# BGP Attributes and Path Selection

## AFNOG 2013 AR-E Workshop

# BGP Attributes

The "tools" available for the job
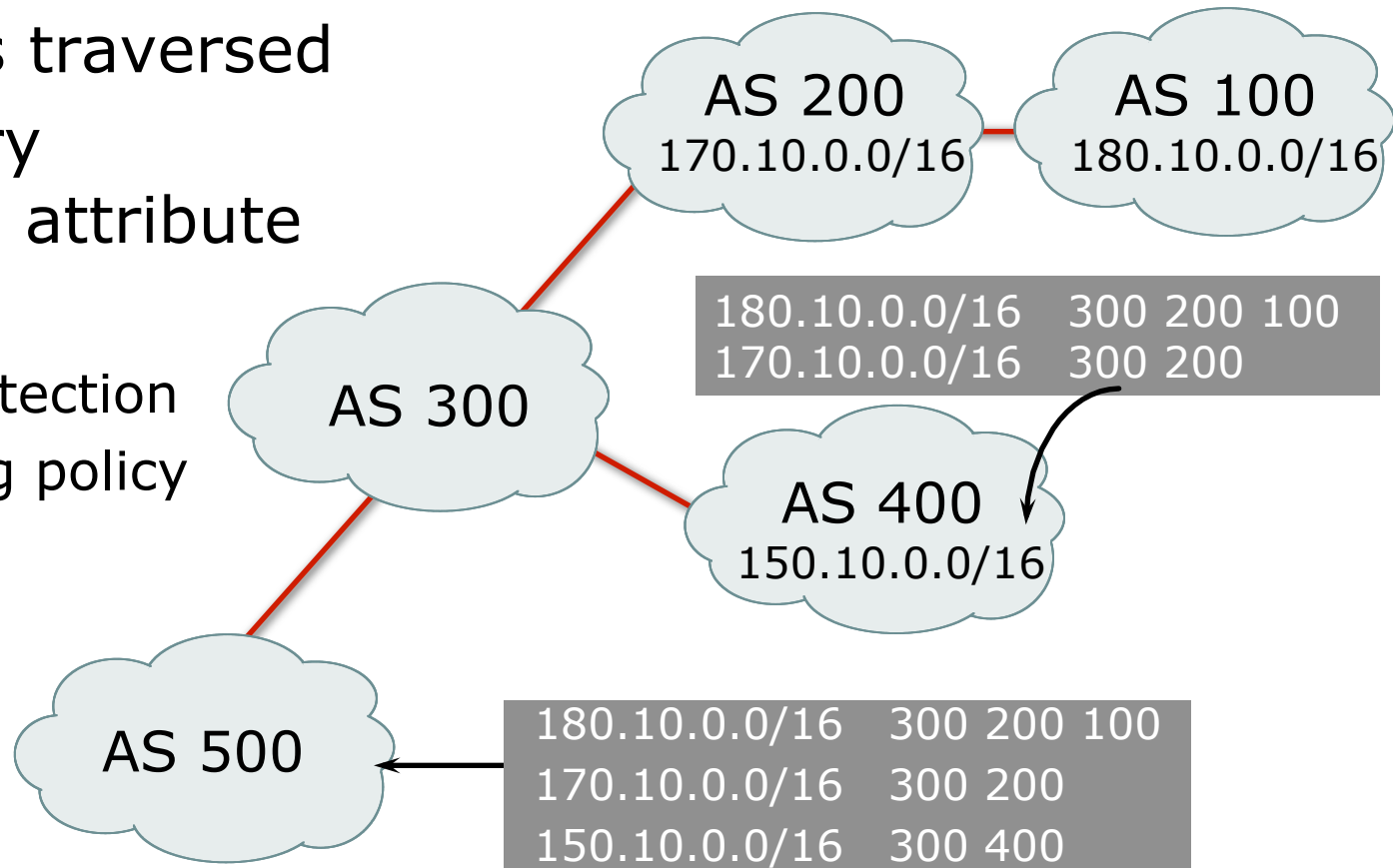
# What Is an Attribute?

| ... | Next Hop | AS Path | MED | ... | ... |
|-----|----------|---------|-----|-----|-----|

- Part of a BGP Update
- Describes the characteristics of prefix
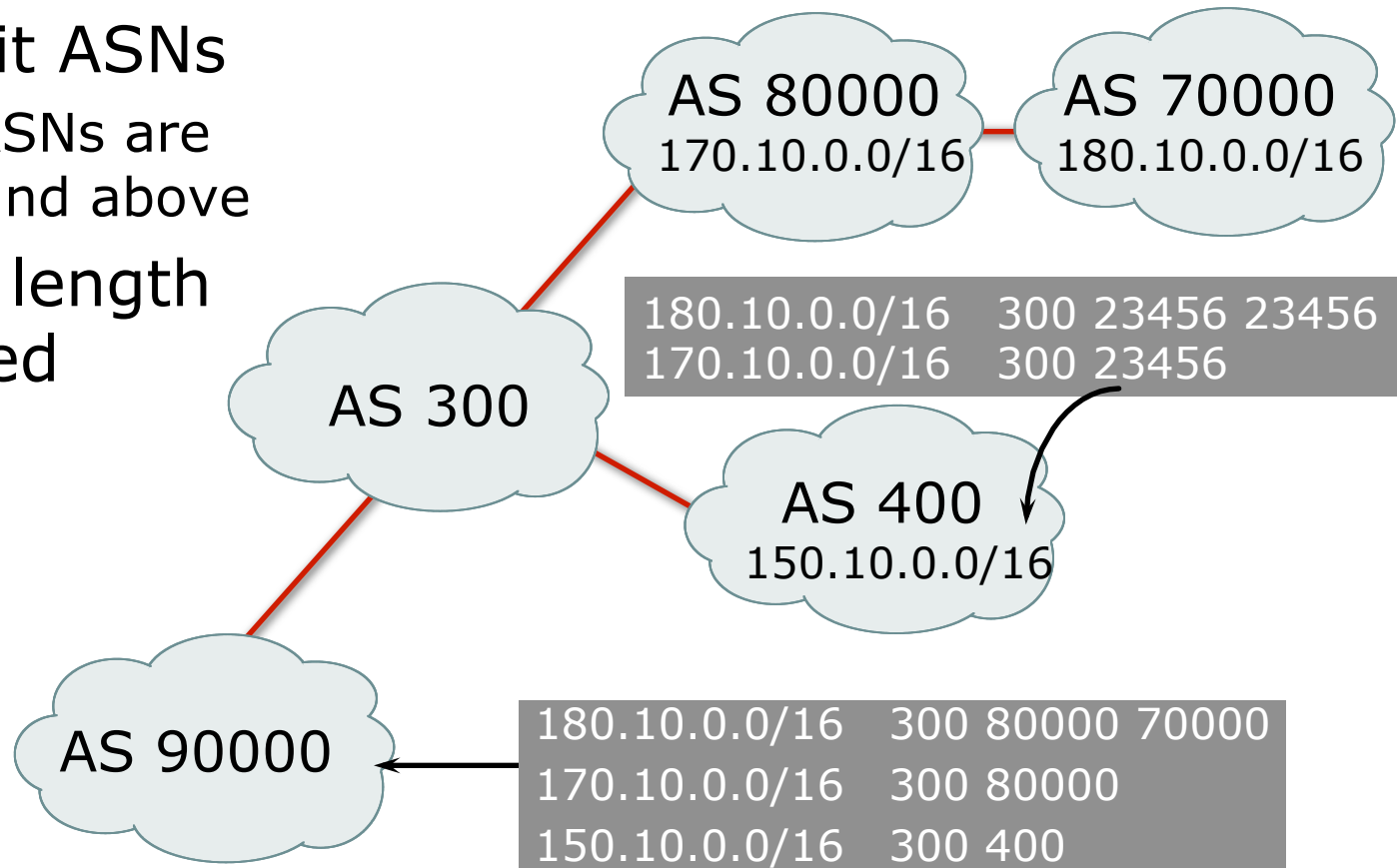- Can either be transitive or non-transitive
- Some are mandatory

