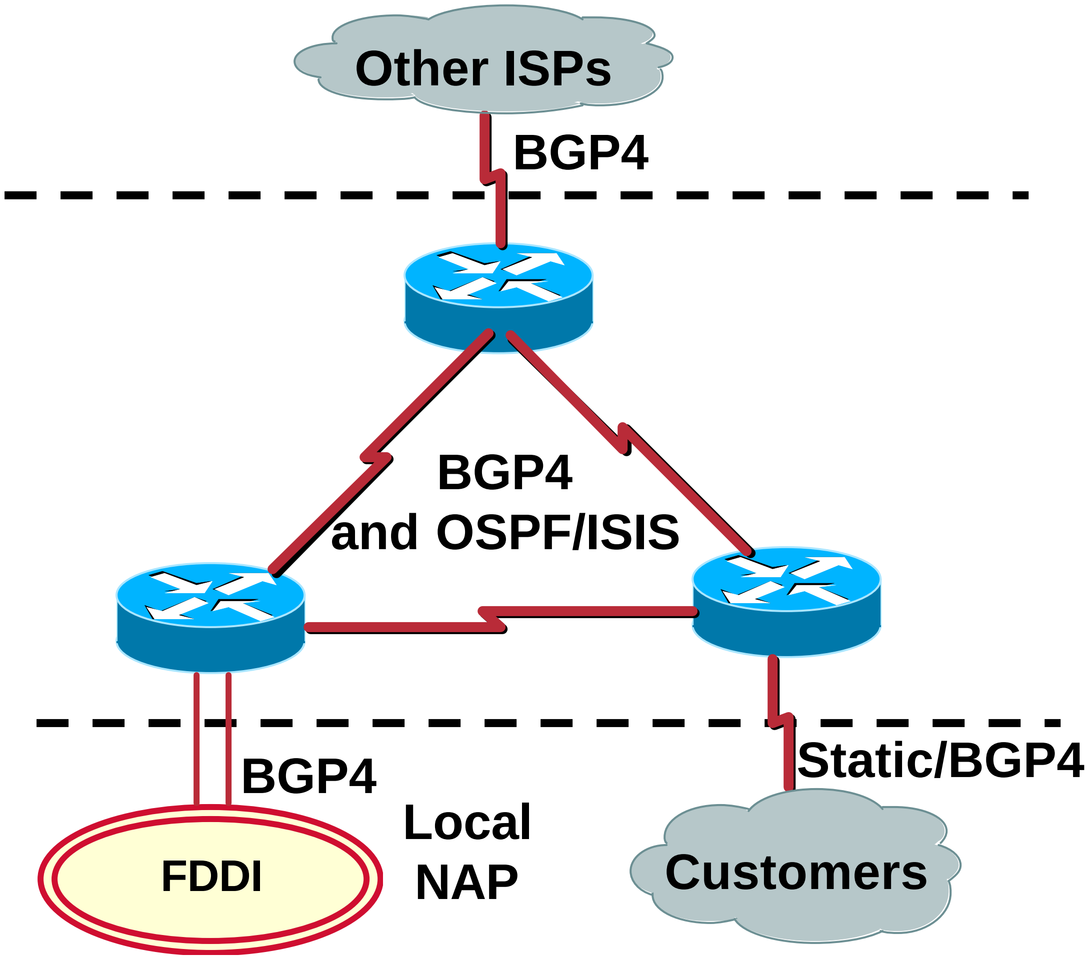
- **Exterior**
  - specifically configured peers
  - connecting with outside networks
  - set administrative boundaries
  - binds AS's together

Interior versus Exterior Routing Protocols

Cisco.com

- **Interior**
  - Carries ISP infrastructure addresses only
  - ISPs aim to keep the IGP small for efficiency and scalability
- **Exterior**
  - Carries customer prefixes
  - Carries Internet prefixes
  - EGPs are independent of ISP network topology

Hierarchy of Routing Protocols



Default Administrative Distances

Route Source	Default Distance
Connected Interface	0
Static Route	1
Enhanced IGRP Summary Route	5
External BGP	20
Internal Enhanced IGRP	90
IGRP	100
OSPF	110
IS-IS	115
RIP	120
EGP	140
External Enhanced IGRP	170
Internal BGP	200
Unknown	255

BGP for Internet Service Providers

- Routing Basics
- BGP Basics
- BGP Attributes
- BGP Path Selection
- BGP Policy
- BGP Capabilities
- Scaling BGP

BGP Basics

What is this BGP thing?

Border Gateway Protocol

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

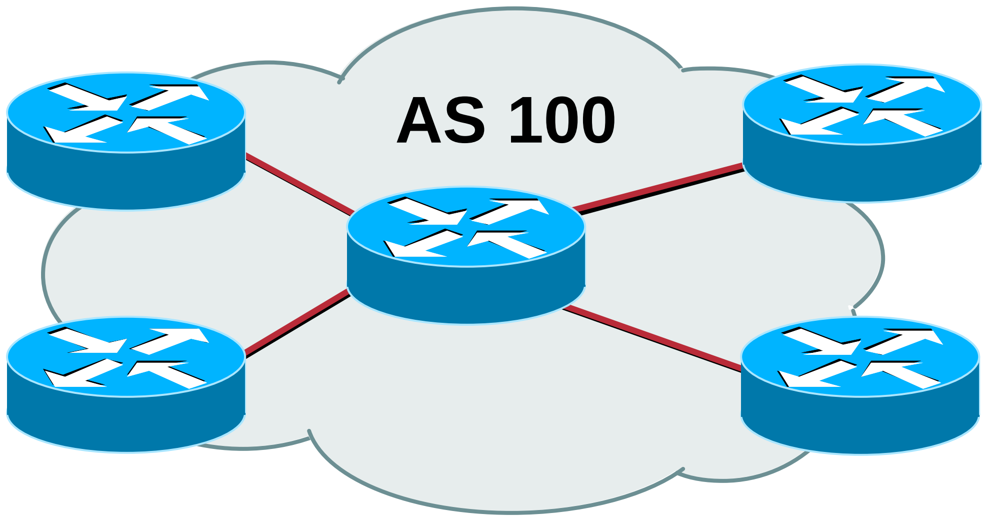
  exterior gateway protocol

- **Described in RFC1771**

  work in progress to update

  www.ietf.org/internet-drafts/draft-ietf-idr-bgp4-18.txt

Autonomous System (AS)

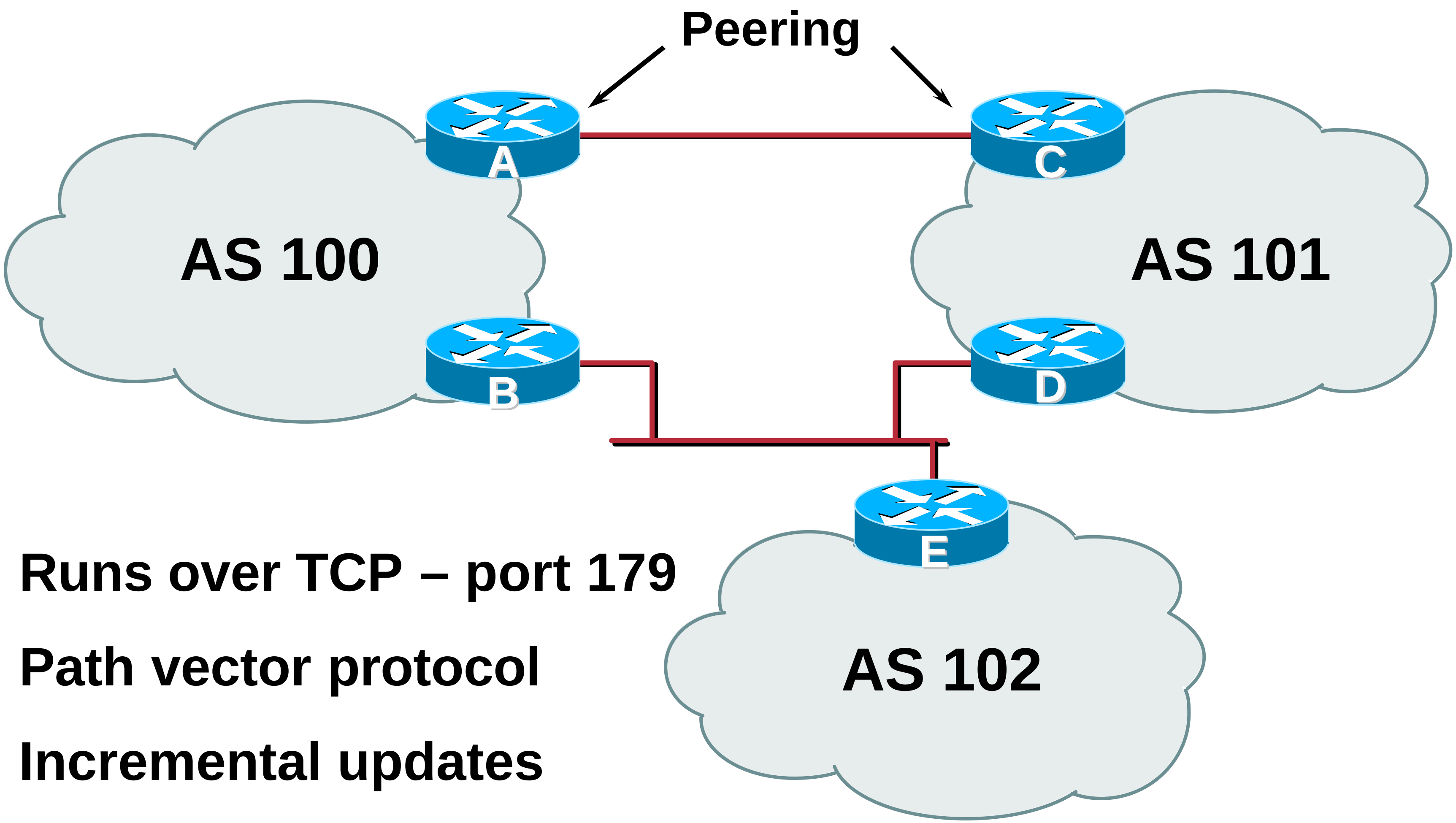


- Collection of networks with same routing policy
- Single routing protocol
- Usually under single ownership, trust and administrative control
- Identified by a unique number

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-06.txt**
- **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 29695 have been made to the RIRs**

