BGP Attributes and Path Selection

ISP Workshops

Last updated 25th October 2021

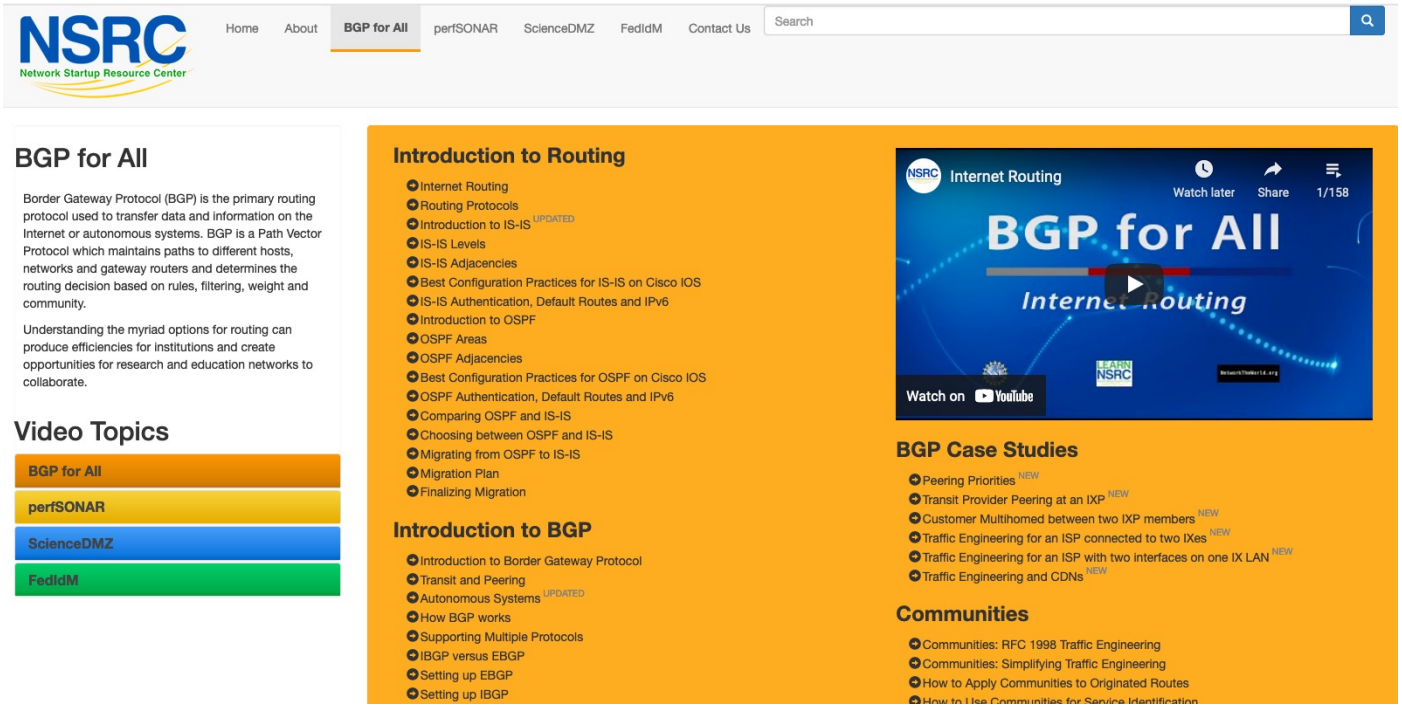
Acknowledgements

- This material originated from the Cisco ISP/IXP Workshop Programme developed by Philip Smith & Barry Greene
- Use of these materials is encouraged as long as the source is fully acknowledged and this notice remains in place
- Bug fixes and improvements are welcomed
  - Please email *workshop (at) bgp4all.com*

Philip Smith

BGP Videos

- NSRC has made a video recording of this presentation, as part of a library of BGP videos for the whole community to use:
  - https://learn.nsrc.org/bgp#bgp_attributes



BGP Attributes

BGP's policy tool kit

What Is an Attribute?

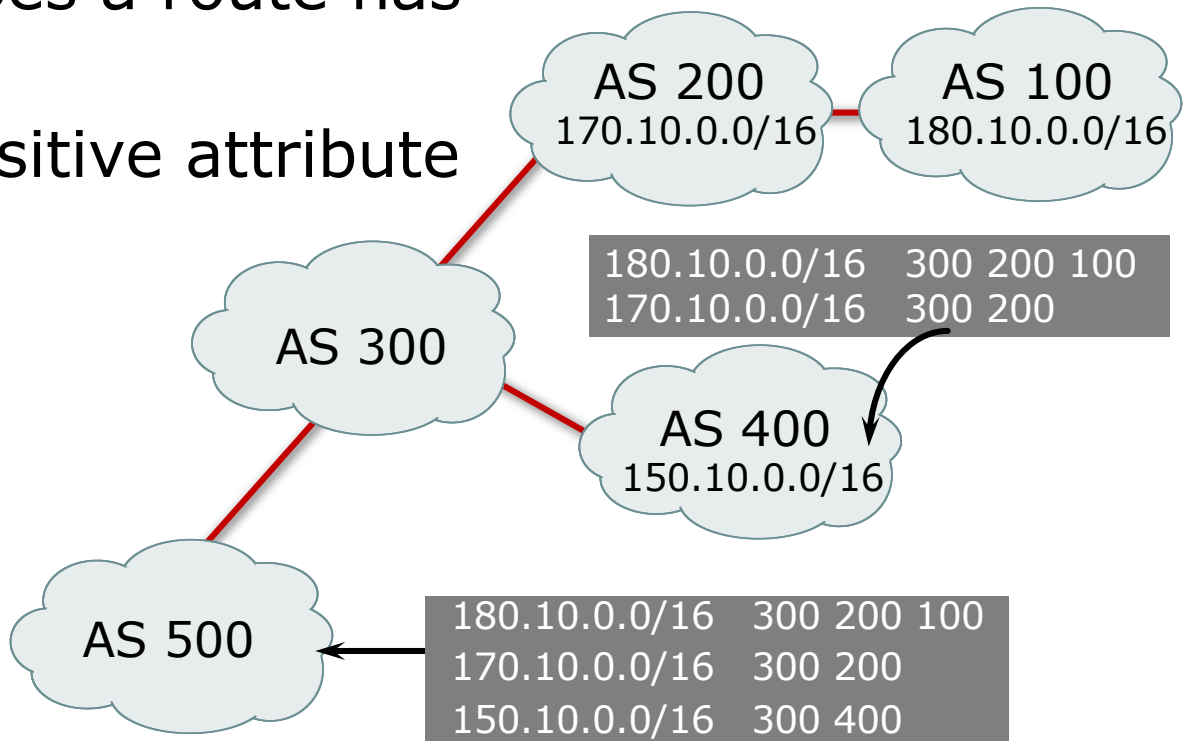
- Part of a BGP Update
- Describes the characteristics of prefix
- Can either be transitive or non-transitive
- Some are mandatory

BGP Attributes

- Carry various information about or characteristics of the prefix being propagated
  - AS-PATH
  - NEXT-HOP
  - ORIGIN
  - AGGREGATOR
  - LOCAL_PREFERENCE
  - Multi-Exit Discriminator
  - (Weight)
  - COMMUNITY