# AS-Path

- Sequence of ASes a route has traversed
- Mandatory transitive attribute
- Used for:
  - Loop detection
  - Applying policy
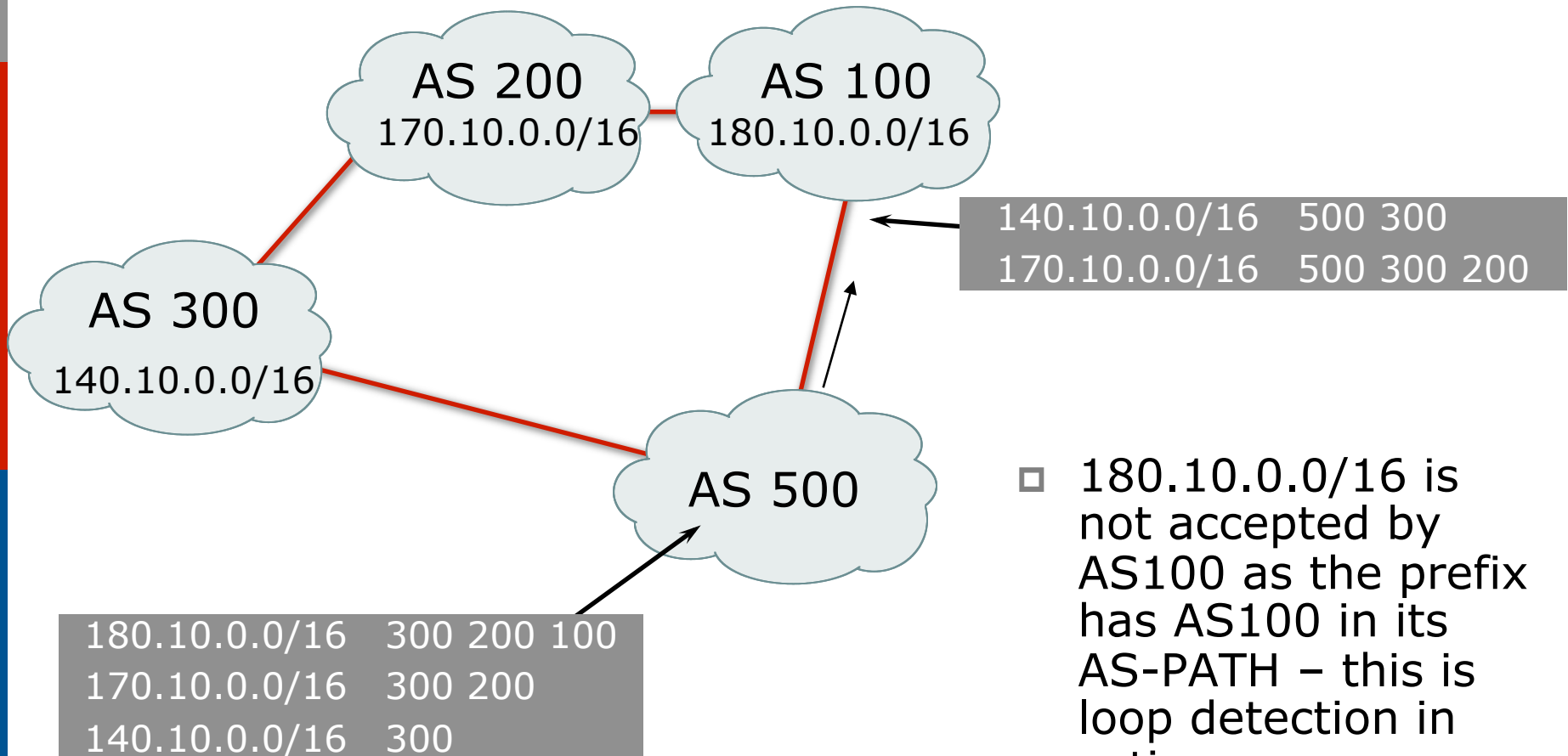
AS 200
170.10.0.0/16

AS 100
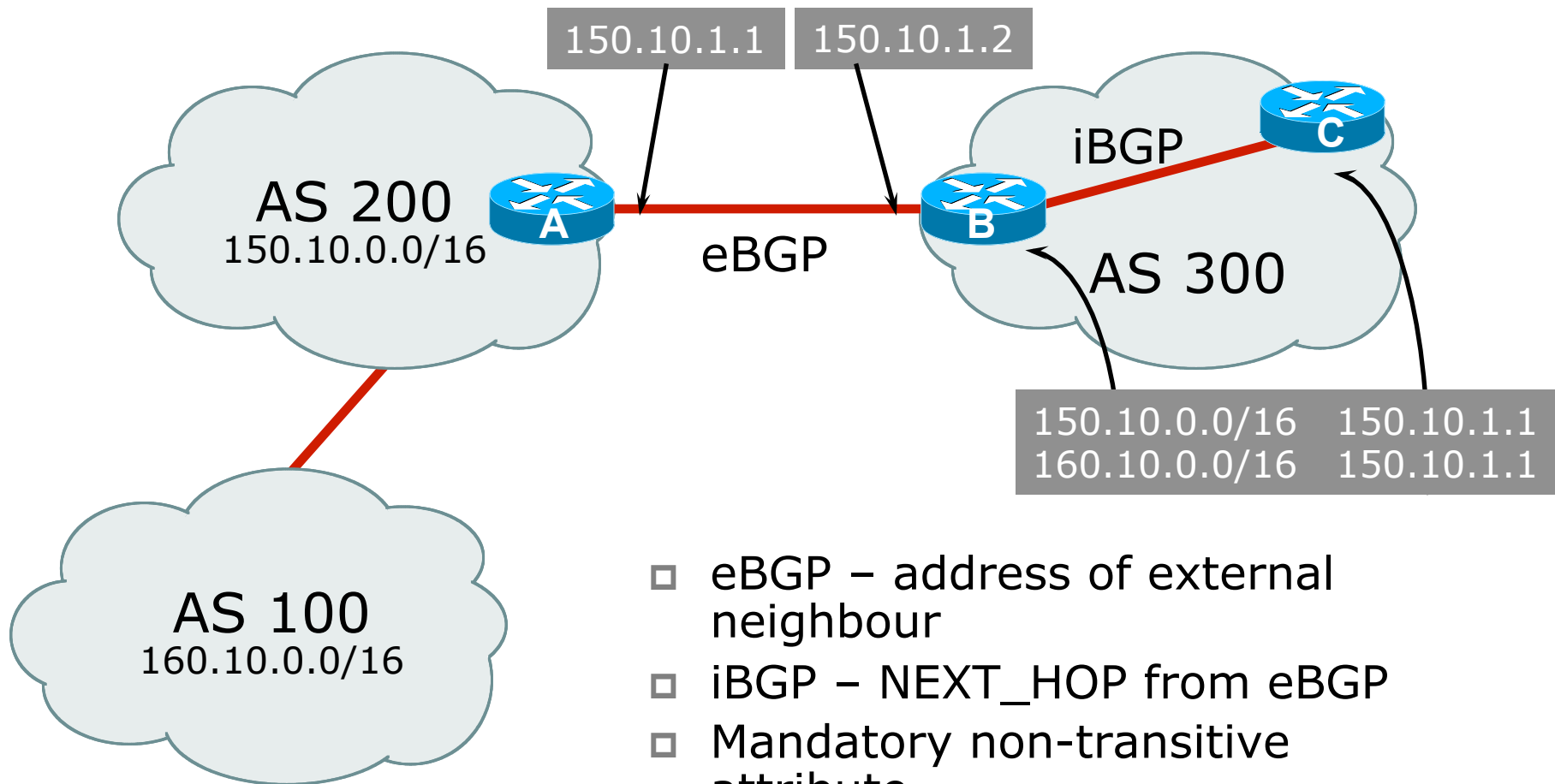180.10.0.0/16

AS 300

AS 400
150.10.0.0/16

| 180.10.0.0/16 | 300 200 100 |
| 170.10.0.0/16 | 300 200 |

AS 500

| 180.10.0.0/16 | 300 200 100 |
| 170.10.0.0/16 | 300 200 |
| 150.10.0.0/16 | 300 400 |

4

# AS-Path (with 16 and 32-bit ASNs)

- Internet with 16-bit and 32-bit ASNs
  - 32-bit ASNs are 65536 and above
- AS-PATH length maintained