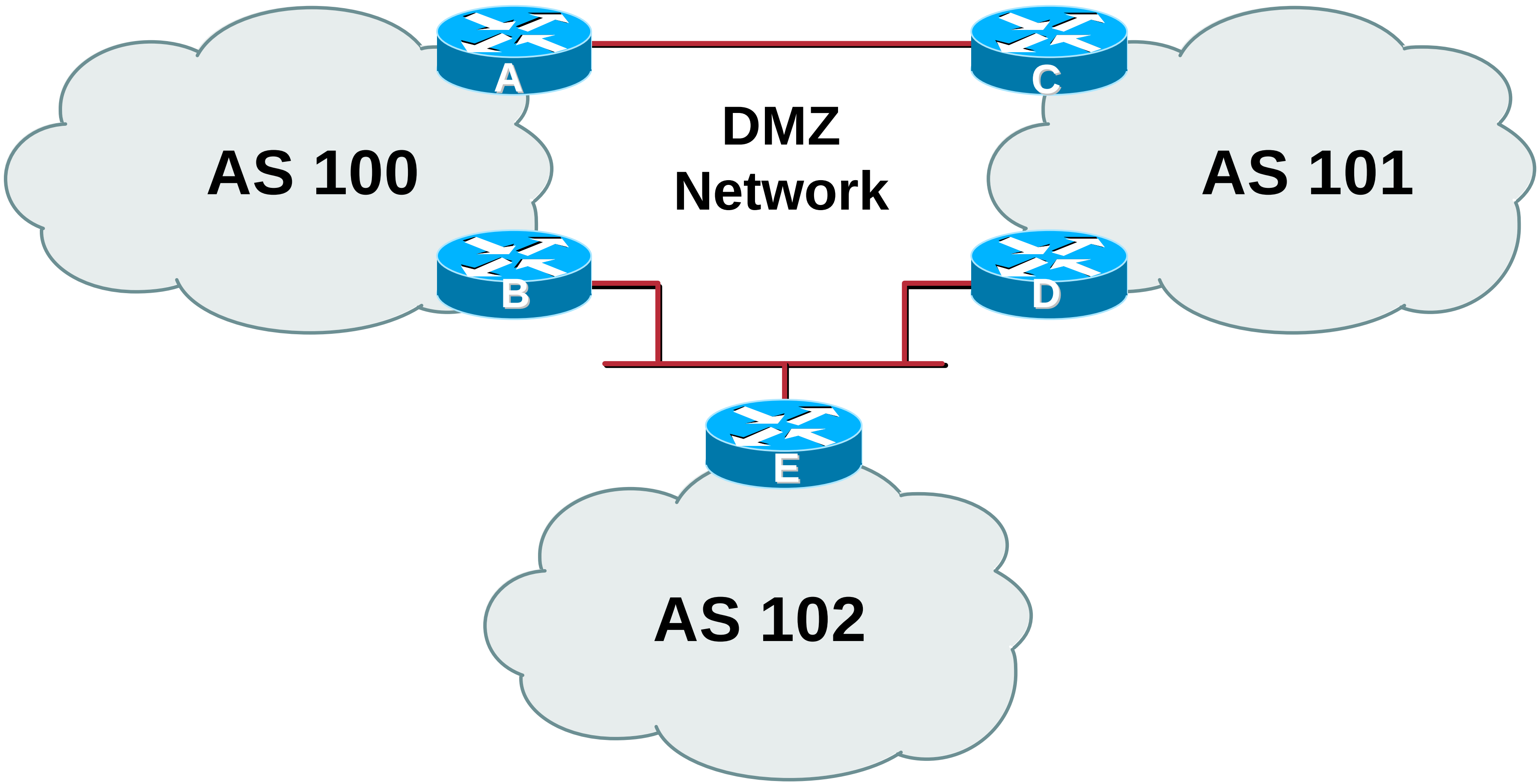
BGP Basics



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

Demarcation Zone (DMZ)

Cisco.com



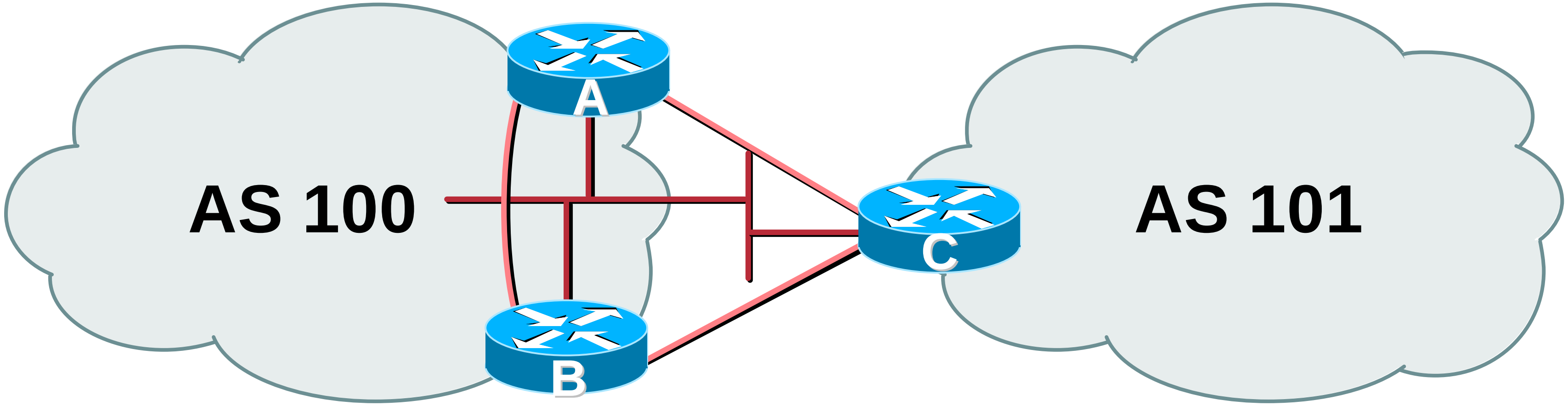
- **Shared network between ASes**

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)

Cisco.com



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

Configuring External BGP

Cisco.com

ip address on ethernet interface

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
!
```

Local ASN

Remote ASN

ip address of Router C ethernet interface

Inbound and outbound filters

Configuring External BGP

ip address on ethernet interface

Router C in AS101

```
interface ethernet 1/0/0
 ip address 222.222.10.1 255.255.255.240
!
router bgp 101
 network 220.220.8.0 mask 255.255.252.0
 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
!
```

Local ASN

Remote ASN

ip address of Router A ethernet interface

Inbound and outbound filters

Internal BGP (iBGP)

- **BGP peer within the same AS**
- **Not required to be directly connected**

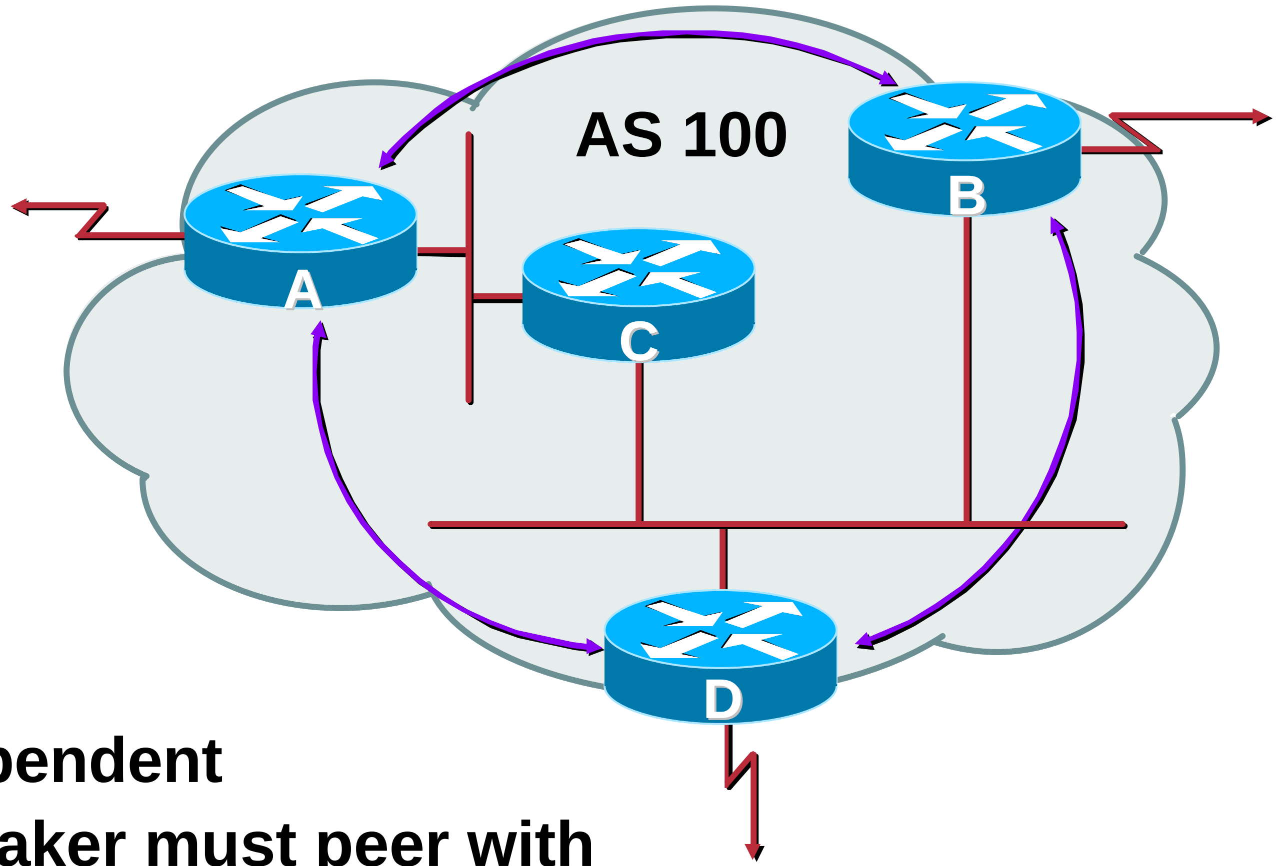
  IGP takes care of inter-BGP speaker connectivity

- **iBGP speakers need to be fully meshed**

  they originate connected networks

  they do not pass on prefixes learned from other iBGP speakers

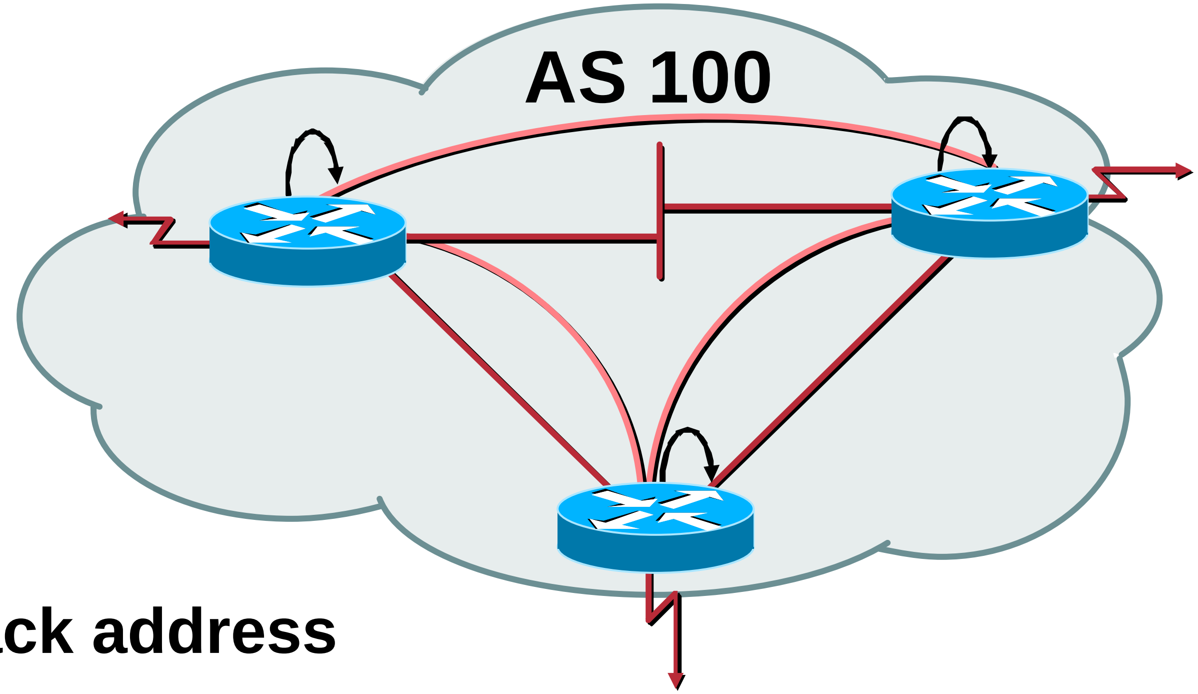
Internal BGP Peering (iBGP)



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

Peering to Loop-back Address

Cisco.com



- **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

Router A in AS100

```
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
!
```

ip address on loopback interface

Local ASN

Local ASN

ip address of Router B loopback interface

Configuring Internal BGP

Router B in AS100