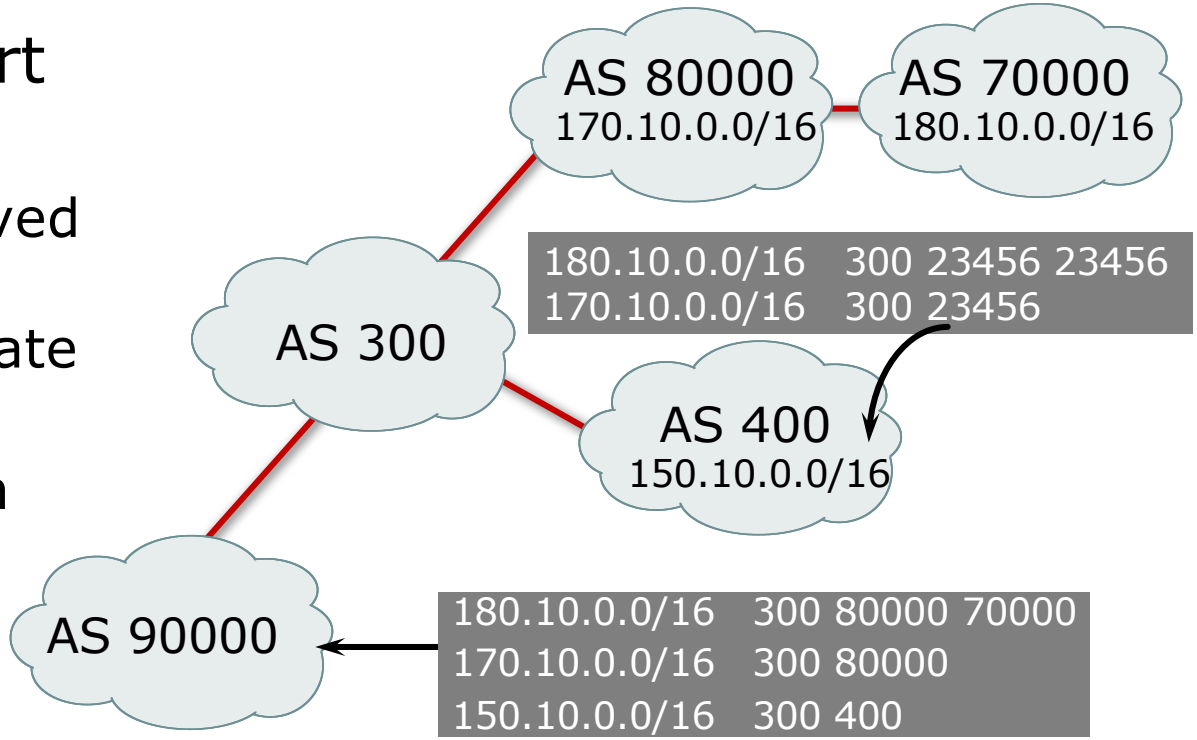
AS-Path

- Sequence of ASes a route has traversed
- Mandatory transitive attribute
- Used for:
  - Loop detection
  - Applying policy

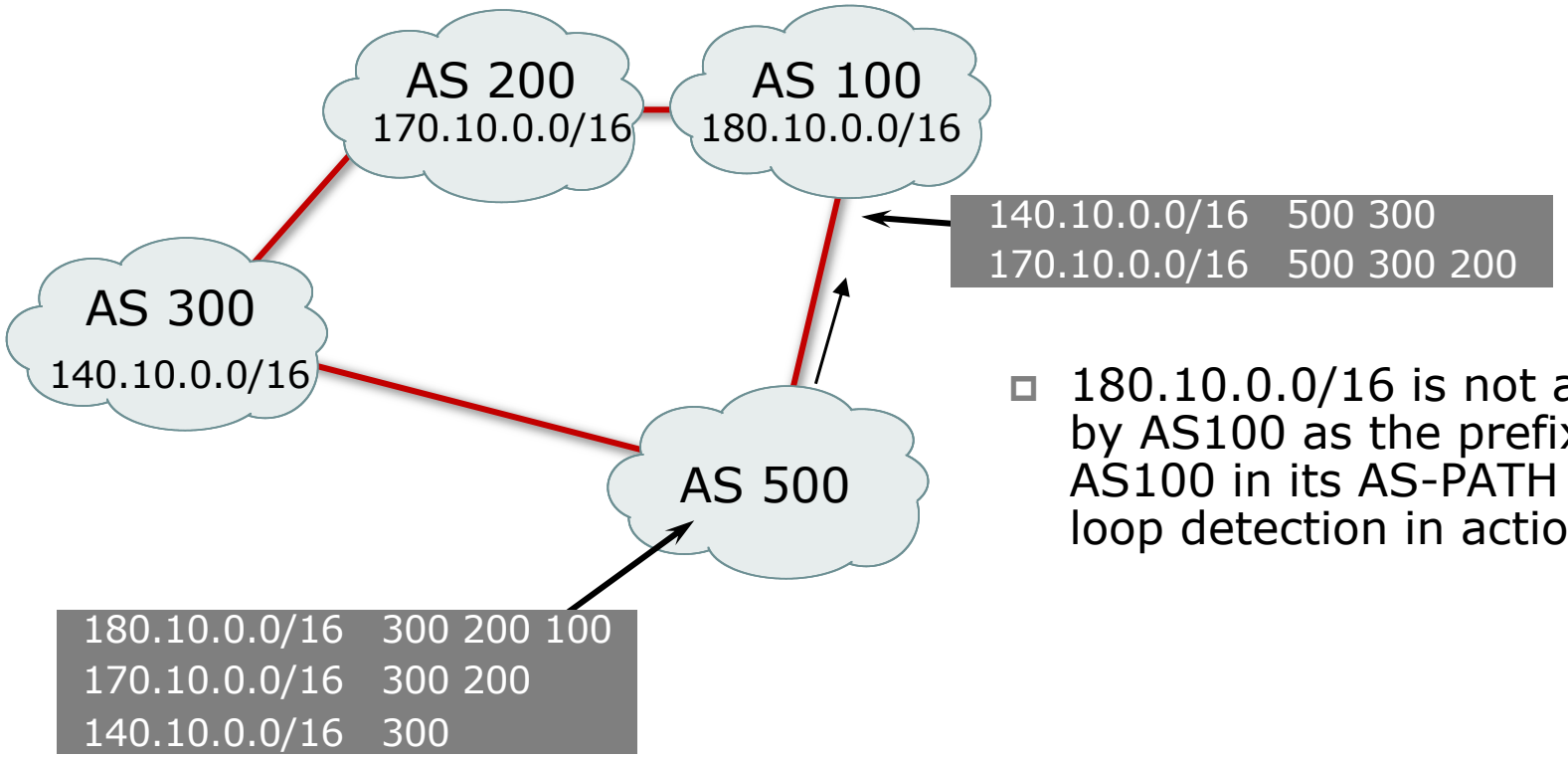


AS-Path (with old router in path)

- AS400 router does not support 32-bit ASNs
  - AS23456 reserved for use on old routers to indicate 32-bit ASNs
  - AS-PATH length maintained