AS 80000
170.10.0.0/16

AS 70000
180.10.0.0/16

AS 300

| | |
|---|---|
| 180.10.0.0/16 | 300 23456 23456 |
| 170.10.0.0/16 | 300 23456 |

AS 400
150.10.0.0/16

AS 90000

| | |
|---|---|
| 180.10.0.0/16 | 300 80000 70000 |
| 170.10.0.0/16 | 300 80000 |
| 150.10.0.0/16 | 300 400 |

# AS-Path loop detection

AS 200
170.10.0.0/16

AS 100
180.10.0.0/16

AS 300

140.10.0.0/16

AS 500

| | |
|---|---|
| 140.10.0.0/16 | 500 300 |
| 170.10.0.0/16 | 500 300 200 |

| | |
|---|---|
| 180.10.0.0/16 | 300 200 100 |
| 170.10.0.0/16 | 300 200 |
| 140.10.0.0/16 | 300 |

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

# Next Hop

150.10.1.1    150.10.1.2

iBGP    C

AS 200
150.10.0.0/16    A    B

eBGP    AS 300

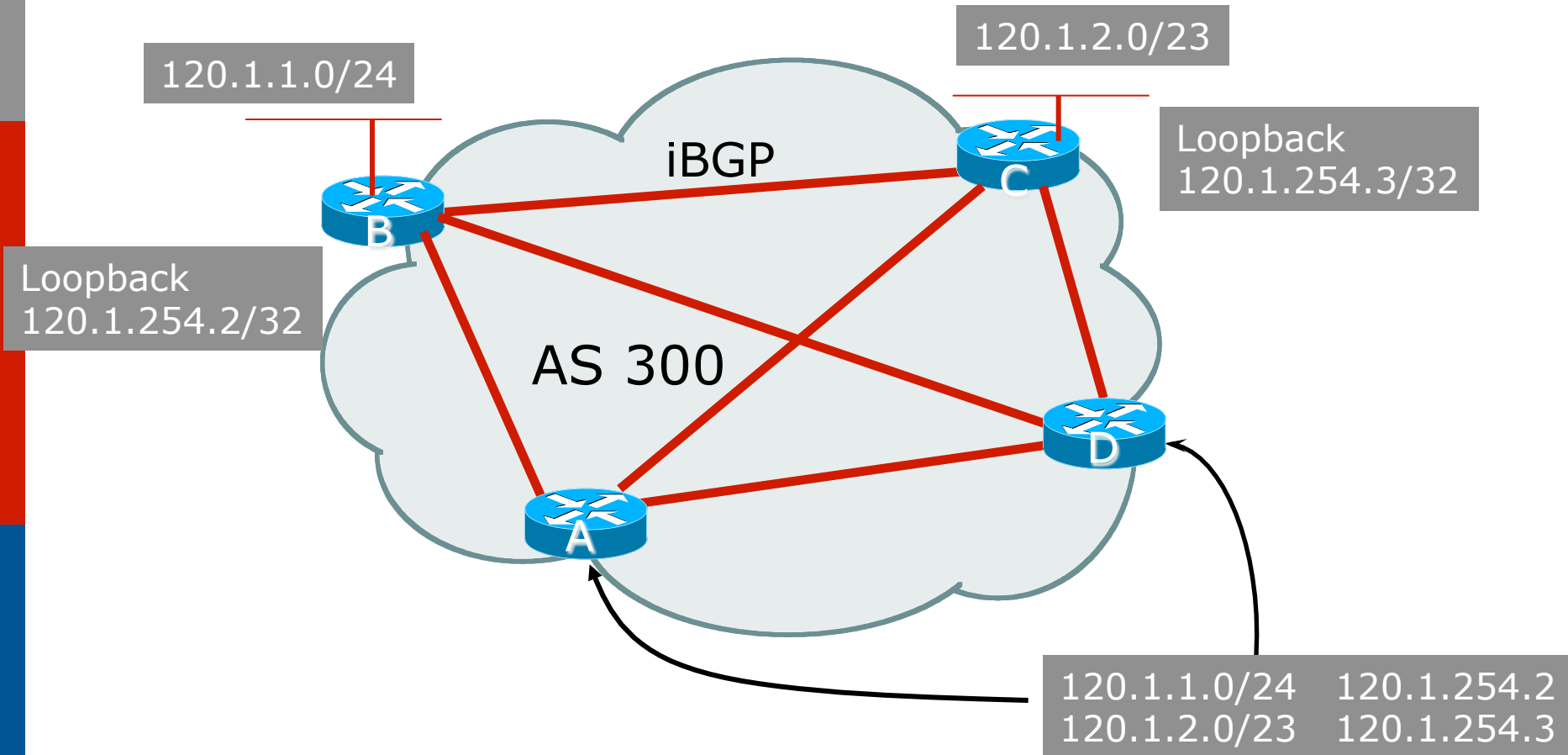150.10.0.0/16    150.10.1.1
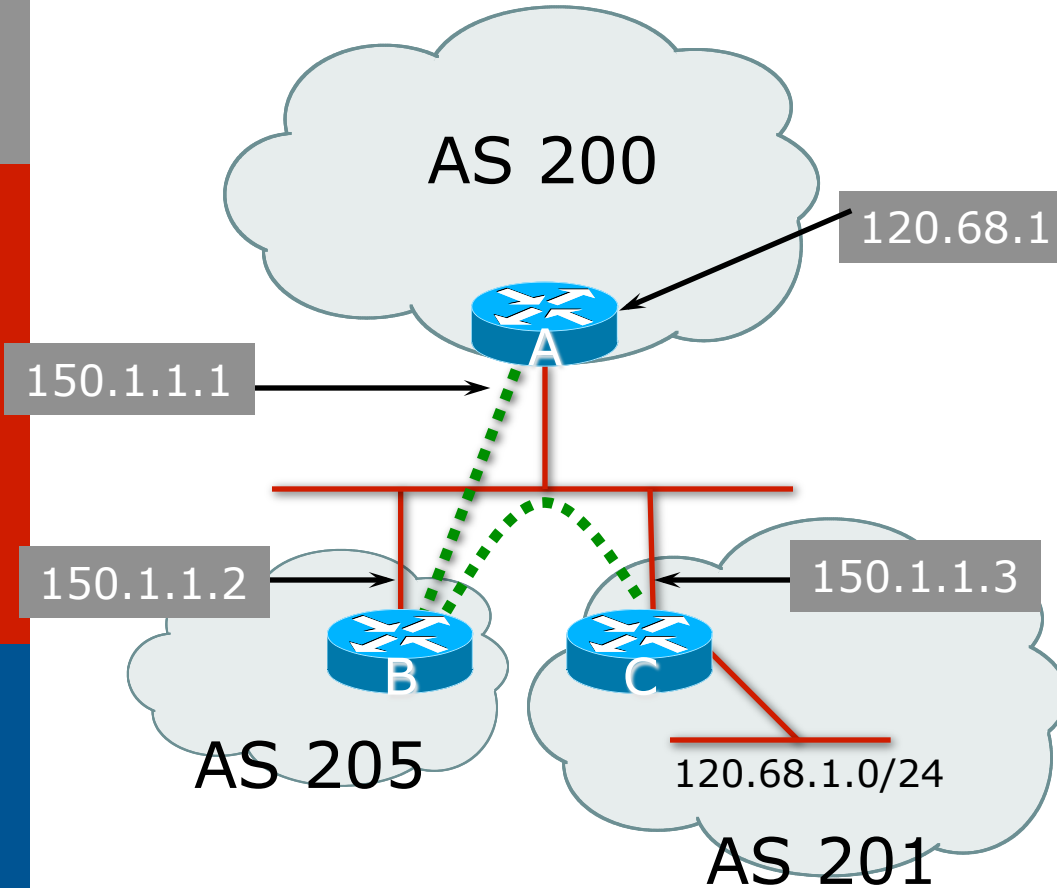160.10.0.0/16    150.10.1.1

AS 100
160.10.0.0/16

- eBGP – address of external neighbour
- iBGP – NEXT_HOP from eBGP
- Mandatory non-transitive attribute

# iBGP Next Hop



- Next hop is ibgp router loopback address
- Recursive route look-up

8

# Third Party Next Hop

AS 200

120.68.1.0/24      150.1.1.3

150.1.1.1

A

150.1.1.2

150.1.1.3

B

C

AS 205

120.68.1.0/24

AS 201

- eBGP between Router A and Router B
- eBGP between Router B and Router C
- 120.68.1/24 prefix has next hop address of 150.1.1.3 – this is used by Router A instead of 150.1.1.2 as it is on same subnet as Router B
- More efficient
- No extra config needed[9]