```
interface loopback 0
 ip address 215.10.7.2 255.255.255.255
!
router bgp 100
 network 220.220.1.0
 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
!
```

ip address on loopback interface

Local ASN

Local ASN

ip address of Router A loopback interface

BGP for Internet Service Providers

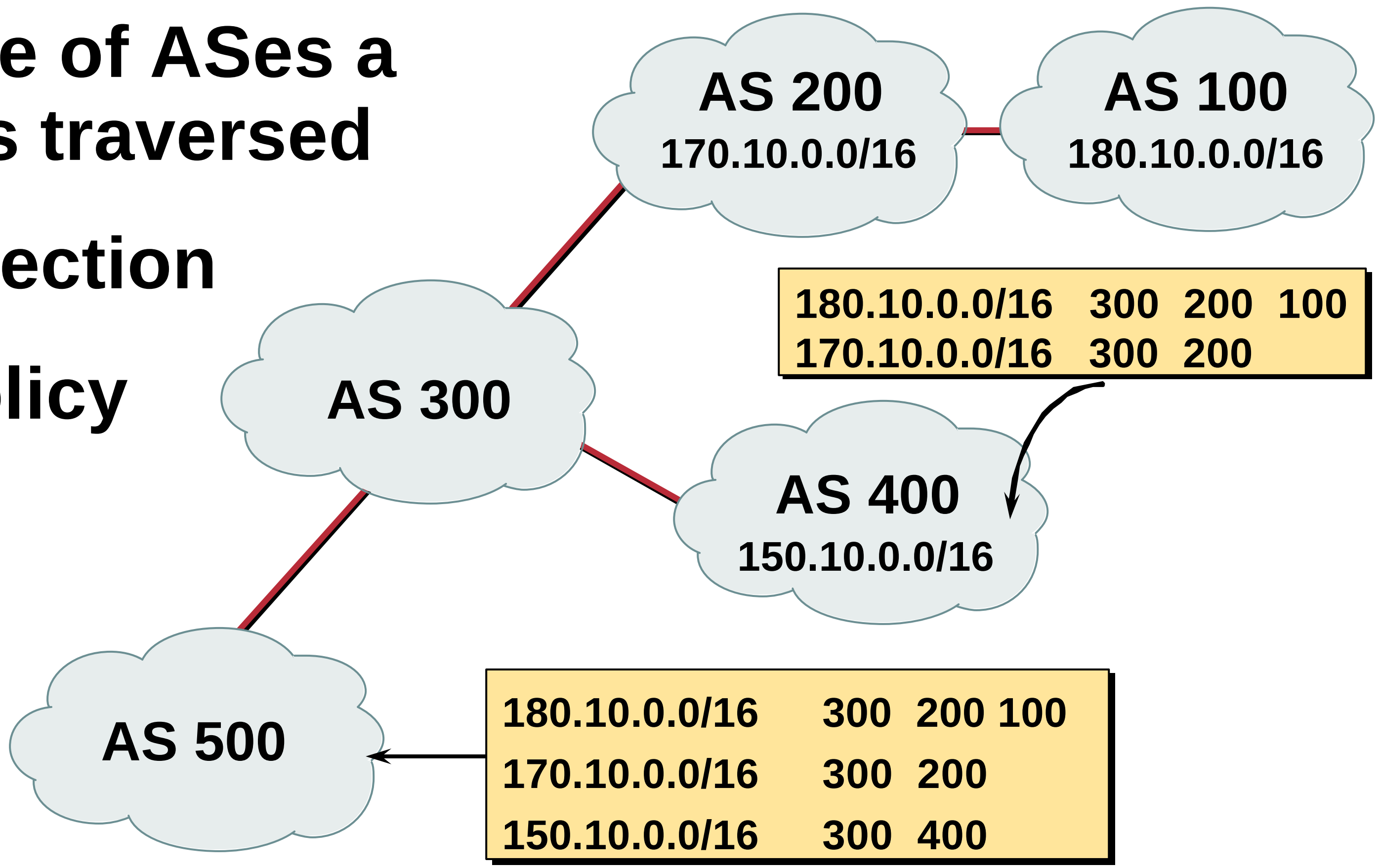
- Routing Basics
- BGP Basics
- BGP Attributes
- BGP Path Selection
- BGP Policy
- BGP Capabilities
- Scaling BGP

BGP Attributes

Recap

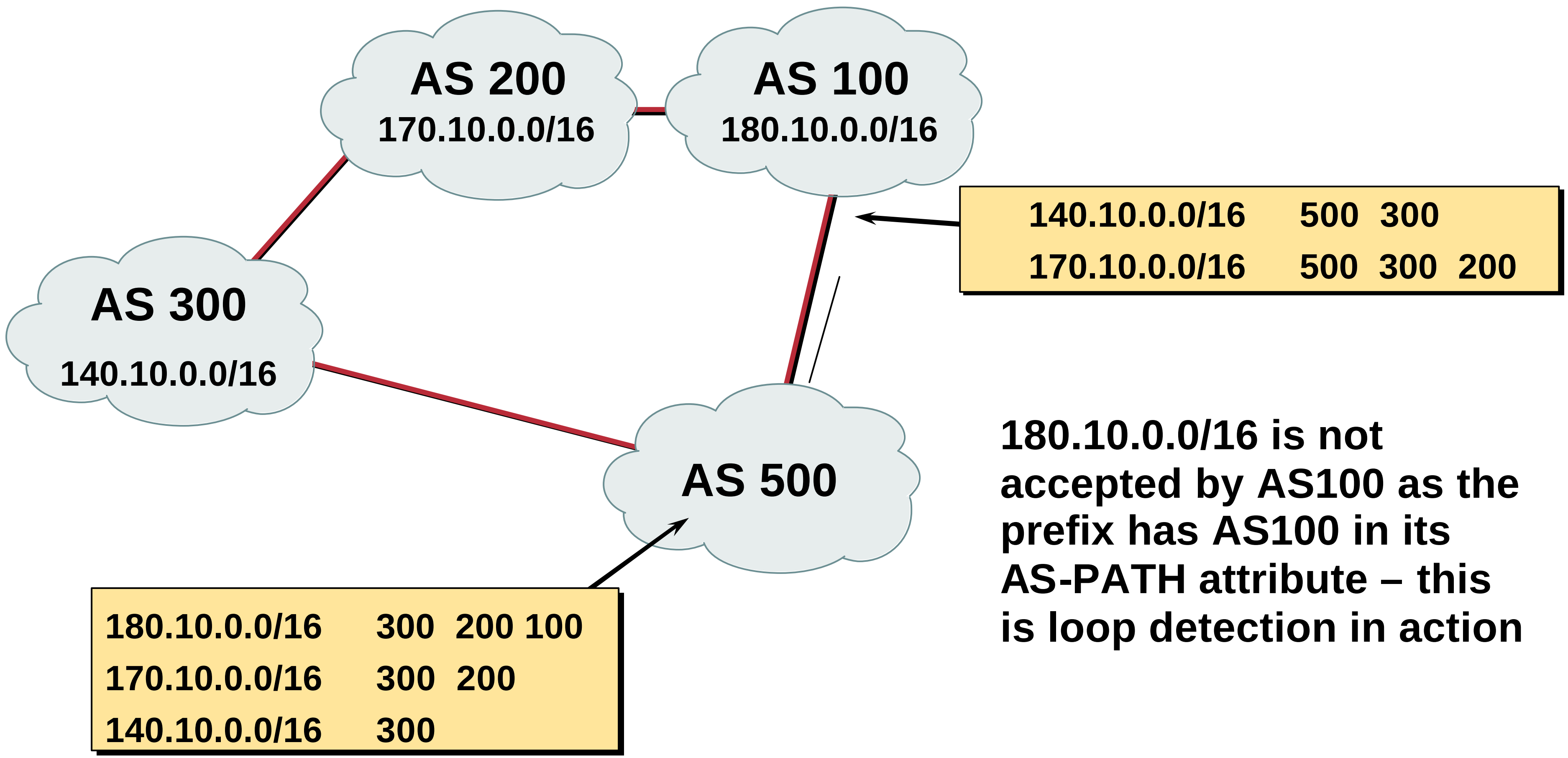
AS-Path

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



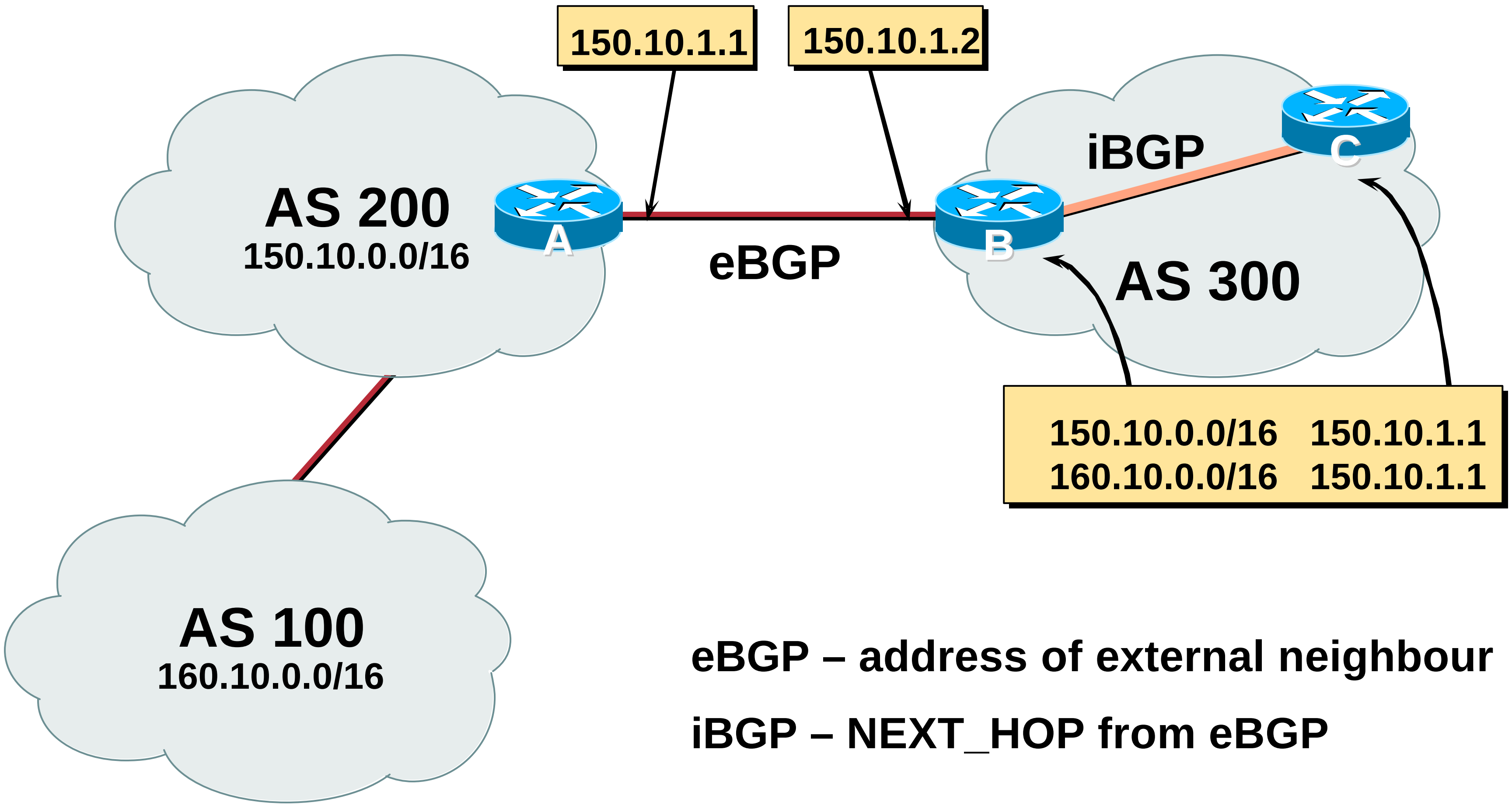
AS-Path loop detection

Cisco.com



180.10.0.0/16 is not accepted by AS100 as the prefix has AS100 in its AS-PATH attribute – this is loop detection in action

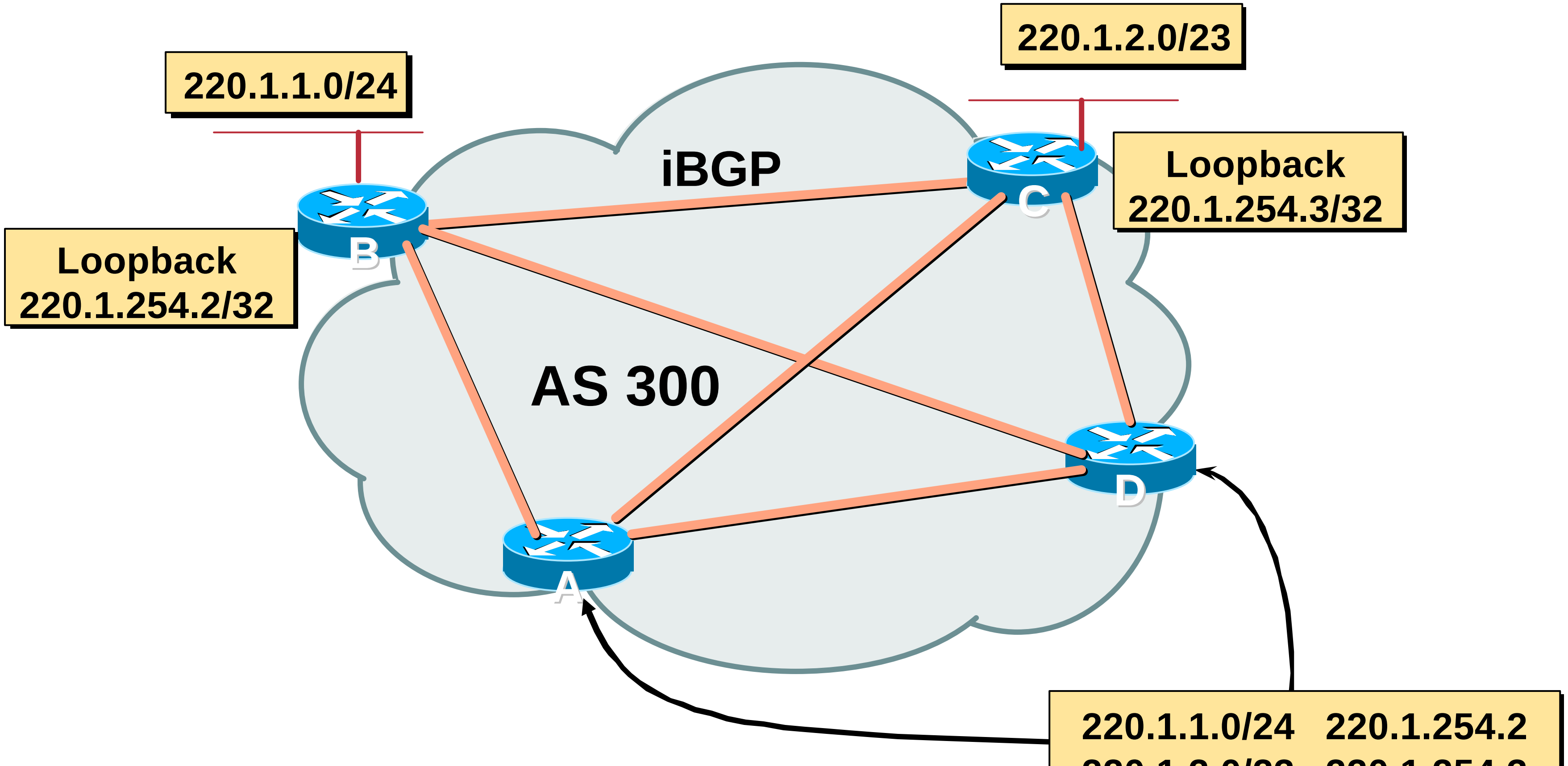
Next Hop



eBGP – address of external neighbour

iBGP – NEXT_HOP from eBGP

iBGP Next Hop



Next hop is ibgp router loopback address

Recursive route look-up

Next Hop (summary)