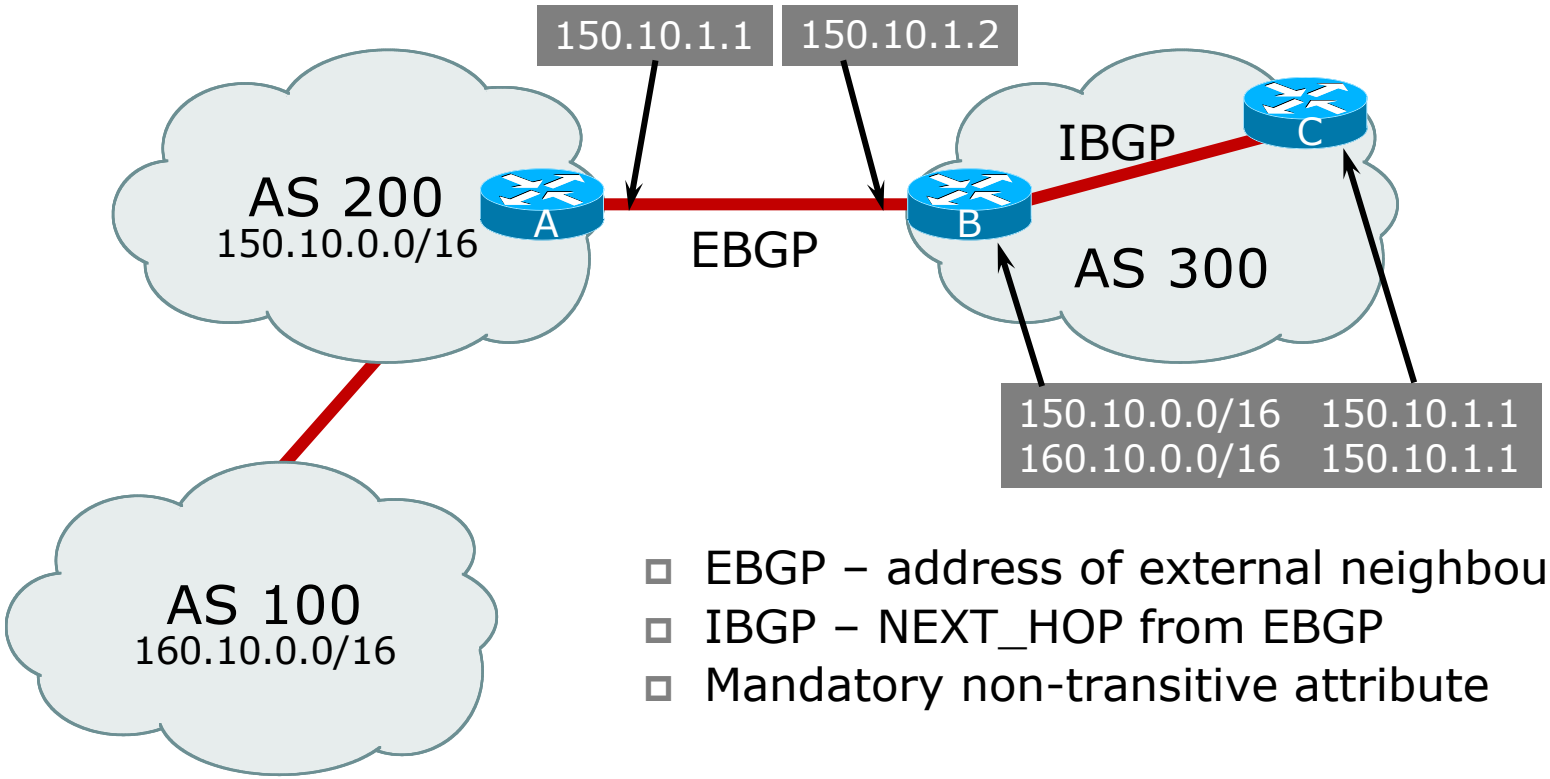


AS-Path loop detection



- 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



- EBGP – address of external neighbour
- IBGP – NEXT_HOP from EBGP
- Mandatory non-transitive attribute

Next Hop Best Practice

- The default behaviour 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

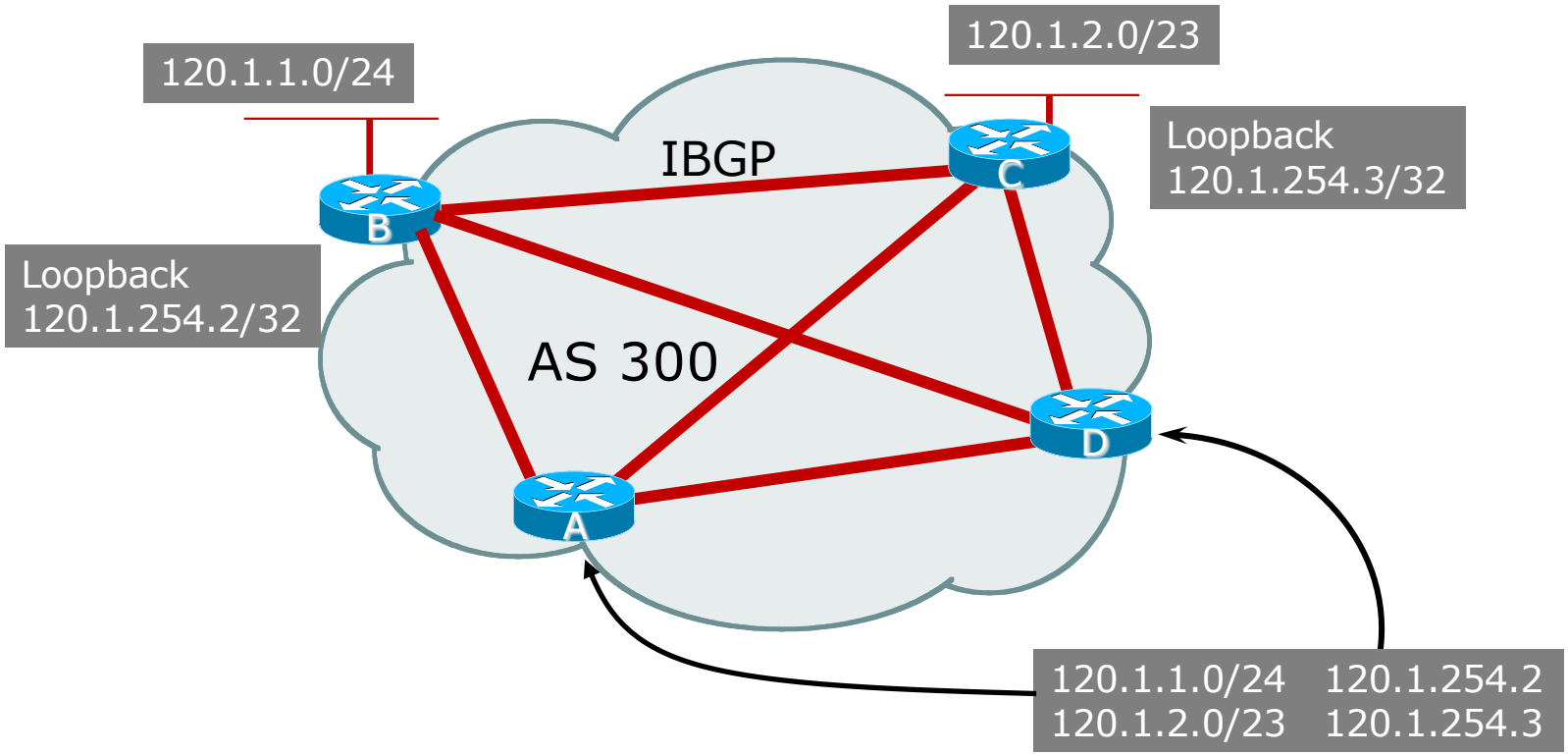
- ISP Best Practice is to change external next-hop to be that of the local router
  - Cisco IOS:

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

  - JunOS:

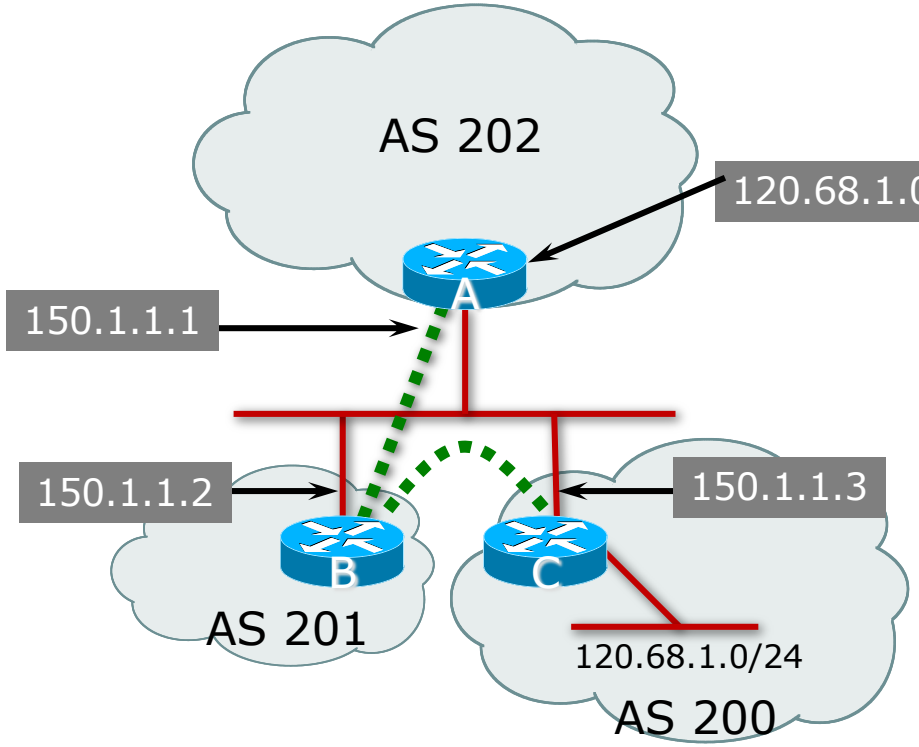
```
set policy-options
   policy-statement <name> term <name> then next-hop self
```