# Next Hop Best Practice

- Cisco IOS default is for external next-hop to be propagated unchanged to iBGP peers
  - This means that IGP has to carry external next-hops
  - Forgetting means external network is invisible
  - With many eBGP peers, it is unnecessary extra load on IGP
- ISP Best Practice is to change external next-hop to be that of the local router

```
neighbor x.x.x.x next-hop-self
```

# Next Hop (Summary)

- IGP should carry route to next hops
- Recursive route look-up
- Unlinks BGP from actual physical topology
- Use "next-hop-self" for external next hops
- Allows IGP to make intelligent forwarding decision

# Origin

- Conveys the origin of the prefix
- <span style="color:red">Historical</span> attribute
  - Used in transition from EGP to BGP
- Transitive and Mandatory Attribute
- Influences best path selection
- Three values: IGP, EGP, incomplete
  - IGP – generated by BGP network statement
  - EGP – generated by EGP
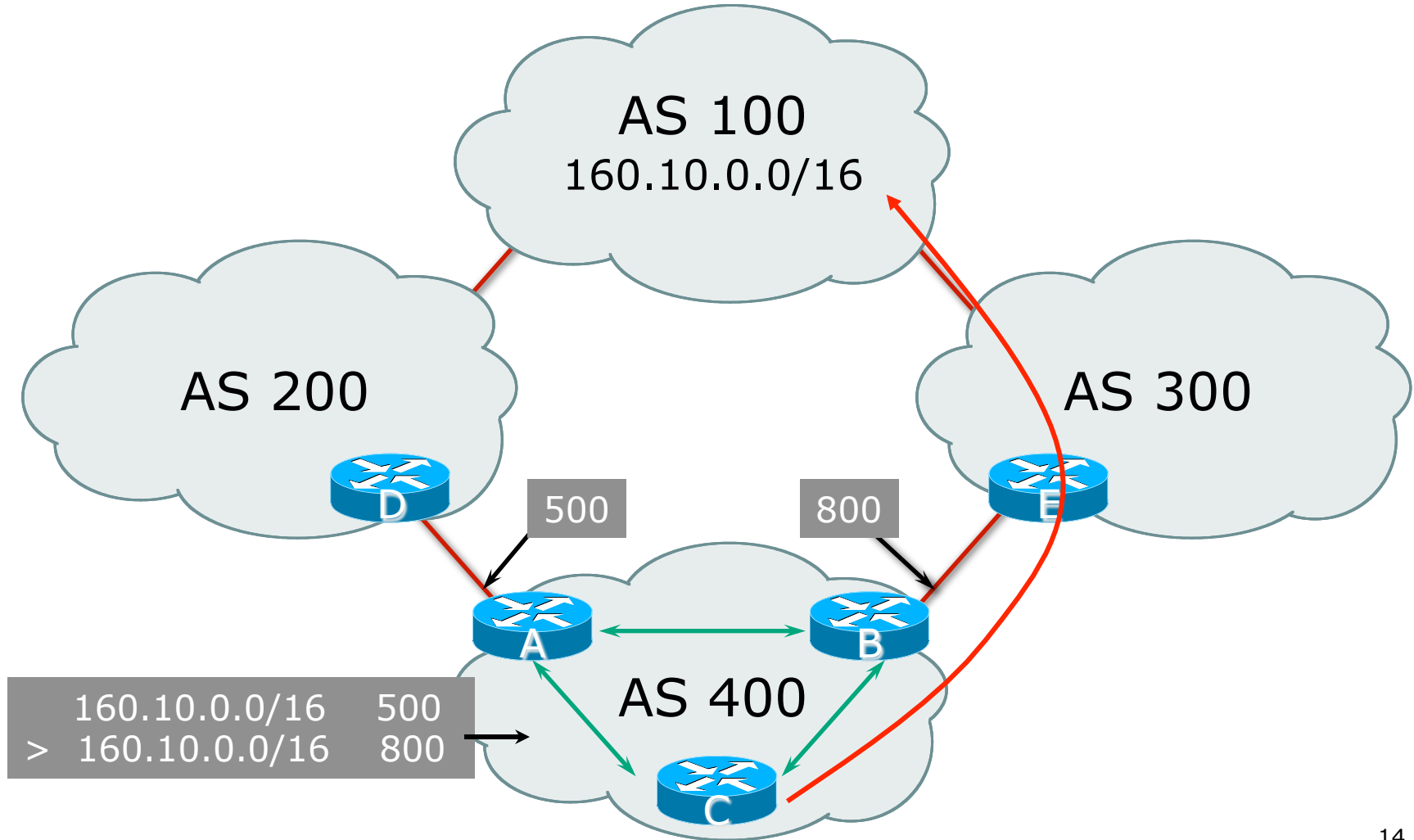  - incomplete – redistributed from another routing protocol

# Aggregator

- Conveys the IP address of the router or BGP speaker generating the aggregate route
- Optional & transitive attribute
- Useful for debugging purposes
- Does not influence best path selection
- Creating aggregate using "aggregate-address" sets the aggregator attribute:

```
router bgp 100
  aggregate-address 100.1.0.0 255.255.0.0
```

# Local Preference



AS 100
160.10.0.0/16

AS 200

AS 300

AS 400

D

A    B

C

E

500

800

| 160.10.0.0/16 | 500 |
|---|---|
| > 160.10.0.0/16 | 800 |

14

# Local Preference

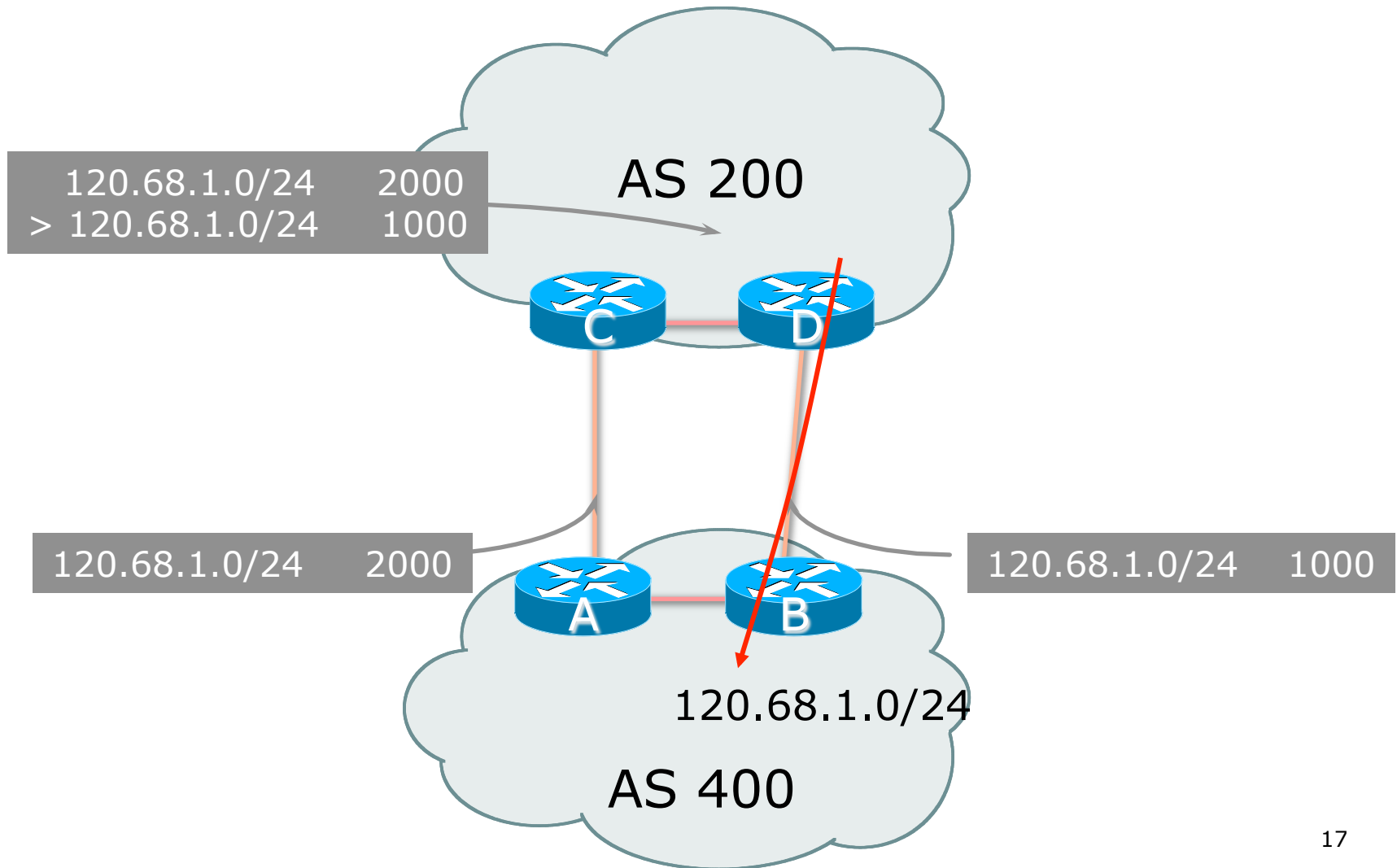- Non-transitive and optional attribute
- Local to an AS only
    - Default local preference is 100 (IOS)
- Used to influence BGP path selection
    - determines best path for *outbound* traffic
- Path with highest local preference wins