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

Origin

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

  IGP – generated by BGP network statement

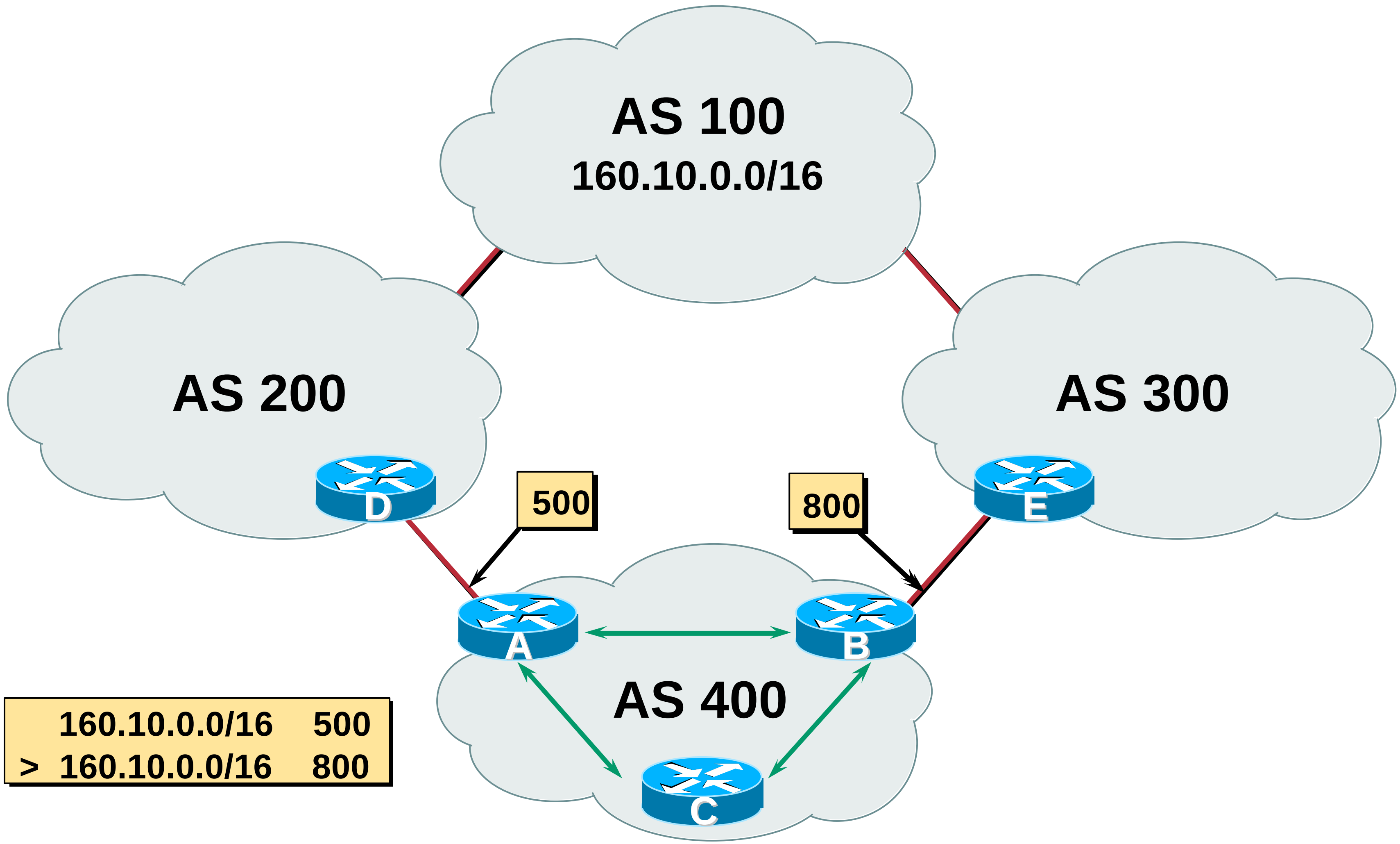
  EGP – generated by EGP

  incomplete – redistributed from another routing protocol

Aggregator

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

Local Preference



Local Preference

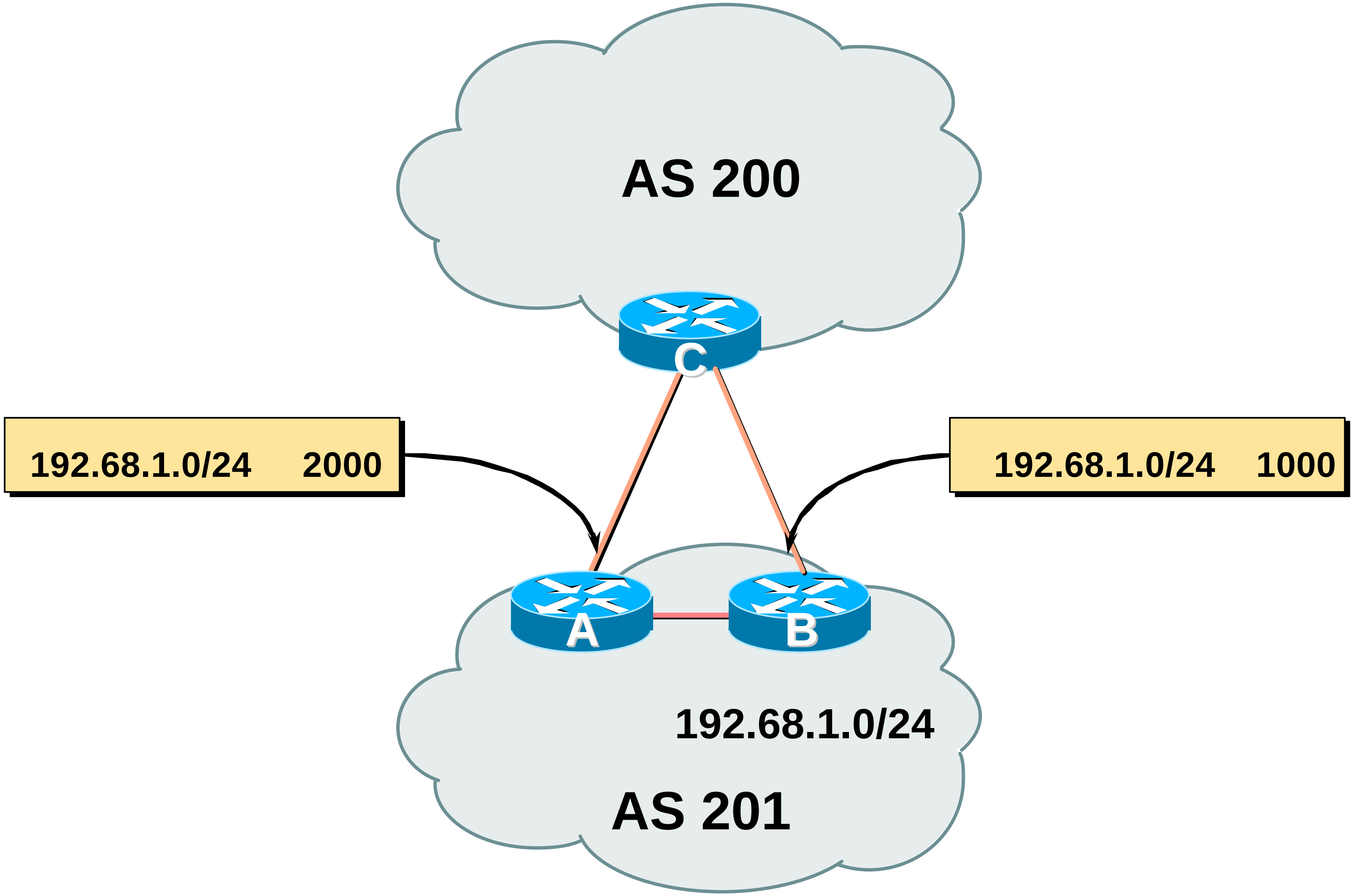
- **Local to an AS – non-transitive**
  - **Default local preference is 100 (IOS)**
- **Used to influence BGP path selection**
  - **determines best path for *outbound* traffic**
- **Path with highest local preference wins**

Local Preference

- **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)



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

- **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

- Not really an attribute – local to router

  Allows policy control, similar to local preference

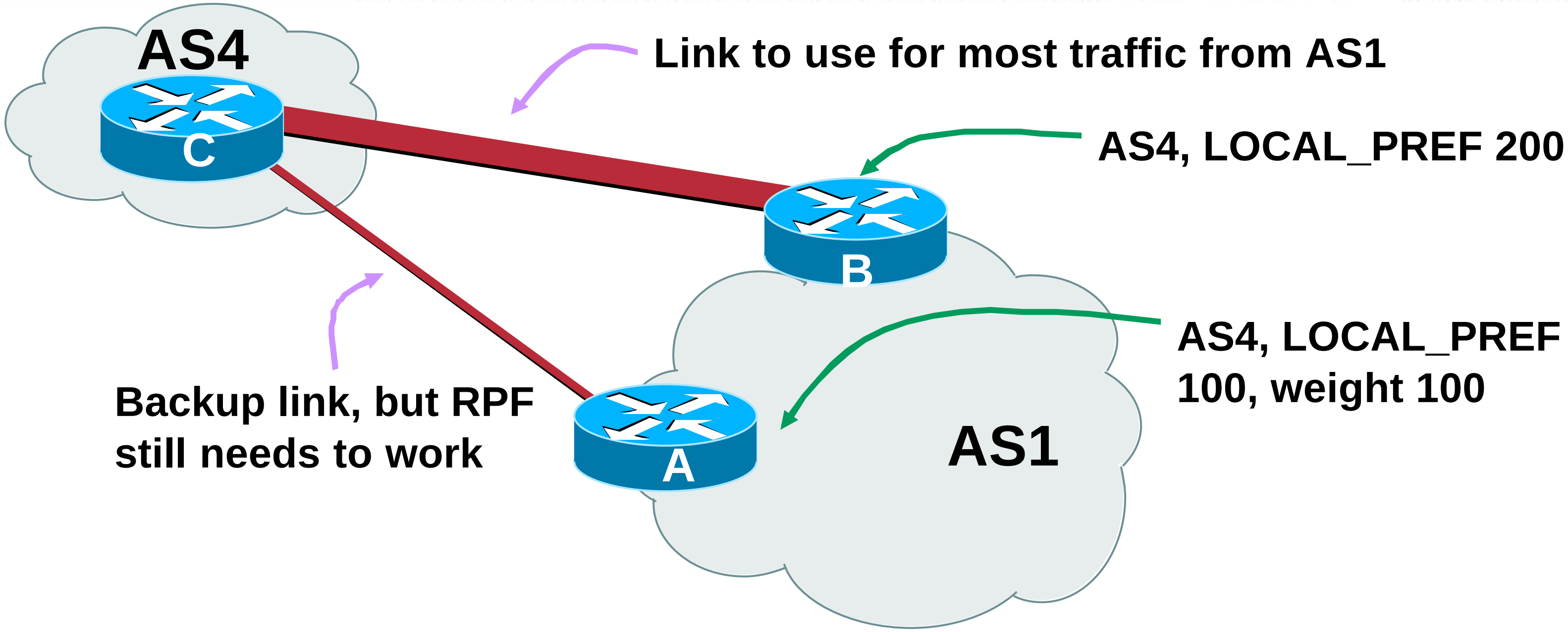
- Highest weight wins
- Applied to all routes from a neighbour