IBGP Next Hop



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

Third Party Next Hop



- 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 configuration needed

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

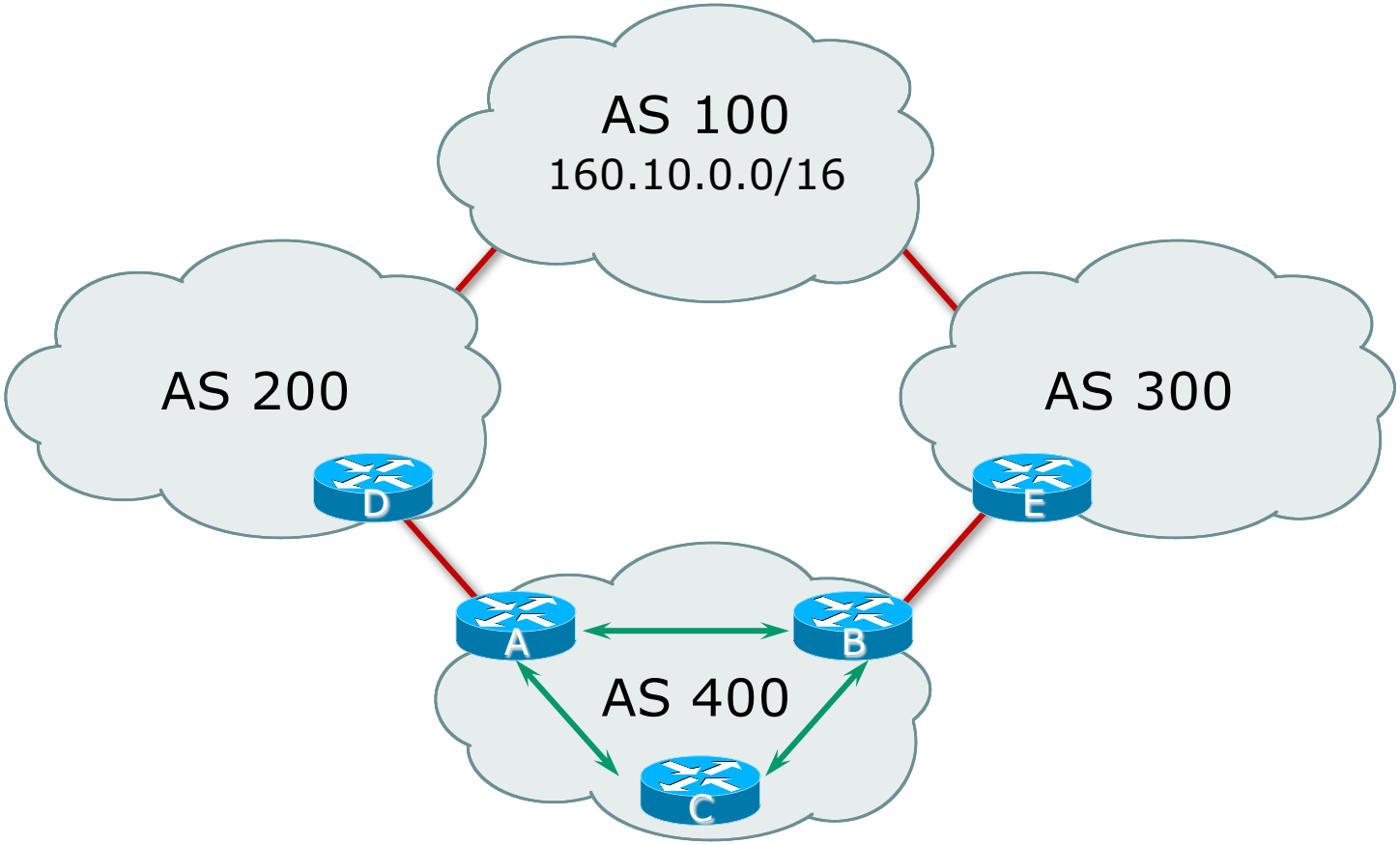
- Conveys the origin of the prefix
- Historical attribute
  - Used in transition from EGP to BGP
- Transitive and Mandatory Attribute
- Influences best path selection
- Three values: IGP, EGP, incomplete
  - IGP – generated by BGP network statement
  - EGP – generated by EGP
  - incomplete – redistributed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

Aggregator

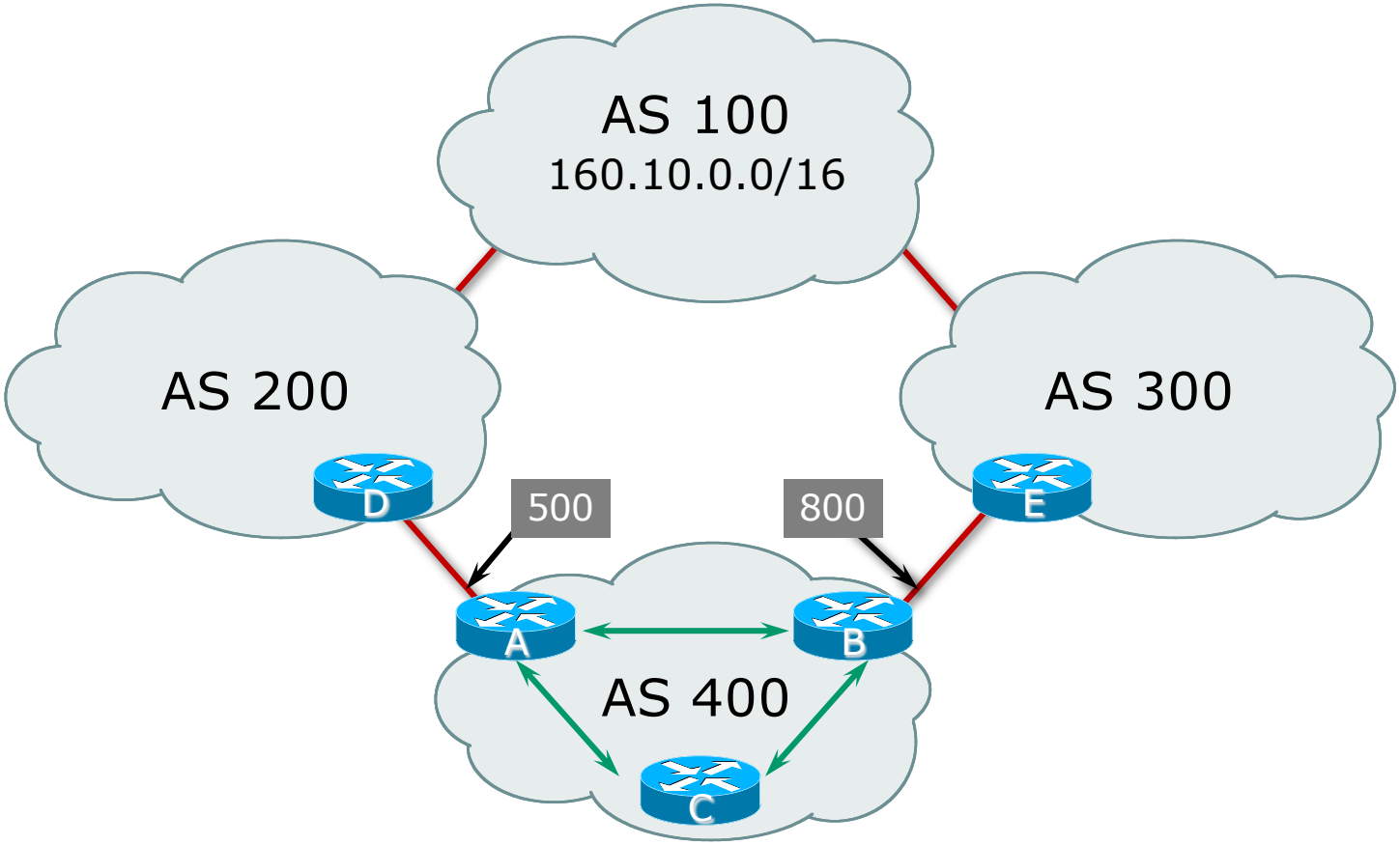
- Conveys the IP address of the router or BGP speaker generating the aggregate route
- Optional & transitive attribute
- Useful for debugging purposes
- Does not influence best path selection
- Creating aggregate using “aggregate-address” sets the aggregator attribute:

```
router bgp 100
 address-family ipv4
  aggregate-address 100.1.0.0 255.255.0.0
```