# Local Preference

- Configuration of Router B:

```
router bgp 400
  neighbor 120.5.1.1 remote-as 300
  neighbor 120.5.1.1 route-map local-pref in
!
route-map local-pref permit 10
 match ip address prefix-list MATCH
 set local-preference 800
route-map local-pref permit 20
!
ip prefix-list MATCH permit 160.10.0.0/16
```

# Multi-Exit Discriminator (MED)

AS 200

| 120.68.1.0/24 | 2000 |
| > 120.68.1.0/24 | 1000 |

C     D

| 120.68.1.0/24 | 2000 |

A     B

| 120.68.1.0/24 | 1000 |

120.68.1.0/24

AS 400

17

# Multi-Exit Discriminator

- Inter-AS – non-transitive & optional attribute
- Used to convey the relative preference of entry points
    - determines best path for inbound traffic
- Comparable if paths are from same AS
    - `bgp always-compare-med` allows comparisons of MEDs from different ASes
- Path with lowest MED wins
- Absence of MED attribute implies MED value of **zero** (RFC4271)

# Deterministic MED

- IOS compares paths in the order they were received
  - Leads to inconsistent decisions when comparing MED
- Deterministic MED
  - Configure on all bgp speaking routers in AS
  - Orders paths according to their neighbouring ASN
  - Best path for each neighbour ASN group is selected
  - Overall bestpath selected from the winners of each group

```
router bgp 100
  bgp deterministic-med
```

# MED & IGP Metric

- IGP metric can be conveyed as MED
  - **`set metric-type internal`** in route-map
    - enables BGP to advertise a MED which corresponds to the IGP metric values
    - changes are monitored (and re-advertised if needed) every 600s
    - **`bgp dynamic-med-interval <secs>`**

# Multi-Exit Discriminator

□ Configuration of Router B:

```
router bgp 400
  neighbor 120.5.1.1 remote-as 200
  neighbor 120.5.1.1 route-map set-med out
 !
route-map set-med permit 10
  match ip address prefix-list MATCH
  set metric 1000
route-map set-med permit 20
 !
ip prefix-list MATCH permit 120.68.1.0/24
```
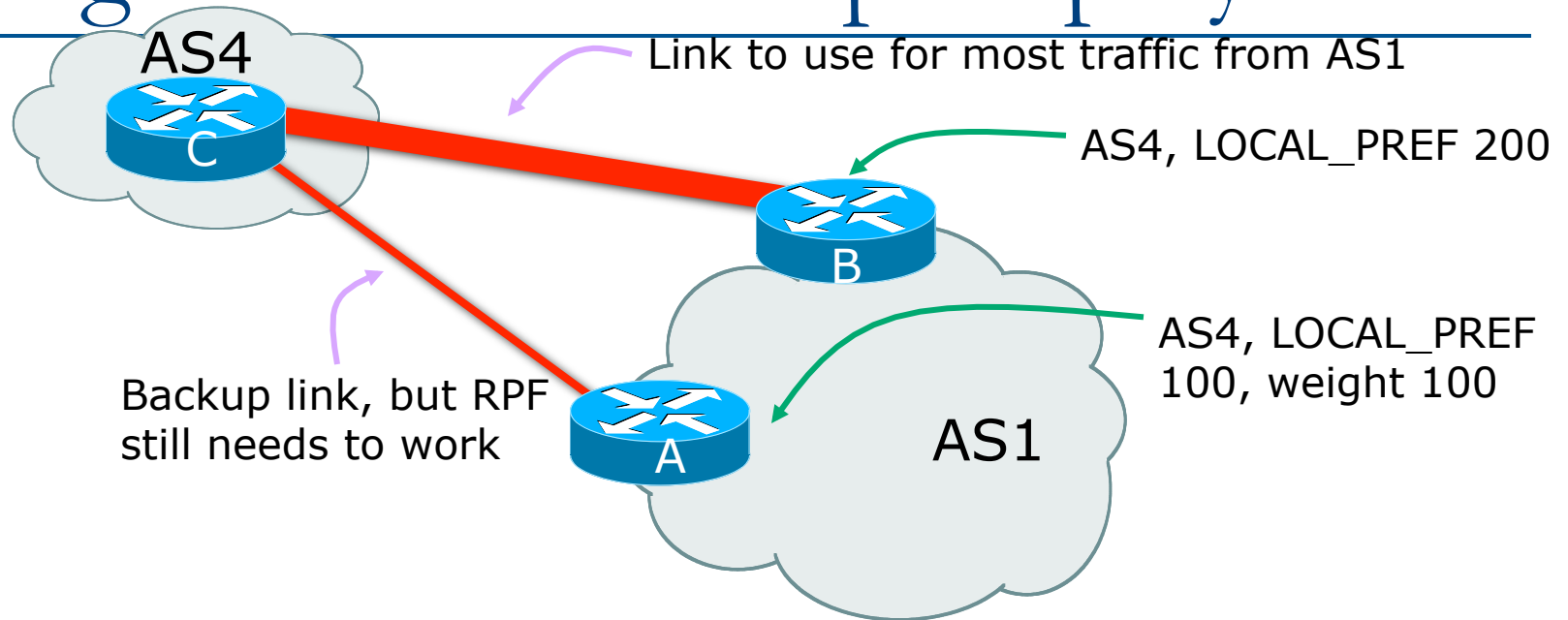
# Weight