```
neighbor 220.5.7.1 weight 100
```

- Weight assigned to routes based on filter

```
neighbor 220.5.7.3 filter-list 3 weight 50
```

Weight – Used to help Deploy RPF



- **Best path to AS4 from AS1 is always via B due to local-pref**
- **But packets arriving at A from AS4 over the direct C to A link will pass the RPF check as that path has a priority due to the weight being set**

  If weight was not set, best path would be via B, and the RPF check would fail

Community

- Communities are described in RFC1997
- 32 bit integer

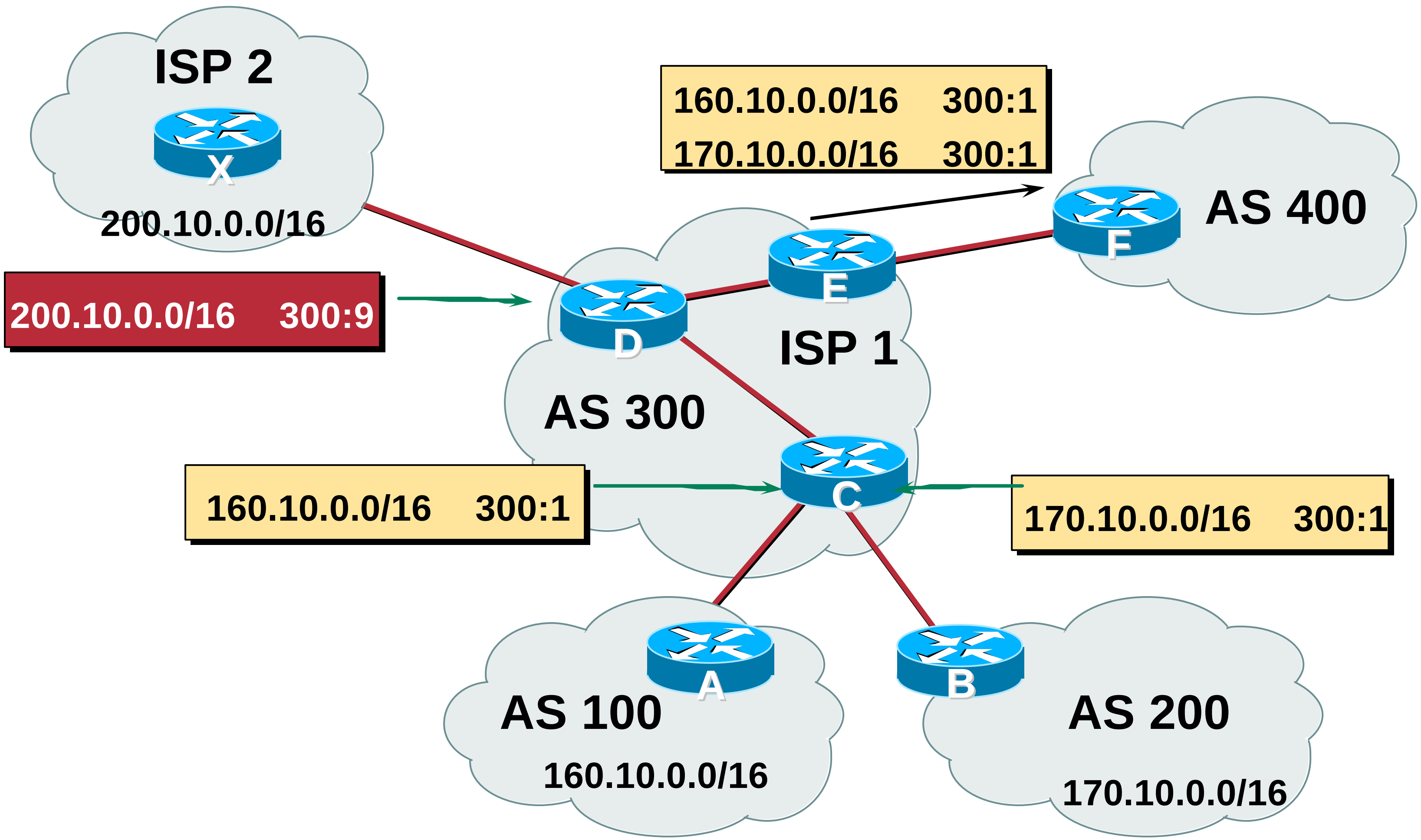
  Represented as two 16 bit integers (RFC1998)

- Used to group destinations

  Each destination could be member of multiple communities

- Community attribute carried across AS's
- Very useful in applying policies

Community


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

Well-Known Communities

- **no-export**

  do not advertise to eBGP peers

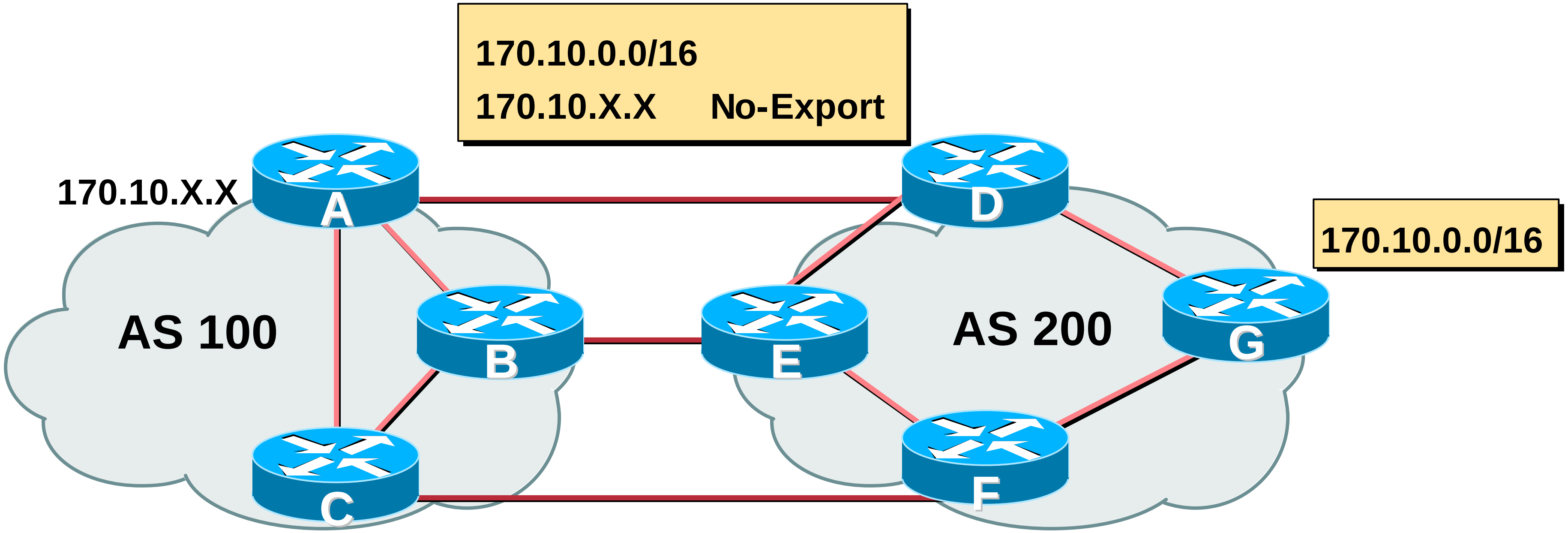
- **no-advertise**

  do not advertise to any peer

- **local-AS**

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

No-Export Community



- **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 for Internet Service Providers

- Routing Basics
- BGP Basics
- BGP Attributes
- BGP Path Selection
- BGP Policy
- BGP Capabilities
- Scaling BGP

BGP Path Selection Algorithm

Why Is This the Best Path?

BGP Path Selection Algorithm
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
Part Two

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
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 for Internet Service Providers

Cisco.com

- Routing Basics
- BGP Basics
- BGP Attributes
- BGP Path Selection
- BGP Policy
- BGP Capabilities
- Scaling BGP

Applying Policy with BGP

Control!

Applying Policy with BGP

- **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

- **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

- **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

Cisco.com

- **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
Regular Expressions

Cisco.com

- **Not so simple Examples**

^[0-9]+$	Match AS_PATH length of one
^[0-9]+_[0-9]+$	Match AS_PATH length of two
^[0-9]*_[0-9]+$	Match AS_PATH length of one or two
^[0-9]*_[0-9]*$	Match AS_PATH length of one or two (will also match zero)
^[0-9]+_[0-9]+_[0-9]+$	Match AS_PATH length of three
_(701|1800)_	Match anything which has gone through AS701 or AS1800
_1849(_.+_)12163$	Match anything of origin AS12163 and passed through AS1849

Policy Control
Route Maps

- **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

- **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

- 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 for Internet Service Providers

- Routing Basics
- BGP Basics
- BGP Attributes
- BGP Path Selection
- BGP Policy
- BGP Capabilities
- Scaling BGP

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
- 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

Cisco.com

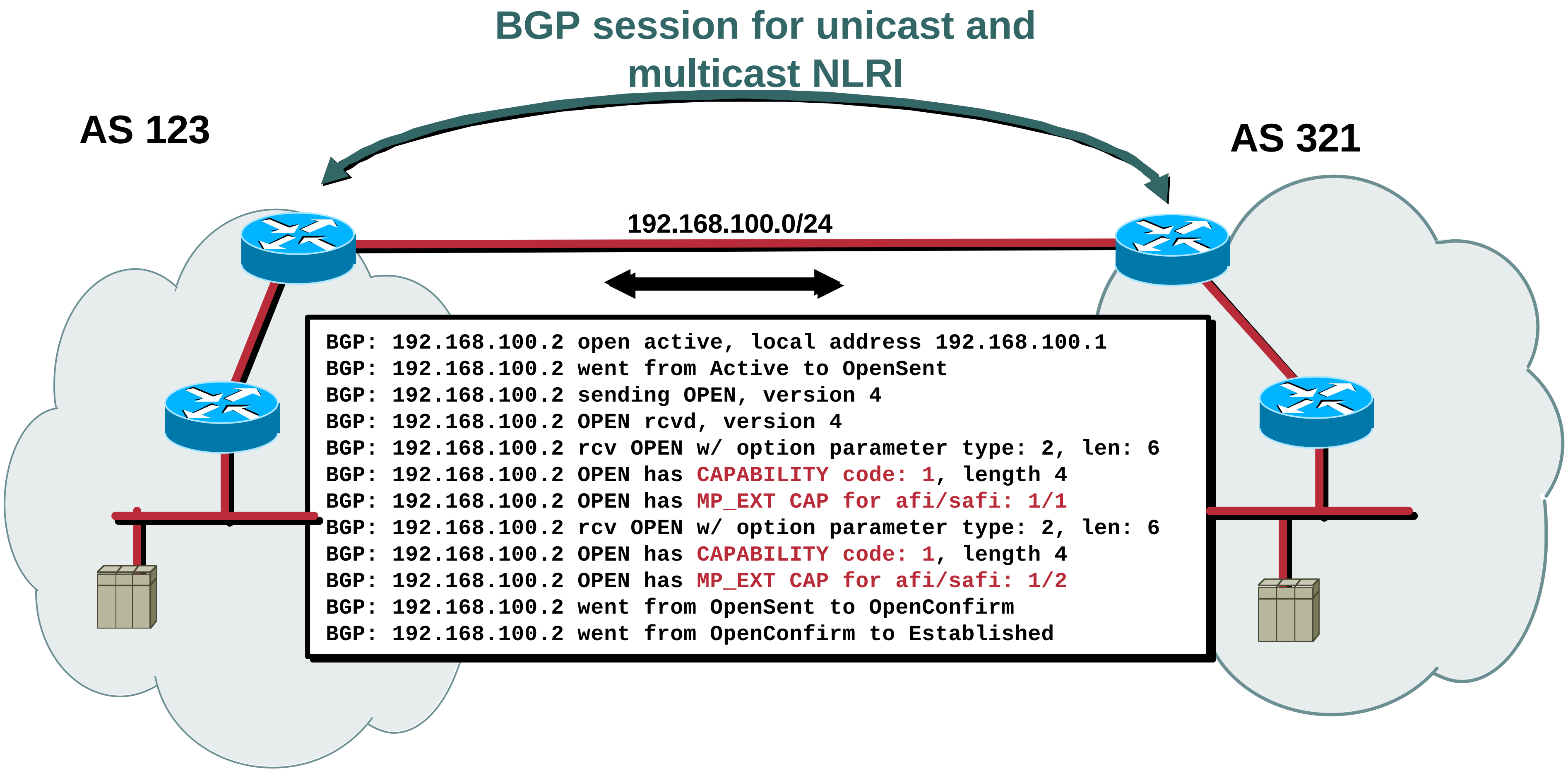
Current capabilities are:

```
 0   Reserved                                  [RFC3392]
 1   Multiprotocol Extensions for BGP-4        [RFC2858]
 2   Route Refresh Capability for BGP-4        [RFC2918]
 3   Cooperative Route Filtering Capability    []
 4   Multiple routes to a destination capability [RFC3107]
64   Graceful Restart Capability               []
65   Support for 4 octet ASNs                  []
66   Support for Dynamic Capability            []
```

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

BGP Capabilities Negotiation

Cisco.com




```
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
```

- Routing Basics
- BGP Basics
- BGP Attributes
- BGP Path Selection
- BGP Policy
- BGP Capabilities
- Scaling BGP

BGP Scaling Techniques

BGP Scaling Techniques

Cisco.com

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

BGP Scaling Techniques

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

Cisco.com

Route Refresh

Route Refresh

Problem:

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

  Tears down BGP peering

  Consumes CPU

  Severely disrupts connectivity for all networks

Solution:

- **Route Refresh**

Route Refresh Capability

Cisco.com

- Facilitates non-disruptive policy changes
- No configuration is needed

  Automatically negotiated at peer establishment
- 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

Cisco.com

- **Router normally stores prefixes which have been received from peer after policy application**
  - **Enabling soft-reconfiguration means router also stores prefixes/attributes 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**
- **Also advantageous when operator requires to know which prefixes have been sent to a router prior to the application of any inbound policy**

Configuring Soft Reconfiguration


```
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

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


```
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
```

Peer Groups

- **Always configure peer-groups for iBGP**
  - Even if there are only a few iBGP peers
  - Easier to scale network in the future
  - Makes template configuration much easier
- **Consider using peer-groups for eBGP**
  - Especially useful for multiple BGP customers using same AS (RFC2270)
  - Also useful at Exchange Points where ISP policy is generally the same to each peer

Route Flap Damping

Stabilising the Network

Route Flap Damping

Cisco.com

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

Route Flap Damping (continued)

- **Requirements**

  Fast convergence for normal route changes

  History predicts future behaviour

  Suppress oscillating routes

  Advertise stable routes

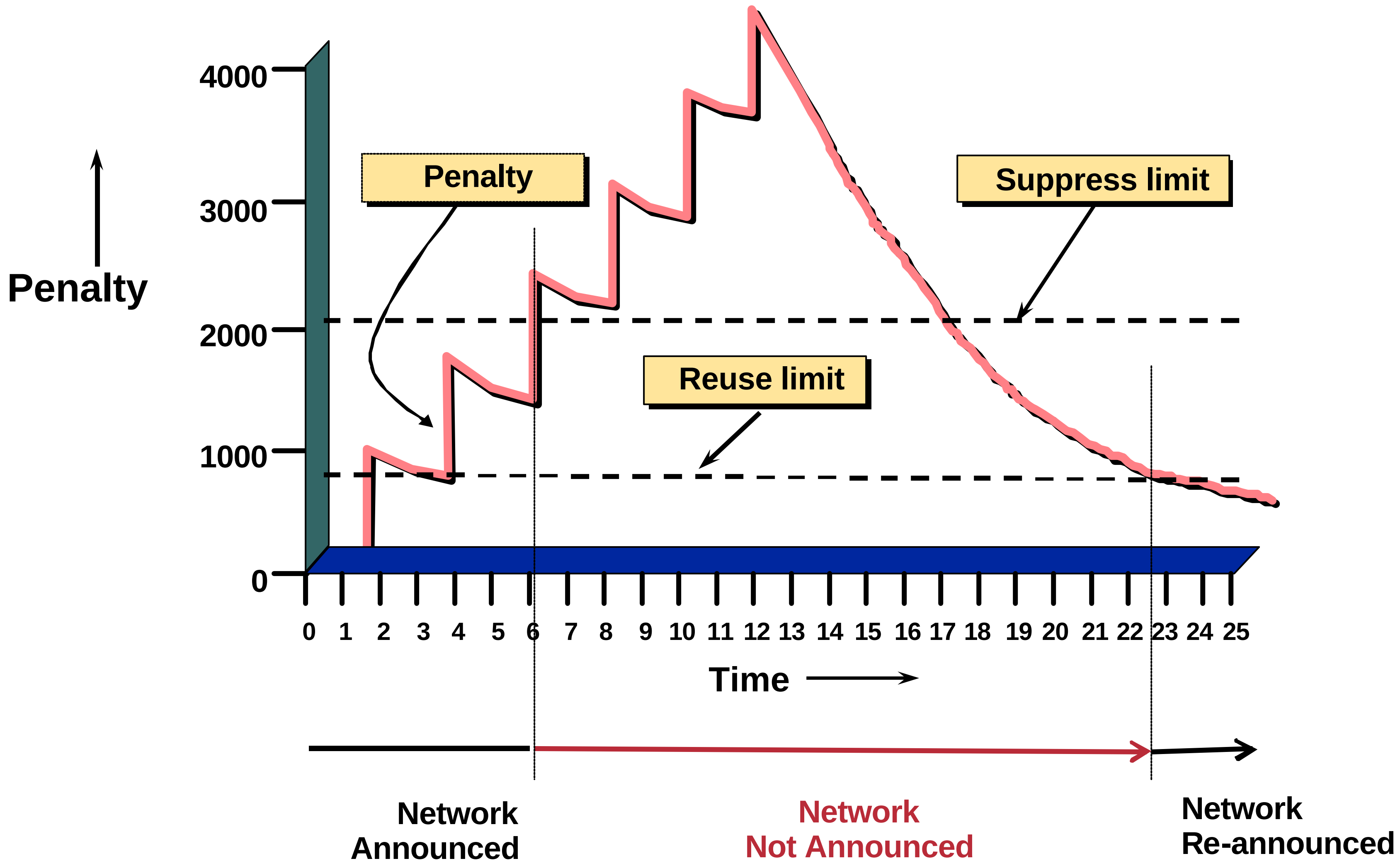
- **Documented in RFC2439**

Operation

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

Operation

- **Only applied to inbound announcements from eBGP peers**
- **Alternate paths still usable**
- **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**