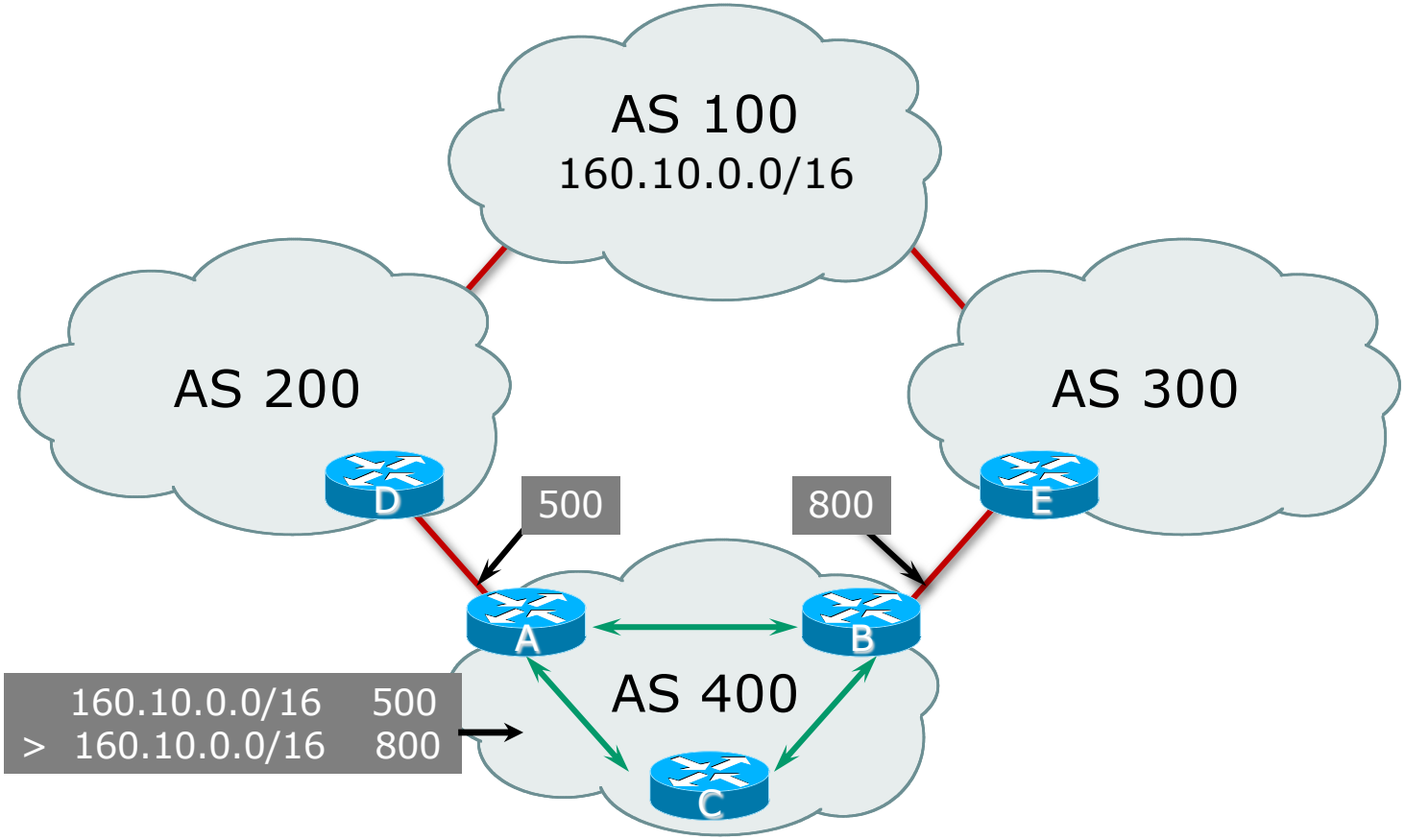
Local Preference



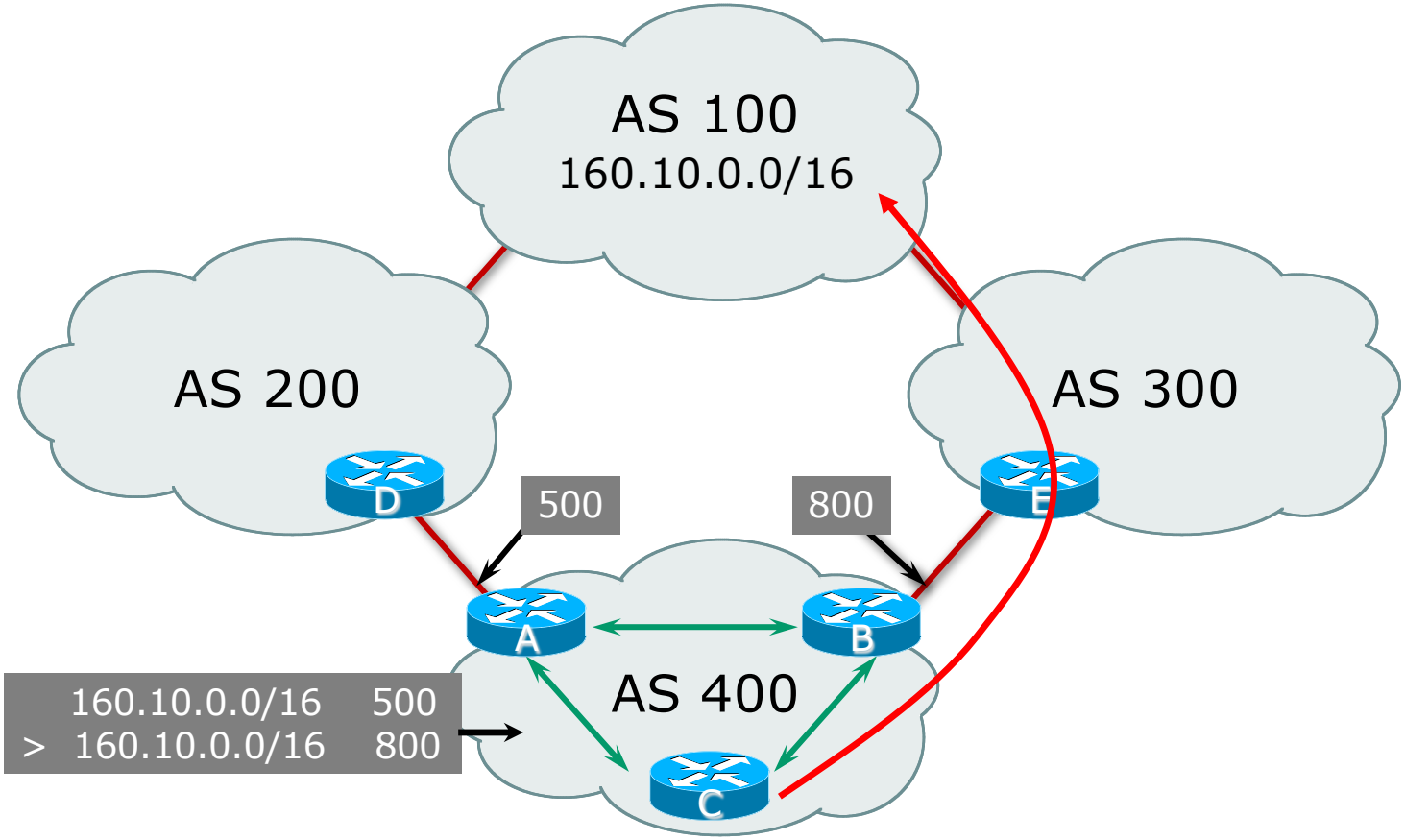
Local Preference



Local Preference



Local Preference



Local Preference

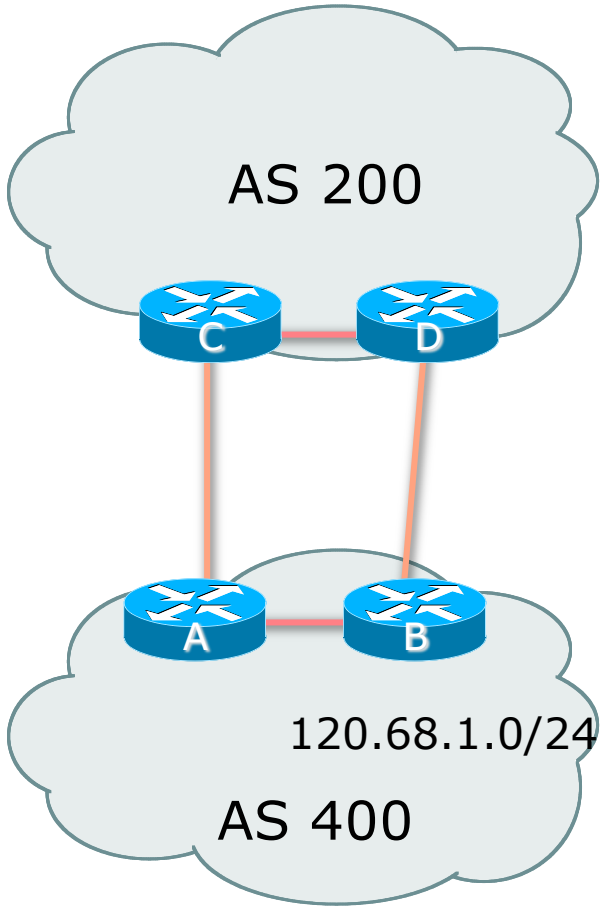
- Non-transitive and optional attribute
- Local to an AS only
  - Default local preference is 100 (IOS)
- Used to influence BGP path selection
  - Determines best path for *outbound* traffic
- Path with highest local preference wins