- Not really an attribute – local to router
- Highest weight wins
- Applied to all routes from a neighbour

```
neighbor 120.5.7.1 weight 100
```
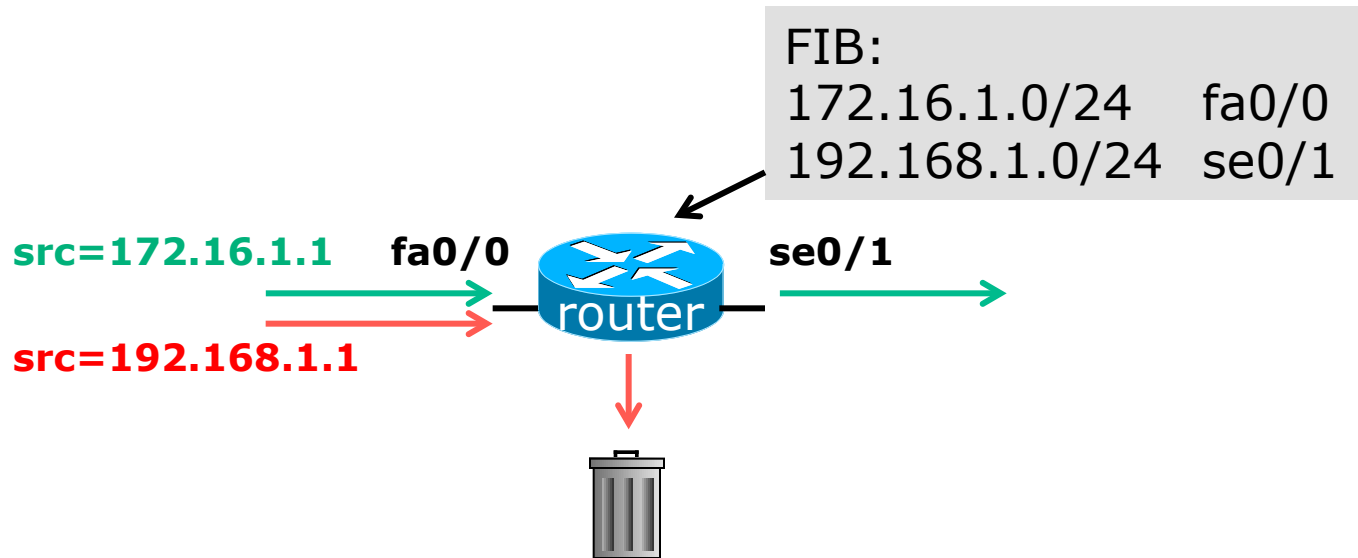
- Weight assigned to routes based on filter

```
neighbor 120.5.7.3 filter-list 3 weight 50
```
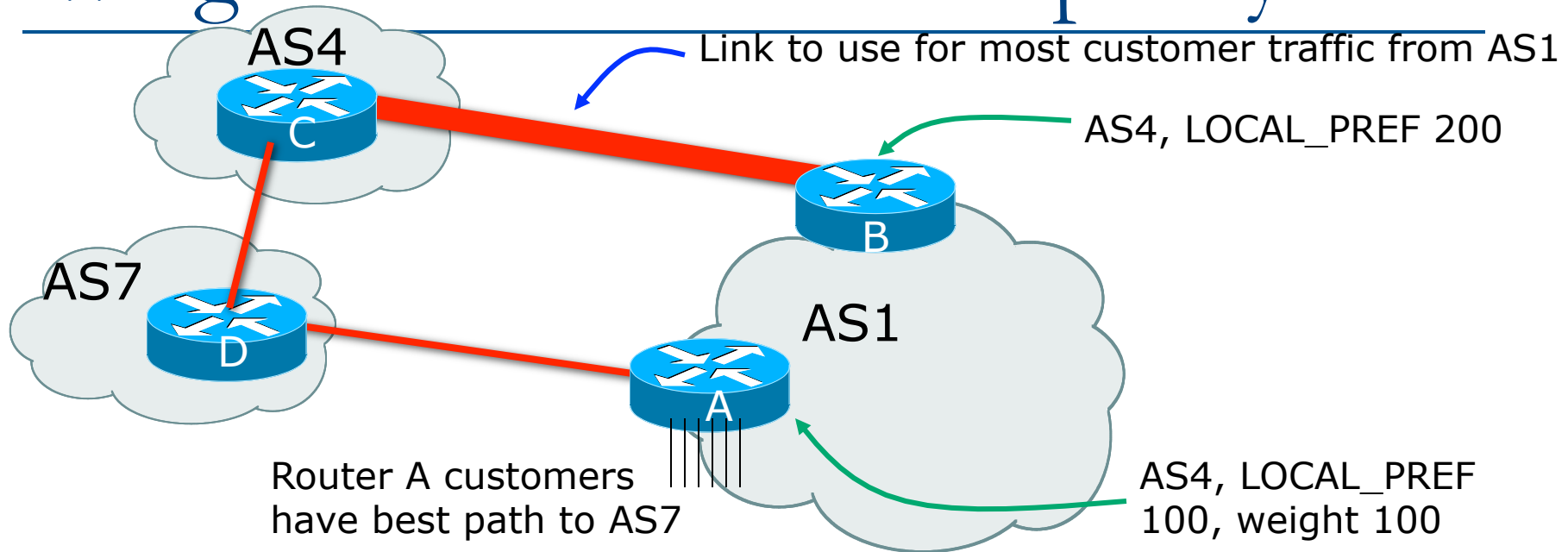
# Weight – Used to help Deploy RPF

AS4

Link to use for most traffic from AS1

C

AS4, LOCAL_PREF 200

B

AS4, LOCAL_PREF 100, weight 100

Backup link, but RPF still needs to work

A

AS1

- 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 back to AS4 would be via B, and the RPF check would fail

23

# Aside: What is uRPF?

FIB:
172.16.1.0/24    fa0/0
192.168.1.0/24  se0/1

src=172.16.1.1    fa0/0        se0/1

router

src=192.168.1.1

- Router compares source address of incoming packet with FIB entry
  - If FIB entry interface matches incoming interface, the packet is forwarded
  - If FIB entry interface does not match incoming interface, the packet is dropped