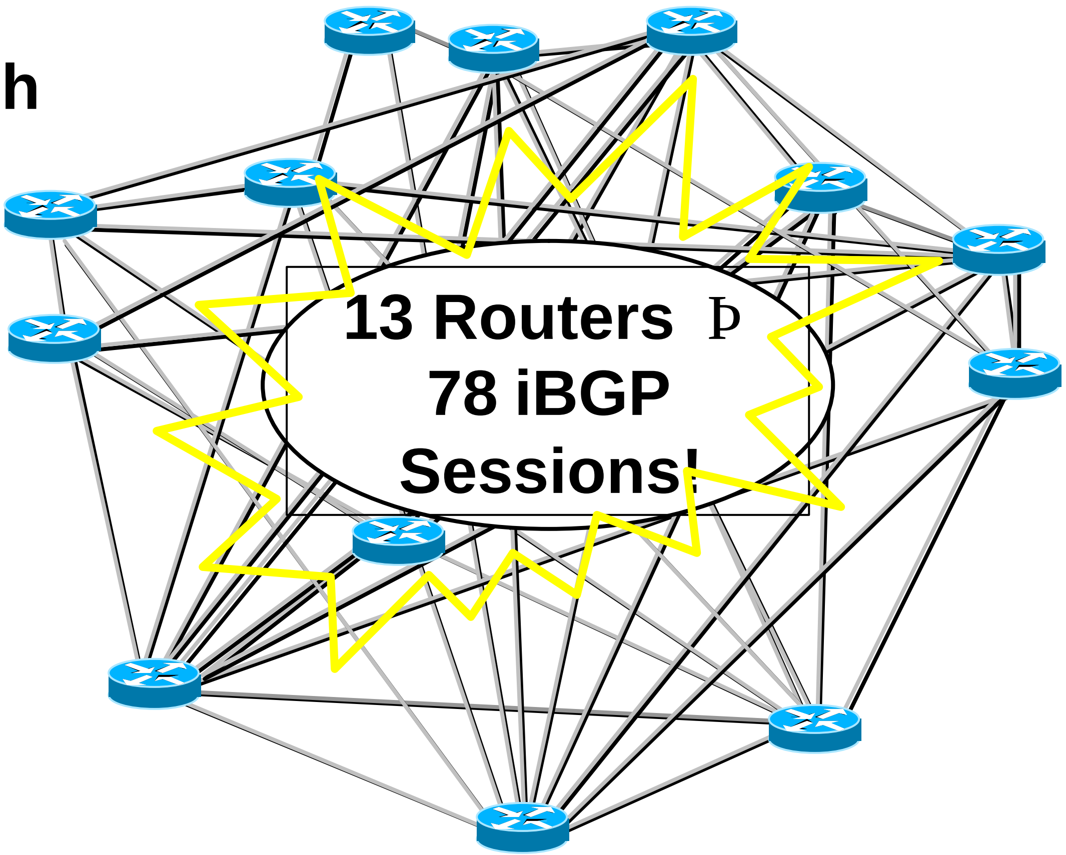
Route Reflectors and Confederations

Scaling iBGP mesh

Cisco.com

Avoid ½n(n-1) iBGP mesh

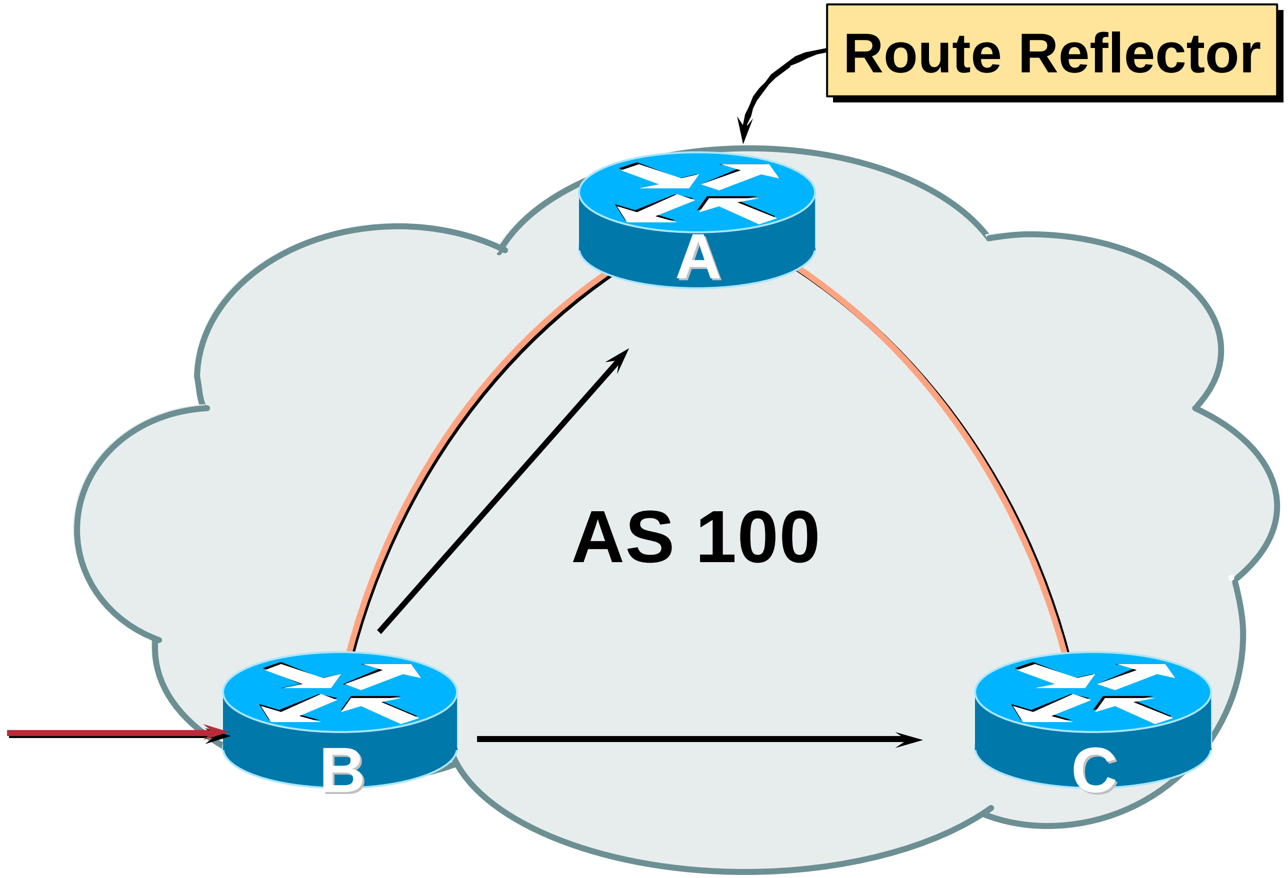
n=1000 Þ nearly half a million ibgp sessions!



Two solutions

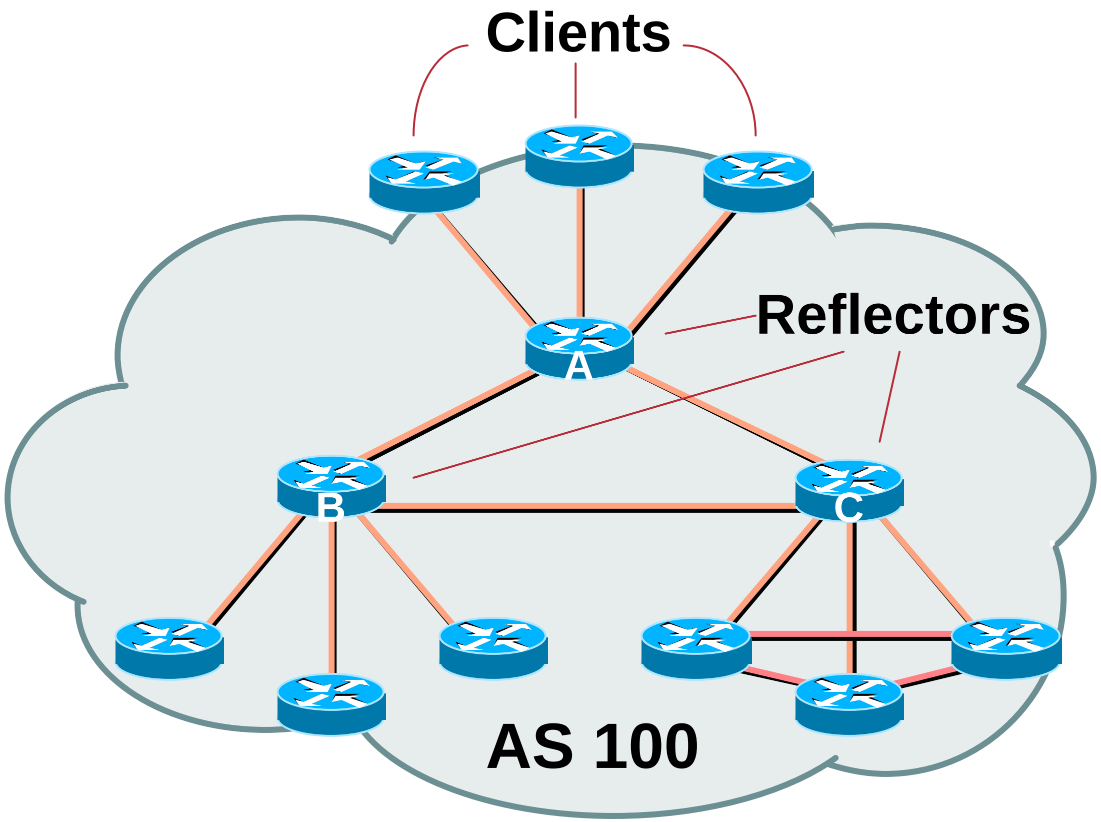
- Route reflector – simpler to deploy and run
- Confederation – more complex, corner case benefits

Route Reflector: Principle



Route Reflector

- Reflector receives path from
  clients and non-clients
- Selects best path
- If best path is from client, reflect to other clients and non-clients
- If best path is from non-client, reflect to clients only
- Non-meshed clients
- Described in RFC2796



Route Reflector Topology

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

Route Reflectors: Loop Avoidance

- **Originator_ID attribute**

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

- **Cluster_list attribute**

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

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

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

Route Reflectors: Redundancy

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

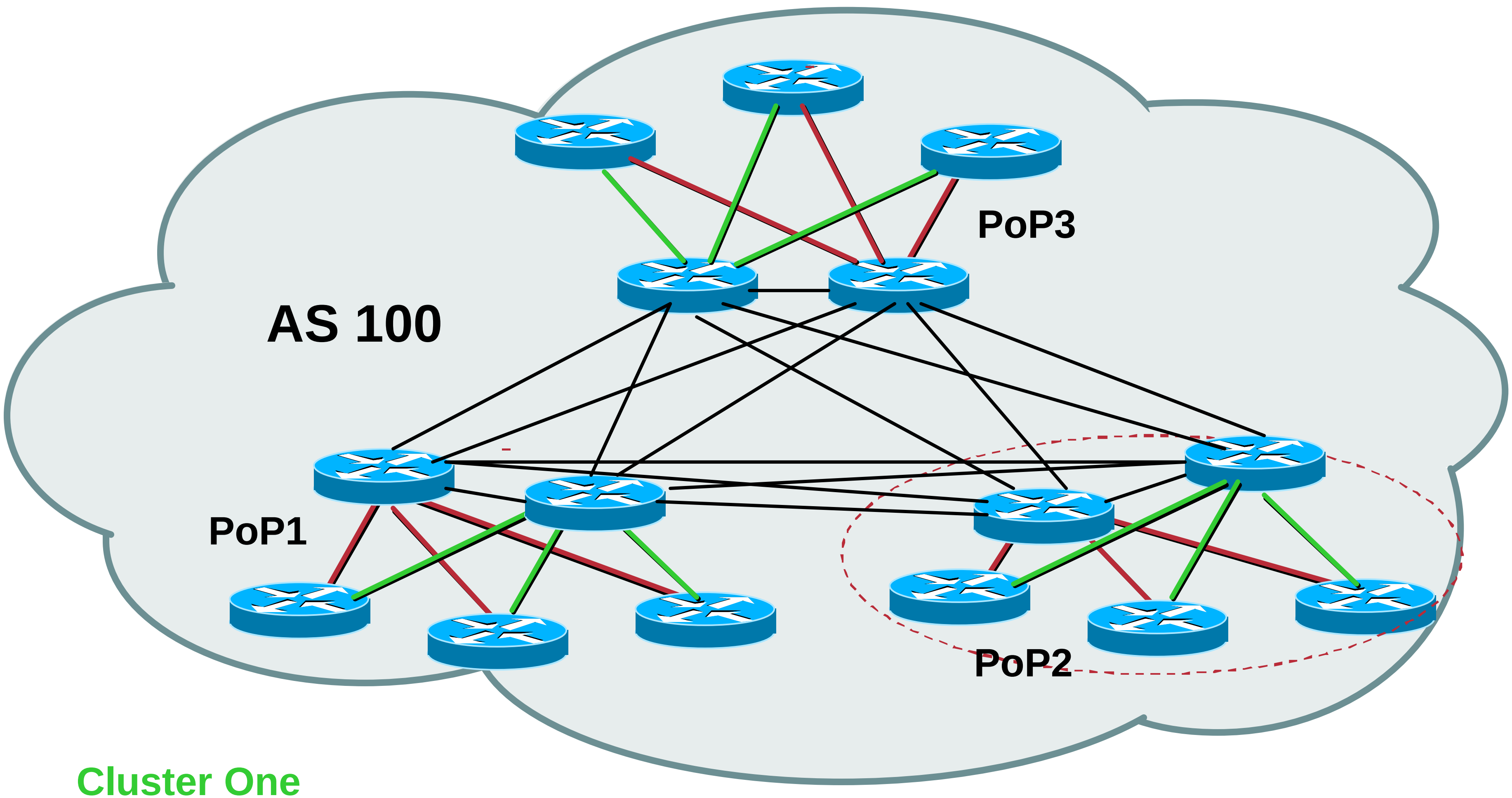
  All RRs in the cluster must have the same cluster-id (otherwise it is a different cluster)