Local Preference

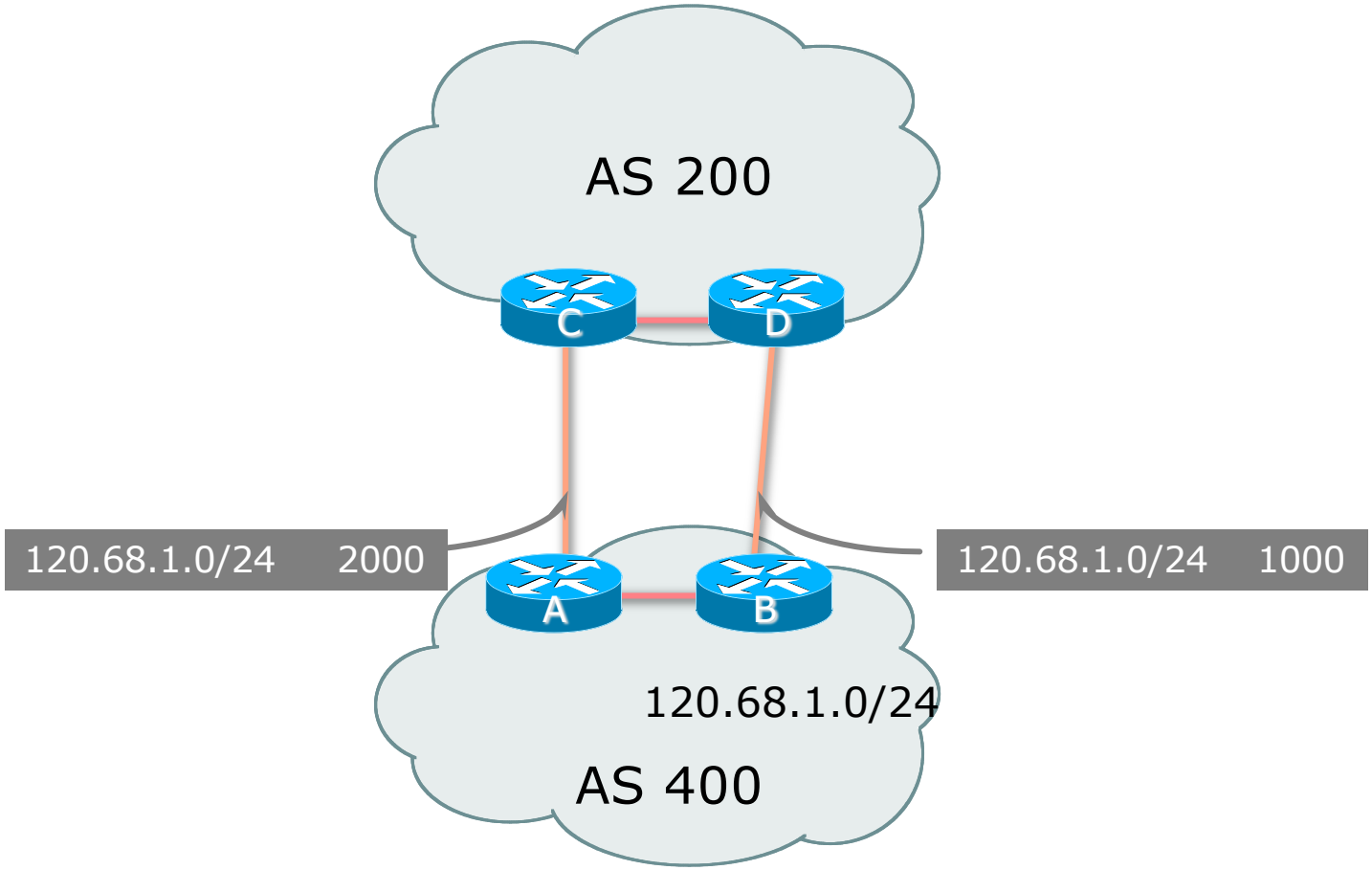
- Configuration of Router B:

```
router bgp 400
 address-family ipv4
  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
```

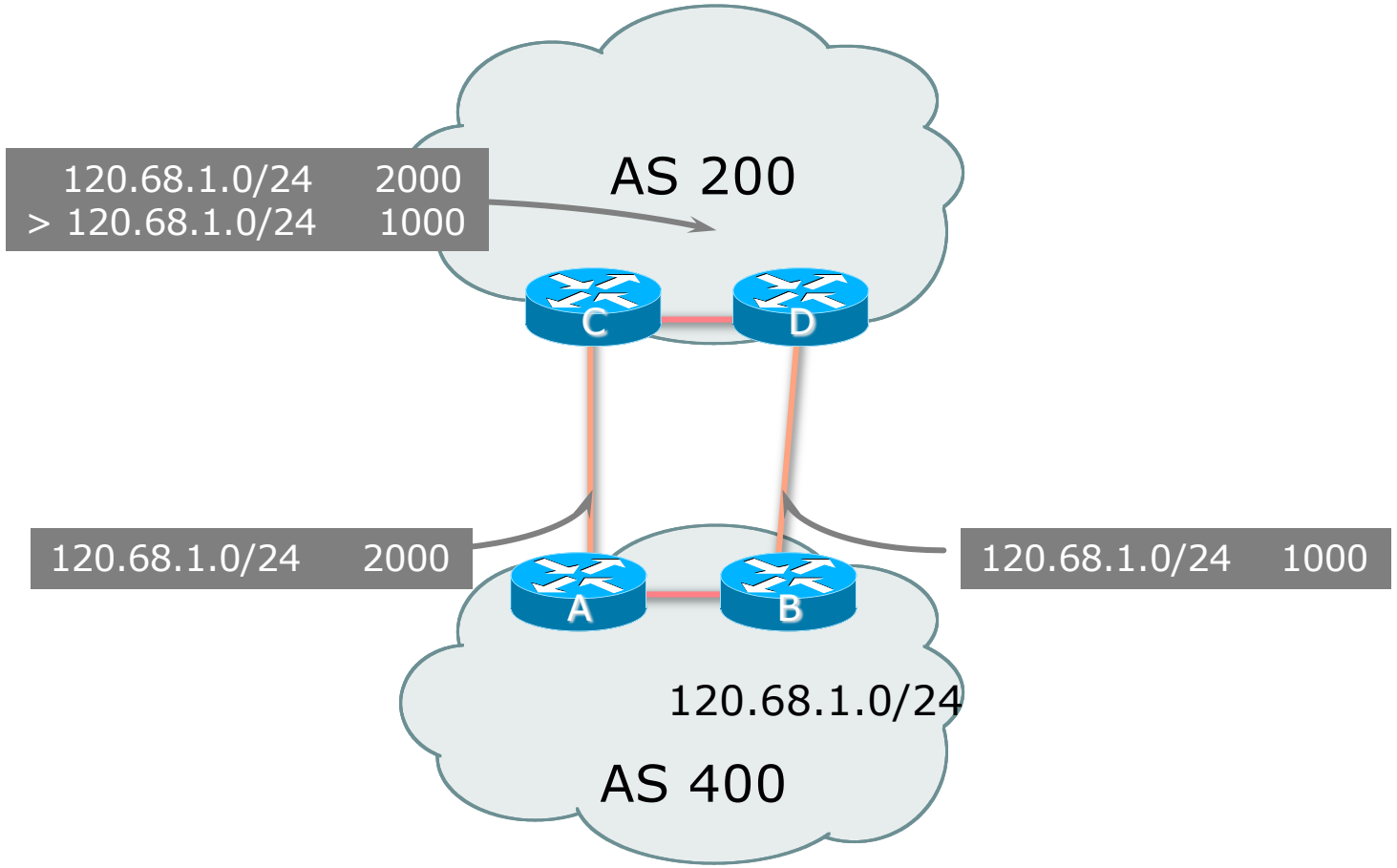
Multi-Exit Discriminator (MED)