24

# Weight – Used for traffic policy

AS4

Link to use for most customer traffic from AS1

AS4, LOCAL_PREF 200

B

AS7

AS1

D

A

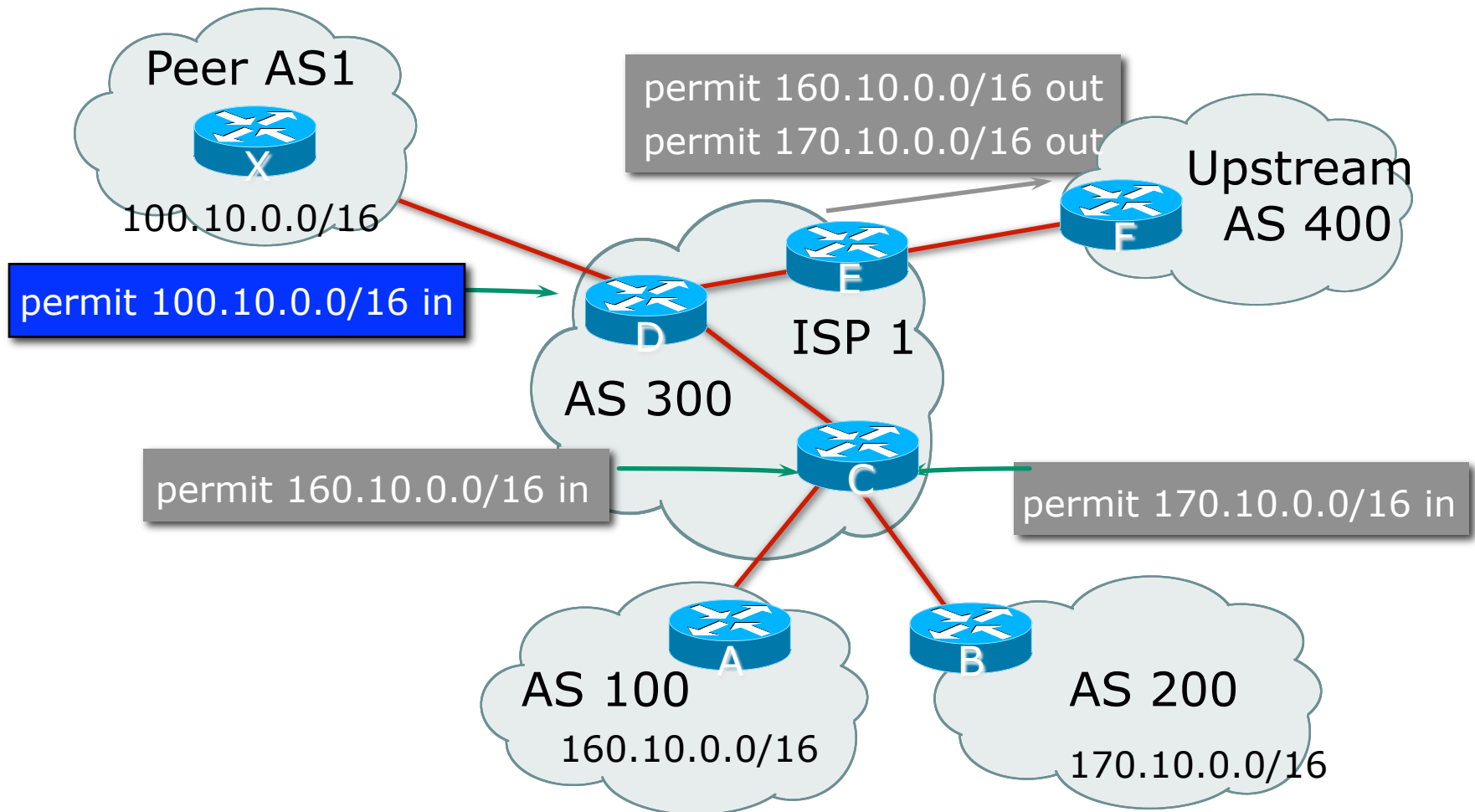Router A customers have best path to AS7

AS4, LOCAL_PREF 100, weight 100

- Best path to AS4 from AS1 is always via B due to local-pref
- But customers connected directly to Router A use the link to AS7 as best outbound path because of the high weight applied to routes heard from AS7
  - If the A to D link goes down, then the Router A customers see best path via Router B and AS4
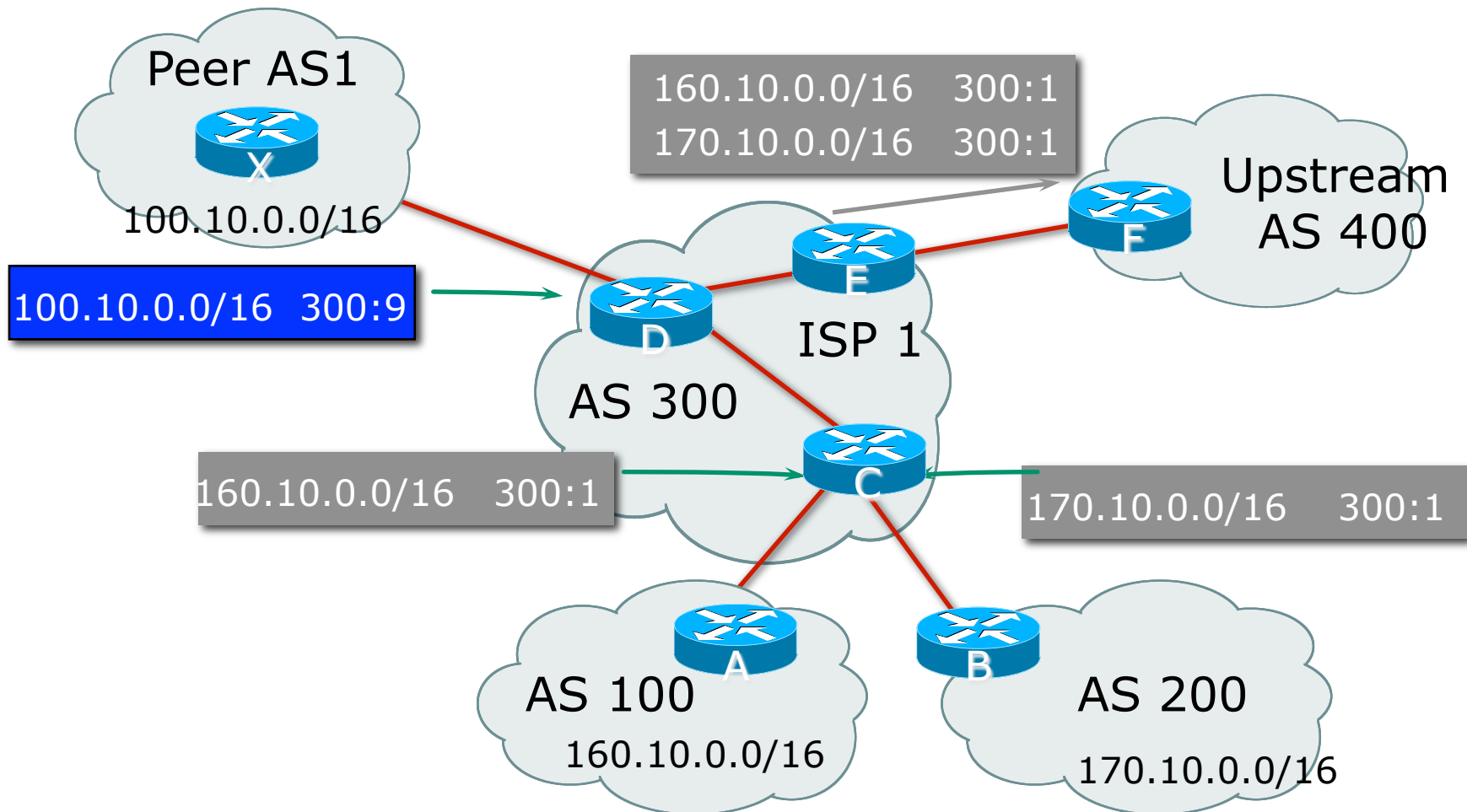
# Community

- Communities are described in RFC1997
  - Transitive and Optional Attribute
- 32 bit integer
  - Represented as two 16 bit integers (RFC1998)
  - Common format is <local-ASN>:xx
  - 0:0 to 0:65535 and 65535:0 to 65535:65535 are reserved
- Used to group destinations
  - Each destination could be member of multiple communities
- Very useful in applying policies within and between ASes