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

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

  ® Each client has two RRs = redundancy

Route Reflectors: Redundancy



Cluster One

Cluster Two

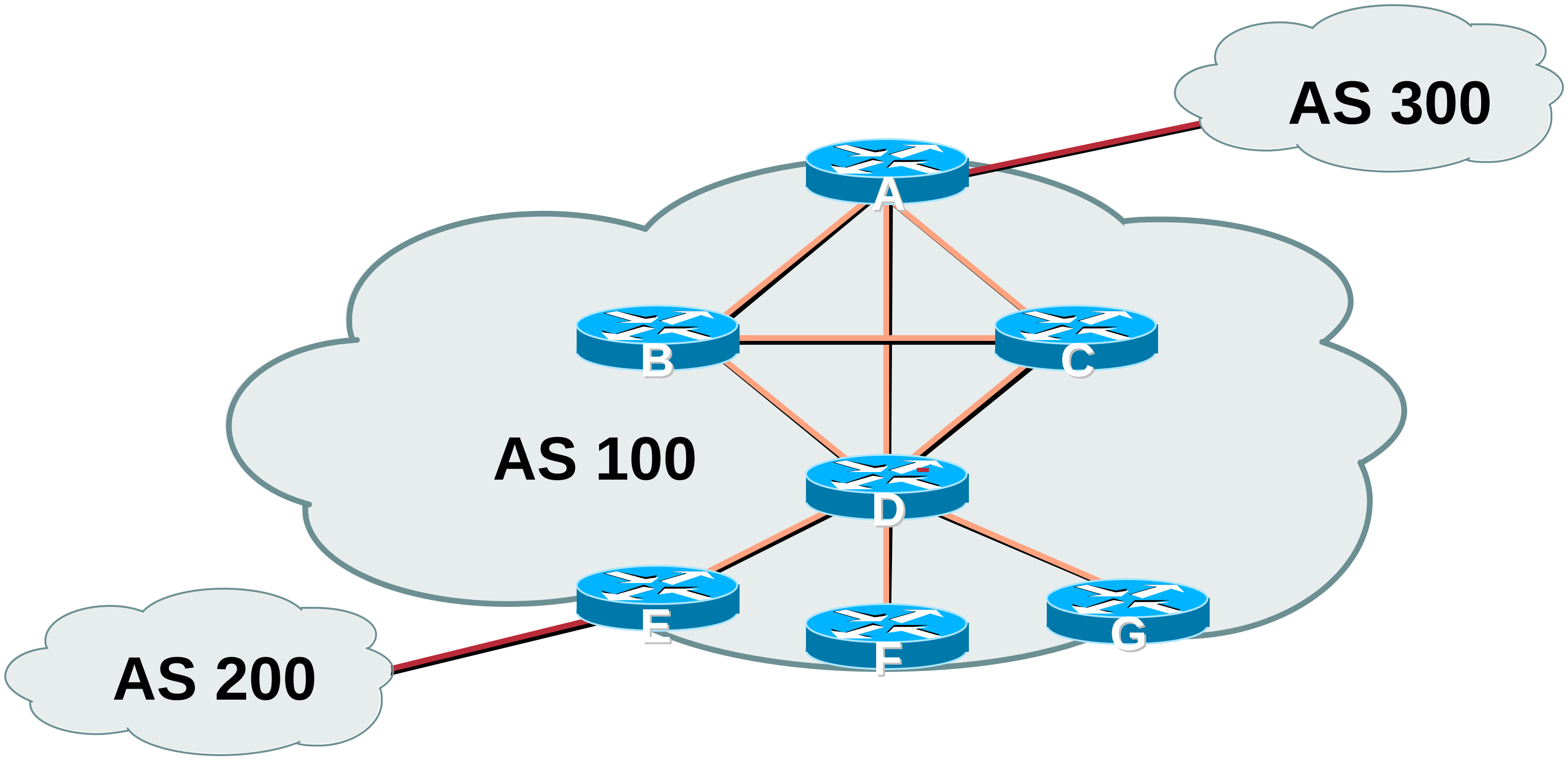
Route Reflectors: Migration

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

Route Reflectors: Migration

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

Route Reflector: Migration



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

Configuring a Route Reflector


```
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

- **Divide the AS into sub-ASes**

  eBGP between sub-ASes, but some iBGP information is kept

  Preserve NEXT_HOP across the sub-AS (IGP carries this information)

  Preserve LOCAL_PREF and MED

- **Usually a single IGP**
- **Described in RFC3065**

Confederations (Cont.)

Cisco.com

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

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

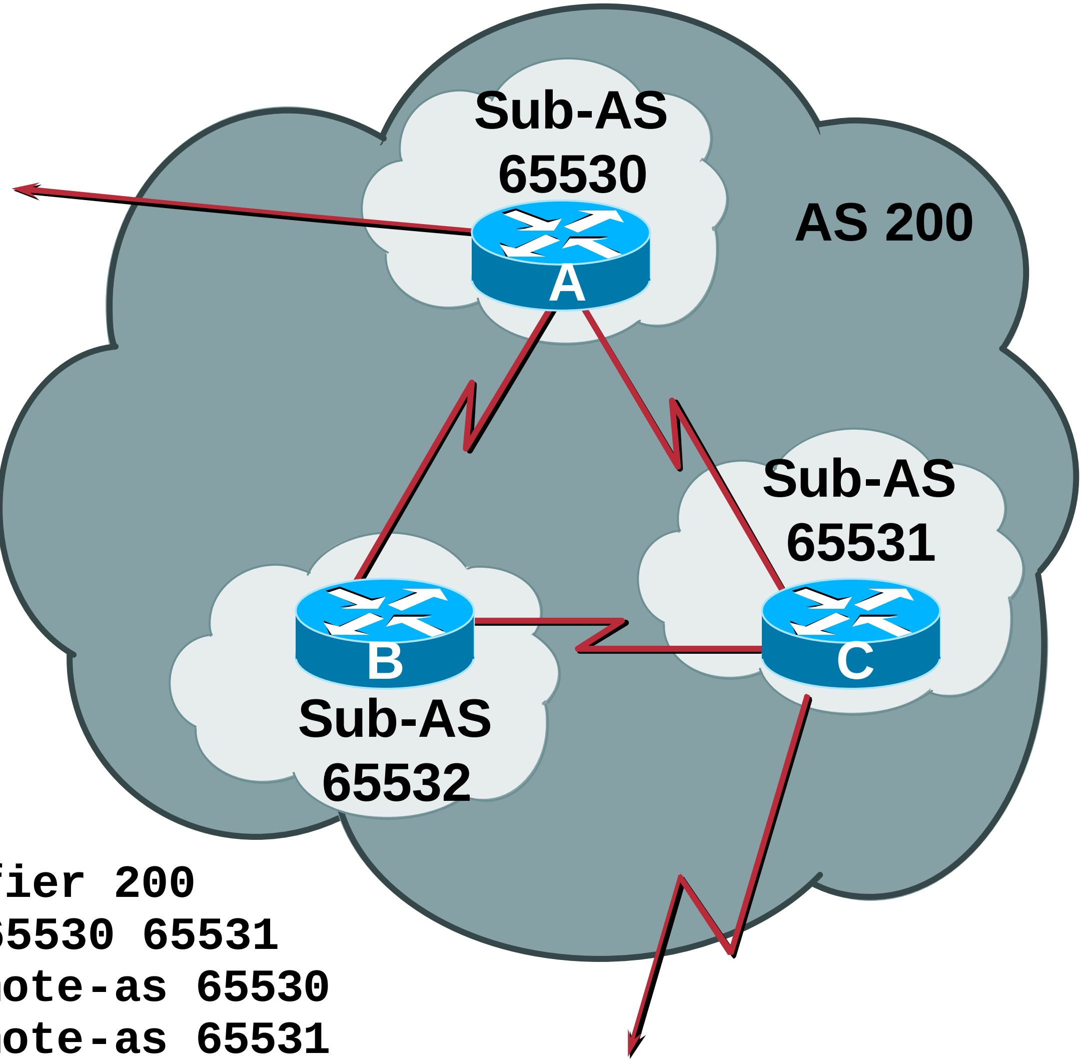
- **iBGP speakers in each sub-AS are fully meshed**

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

  Can also use Route-Reflector within sub-AS

Confederations (cont.)

Cisco.com

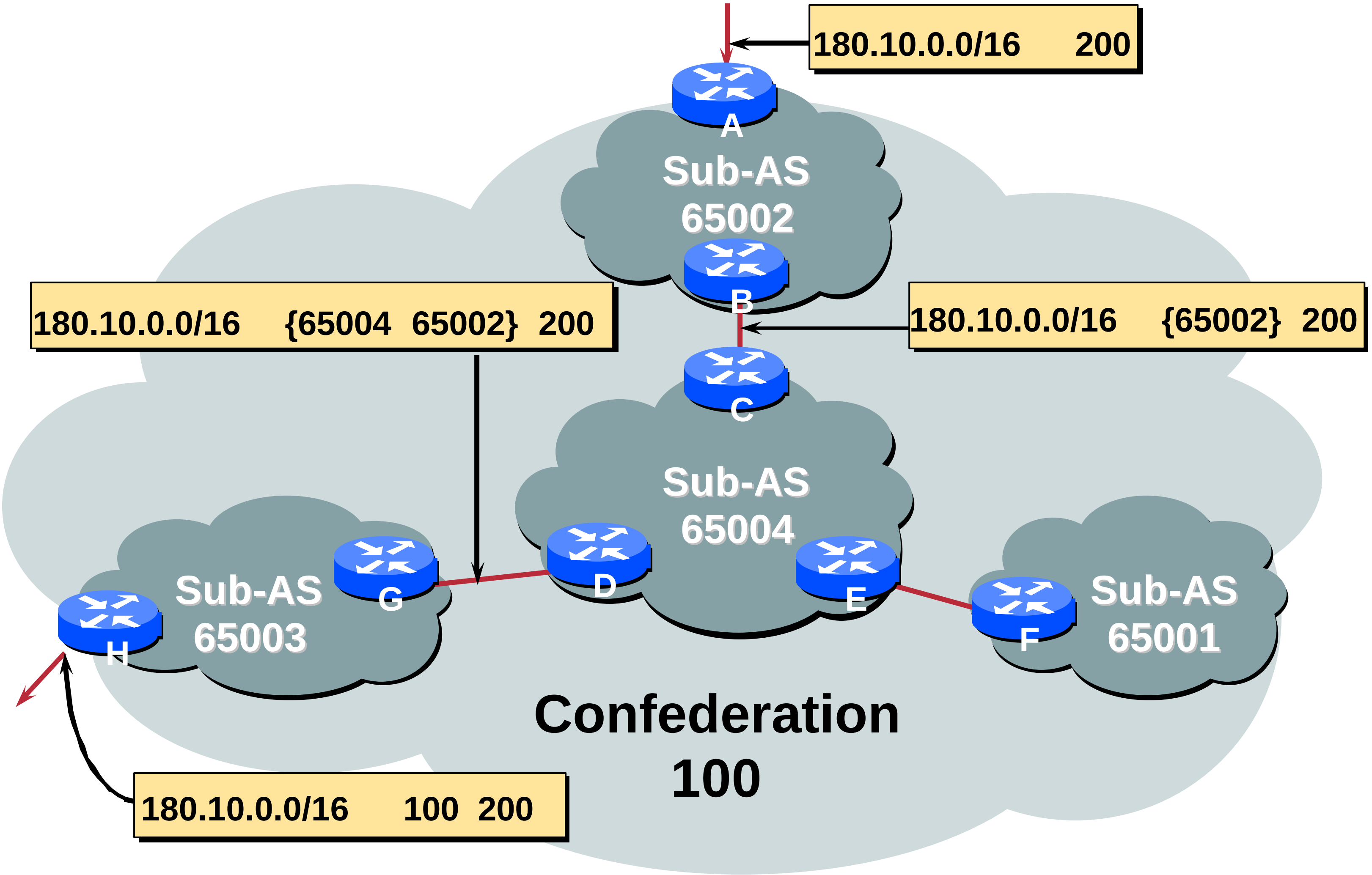


- **Configuration (rtr B):**

```
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
```

Confederations: AS-Sequence

Cisco.com



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**

Confederations (cont.)

- **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
```

Route Reflectors or Confederations?

Cisco.com

	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

- **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

- Routing Basics
- BGP Basics
- BGP Attributes
- BGP Path Selection
- BGP Policy
- BGP Capabilities
- Scaling BGP

BGP Tutorial

End of Part 1 – Introduction

Part 2 – Deployment Techniques is next