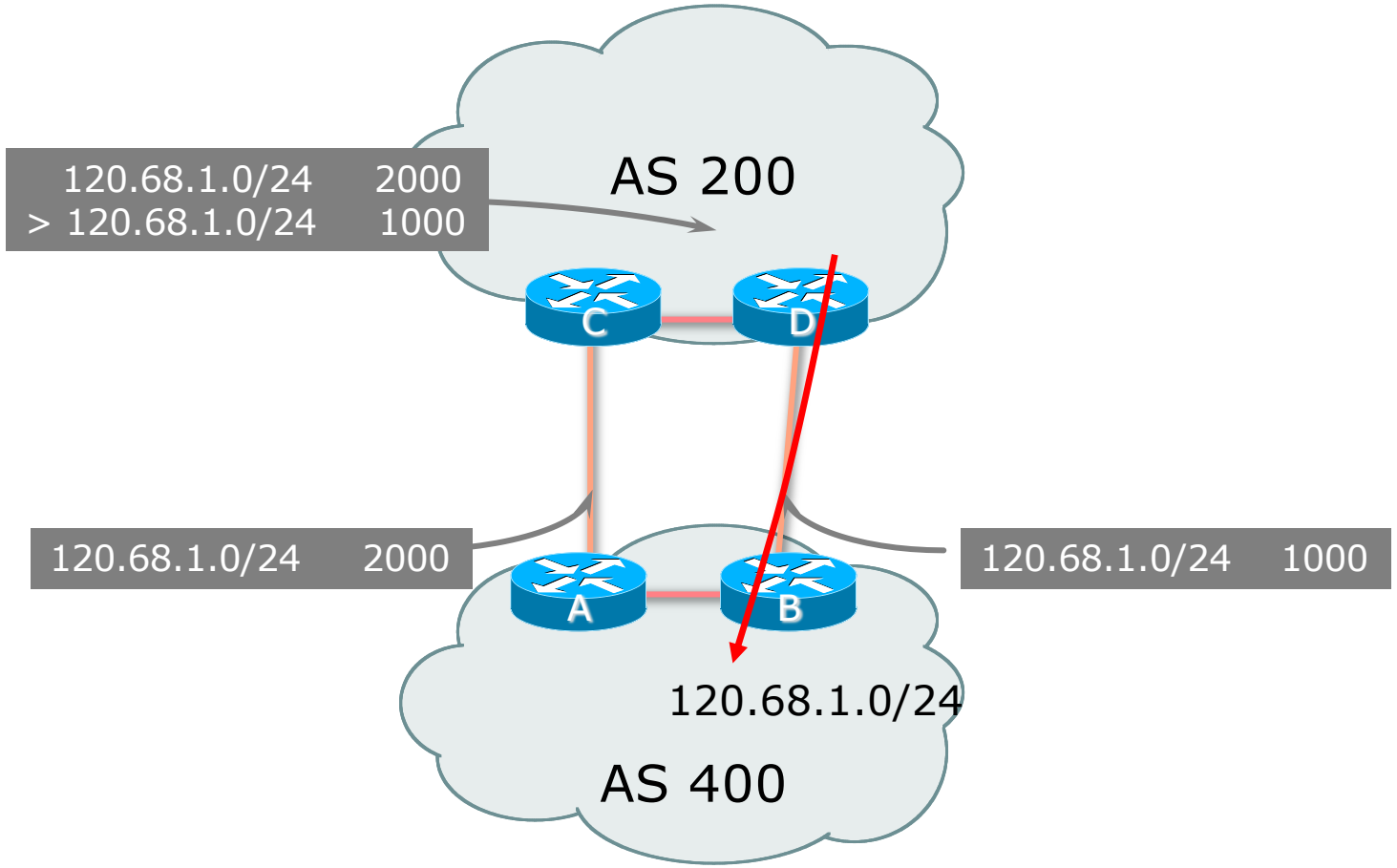
Multi-Exit Discriminator (MED)



Multi-Exit Discriminator (MED)



Multi-Exit Discriminator (MED)



Multi-Exit Discriminator

- Inter-AS – non-transitive & optional attribute
- Used to convey the relative preference of entry points
  - Determines best path for inbound traffic
- Comparable if paths are from same AS
  - `bgp always-compare-med` allows comparisons of MEDs from different ASes
  - Also available in JunOS:

```
set protocols bgp path-selection always-compare-med
```

- Path with lowest MED wins
- Absence of MED attribute implies MED value of **zero** (RFC4271)

Multi-Exit Discriminator

- Configuration of Router B:

```
router bgp 400
 address-family ipv4
  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
```

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 10
 bgp deterministic-med
```

- Deterministic MED is default in JunOS
  - Non-deterministic behaviour enabled with