# Community Example (before)



Peer AS1

X

100.10.0.0/16

permit 100.10.0.0/16 in

permit 160.10.0.0/16 out
permit 170.10.0.0/16 out

Upstream AS 400

F

E

D

ISP 1

AS 300

permit 160.10.0.0/16 in

C

permit 170.10.0.0/16 in

A

B

AS 100

160.10.0.0/16

AS 200

170.10.0.0/16

# Community Example (after)



Peer AS1

X

100.10.0.0/16

160.10.0.0/16    300:1
170.10.0.0/16    300:1

Upstream AS 400

F

100.10.0.0/16  300:9

E

ISP 1

D

AS 300

160.10.0.0/16    300:1

C

170.10.0.0/16    300:1

A

AS 100

160.10.0.0/16

B

AS 200

170.10.0.0/16
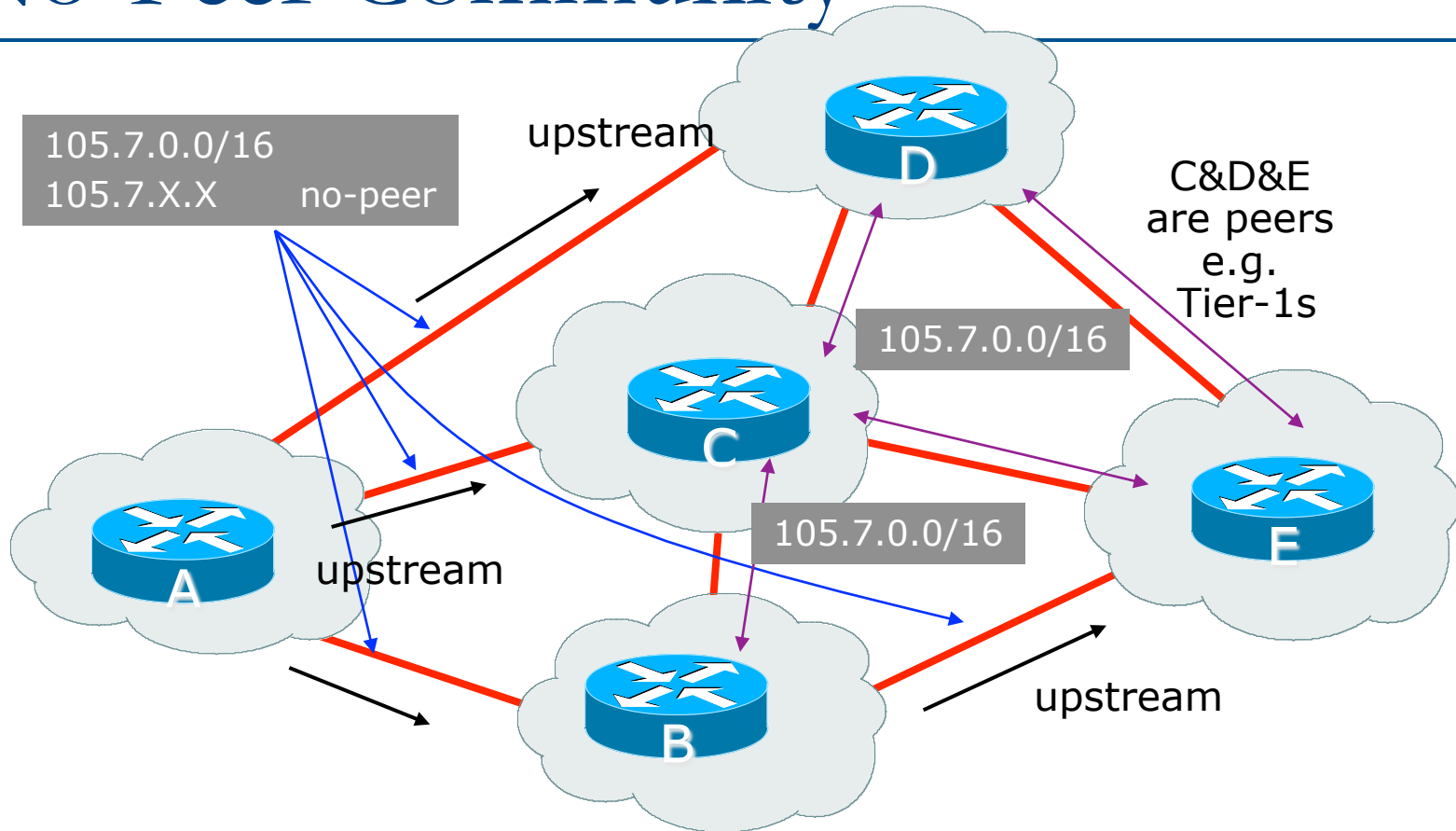
28

# Well-Known Communities

- Several well known communities
  - www.iana.org/assignments/bgp-well-known-communities
- no-export                    65535:65281
  - do not advertise to any eBGP peers
- no-advertise                 65535:65282
  - do not advertise to any BGP peer
- no-export-subconfed          65535:65283
  - do not advertise outside local AS (only used with confederations)
- no-peer                      65535:65284
  - do not advertise to bi-lateral peers (RFC3765)

# No-Export Community



105.7.0.0/16
105.7.X.X          no-export

105.7.X.X

AS 100

AS 200

105.7.0.0/16
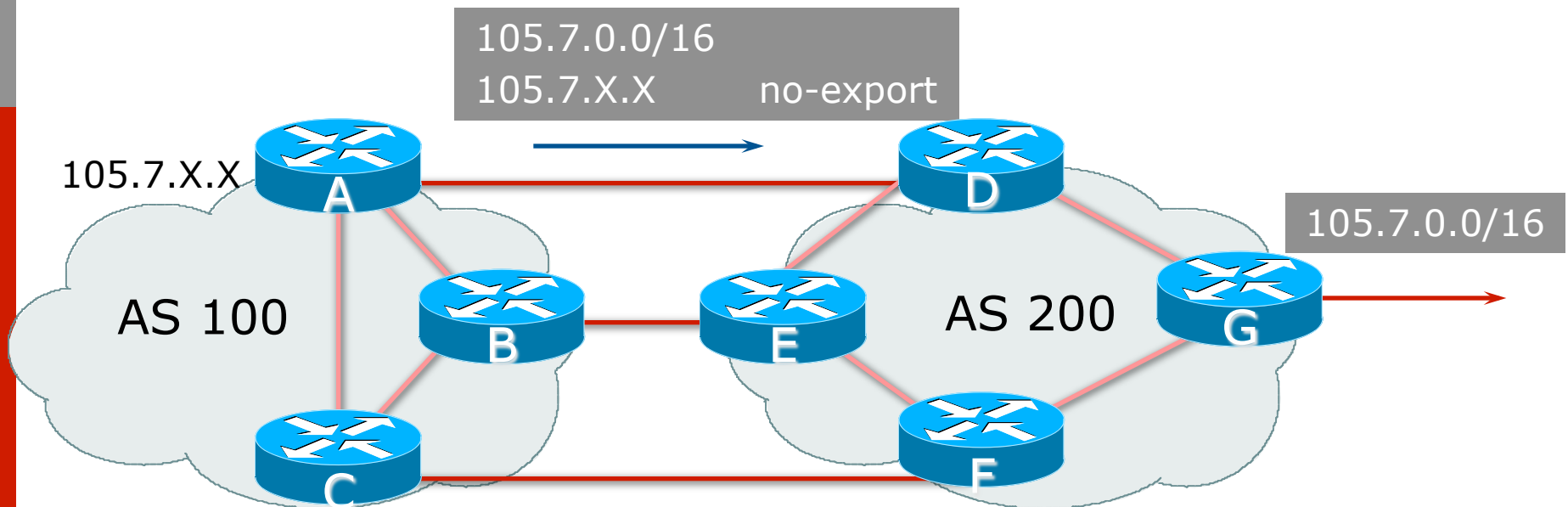
- AS100 announces aggregate and subprefixes
  - Intention is to improve loadsharing by leaking subprefixes
- Subprefixes marked with no-export community
- Router G in AS200 does not announce prefixes with no-export community set

# No-Peer Community



105.7.0.0/16
105.7.X.X       no-peer

upstream

D

C&D&E
are peers
e.g.
Tier-1s

105.7.0.0/16

A

upstream

105.7.0.0/16

E

B

upstream

- Sub-prefixes marked with no-peer community are not sent to bi-lateral peers
    - They are only sent to upstream providers

31

# What about 4-byte ASNs?

- Communities are widely used for encoding ISP routing policy
  - 32 bit attribute
- RFC1998 format is now "standard" practice
  - ASN:number
- Fine for 2-byte ASNs, but 4-byte ASNs cannot be encoded
- Solutions:
  - Use "private ASN" for the first 16 bits
  - Wait for http://datatracker.ietf.org/doc/draft-ietf-idr-as4octet-extcomm-generic-subtype/ to be implemented

# Summary
# Attributes in Action

```
Router6>sh ip bgp
BGP table version is 16, local router ID is 10.0.15.246
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale, m multipath, b backup-path, f RT-Filter,
              x best-external, a additional-path, c RIB-compressed,
Origin codes: i - IGP, e - EGP, ? - incomplete
RPKI validation codes: V valid, I invalid, N Not found
```

| | Network | Next Hop | Metric | LocPrf | Weight | Path |
|---|---|---|---|---|---|---|
| *>i | 10.0.0.0/26 | 10.0.15.241 | 0 | 100 | 0 | i |
| *>i | 10.0.0.64/26 | 10.0.15.242 | 0 | 100 | 0 | i |
| *>i | 10.0.0.128/26 | 10.0.15.243 | 0 | 100 | 0 | i |
| *>i | 10.0.0.192/26 | 10.0.15.244 | 0 | 100 | 0 | i |
| *>i | 10.0.1.0/26 | 10.0.15.245 | 0 | 100 | 0 | i |
| *> | 10.0.1.64/26 | 0.0.0.0 | 0 | | 32768 | i |
| *>i | 10.0.1.128/26 | 10.0.15.247 | 0 | 100 | 0 | i |
| *>i | 10.0.1.192/26 | 10.0.15.248 | 0 | 100 | 0 | i |
| *>i | 10.0.2.0/26 | 10.0.15.249 | 0 | 100 | 0 | i |
| *>i | 10.0.2.64/26 | 10.0.15.250 | 0 | 100 | 0 | i |

...

# BGP Path Selection Algorithm

Why is this the best path?

# BGP Path Selection Algorithm for Cisco IOS: Part One

1. Do not consider path if no route to next hop
2. Do not consider iBGP path if not synchronised (Cisco IOS)
3. Highest weight (local to router)
4. Highest local preference (global within AS)
5. Prefer locally originated route
6. Shortest AS path

# BGP Path Selection Algorithm for Cisco IOS: Part Two

7. Lowest origin code
   - IGP < EGP < incomplete
8. Lowest Multi-Exit Discriminator (MED)
   - If bgp deterministic-med, order the paths by AS number before comparing
   - If bgp always-compare-med, then compare for all paths
   - Otherwise MED only considered if paths are from the same AS (default)

# BGP Path Selection Algorithm for Cisco IOS: Part Three

9. Prefer eBGP path over iBGP path

10. Path with lowest IGP metric to next-hop

11. For eBGP paths:
    - If multipath is enabled, install N parallel paths in forwarding table
    - If router-id is the same, go to next step
    - If router-id is not the same, select the oldest path

# BGP Path Selection Algorithm for Cisco IOS: Part Four

12. Lowest router-id (originator-id for reflected routes)

13. Shortest cluster-list

    - Client must be aware of Route Reflector attributes!

14. Lowest neighbour address

# BGP Attributes and Path Selection

## AFNOG 2013 AR-E Workshop