```
set protocols bgp path-selection cisco-non-deterministic
```

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
    - Monitoring period can be changed using:

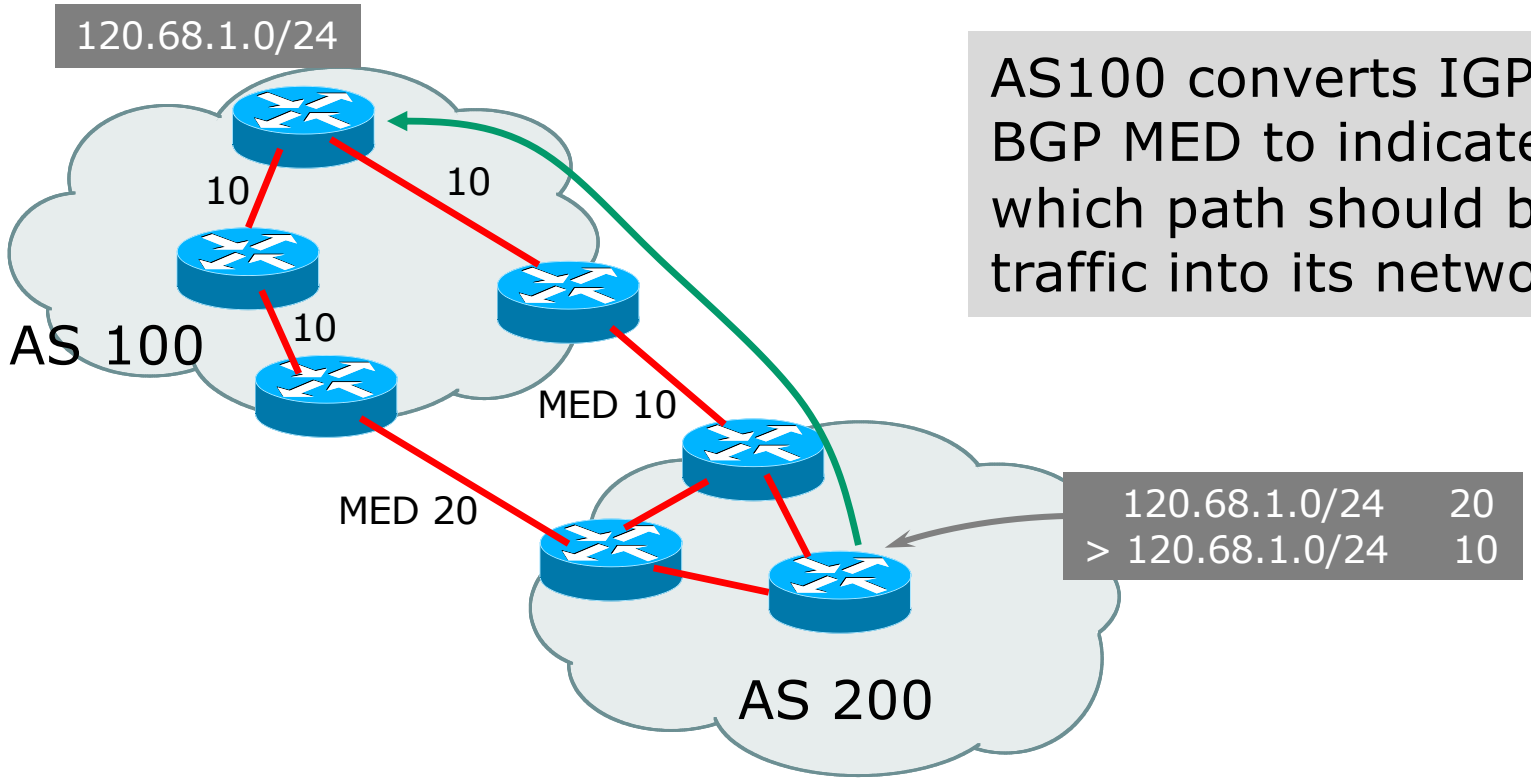
      ```
      bgp dynamic-med-interval <secs>
      ```

  - Also available in JunOS:

    ```
    set protocols bgp path-selection med-plus-igp
    ```

MED & IGP Metric

- Example: IGP metric conveyed as MED



AS100 converts IGP metric to BGP MED to indicate to AS200 which path should be used for traffic into its network

Weight

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

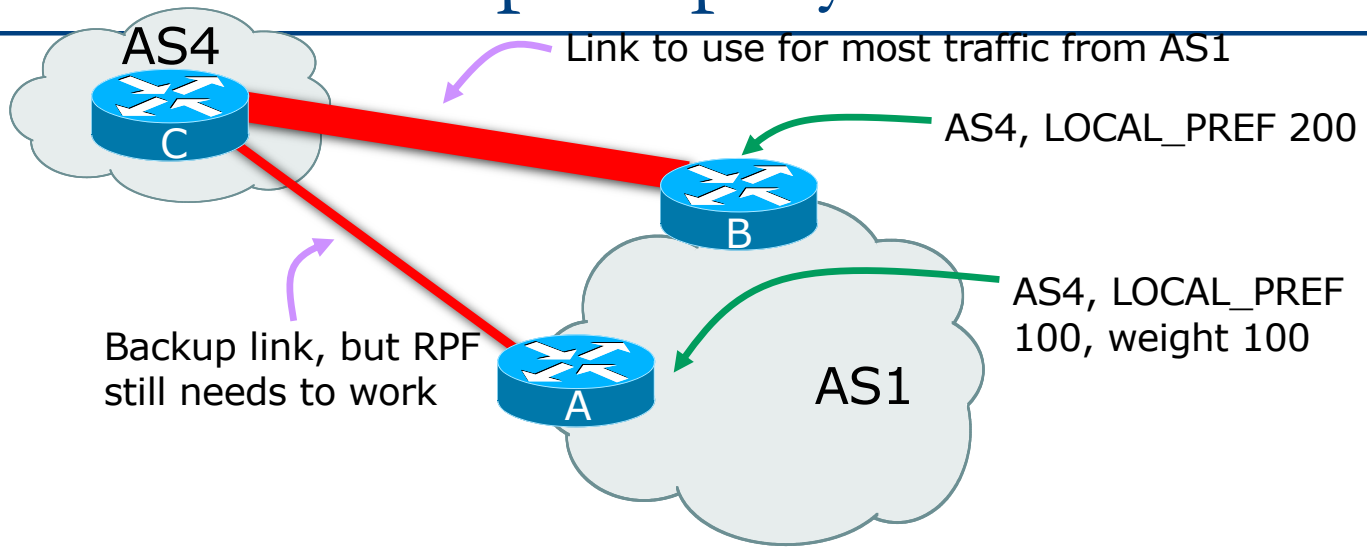
```
neighbor 120.5.7.1 weight 100
```

- Weight assigned to routes based on filter:

```
neighbor 120.5.7.3 filter-list 3 weight 50
```

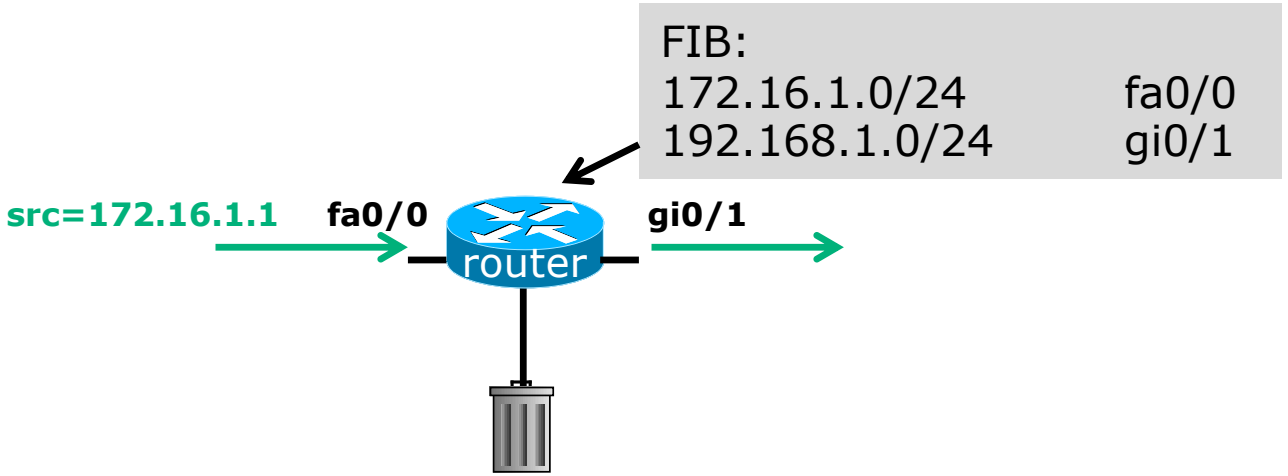
- Note: weight is not supported by every BGP implementation

Weight – Used to help Deploy RPF



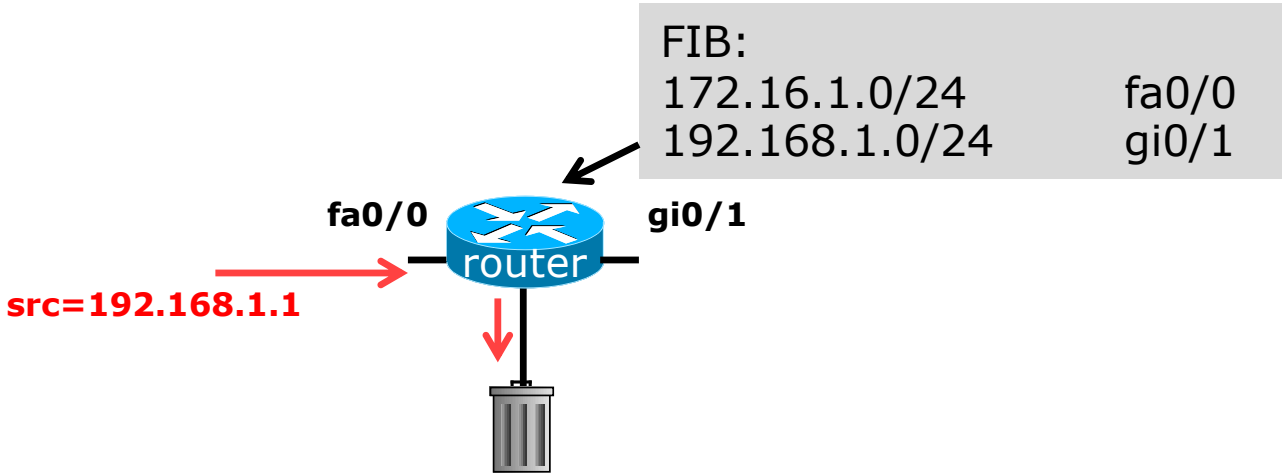
- Best path to AS4 from AS1 is always via B due to local-pref
- But packets arriving at A from AS4 over the direct C to A link will pass the RPF check as that path has a priority due to the weight being set
  - If weight was not set, best path back to AS4 would be via B, and the RPF check would fail

Aside: What is uRPF?



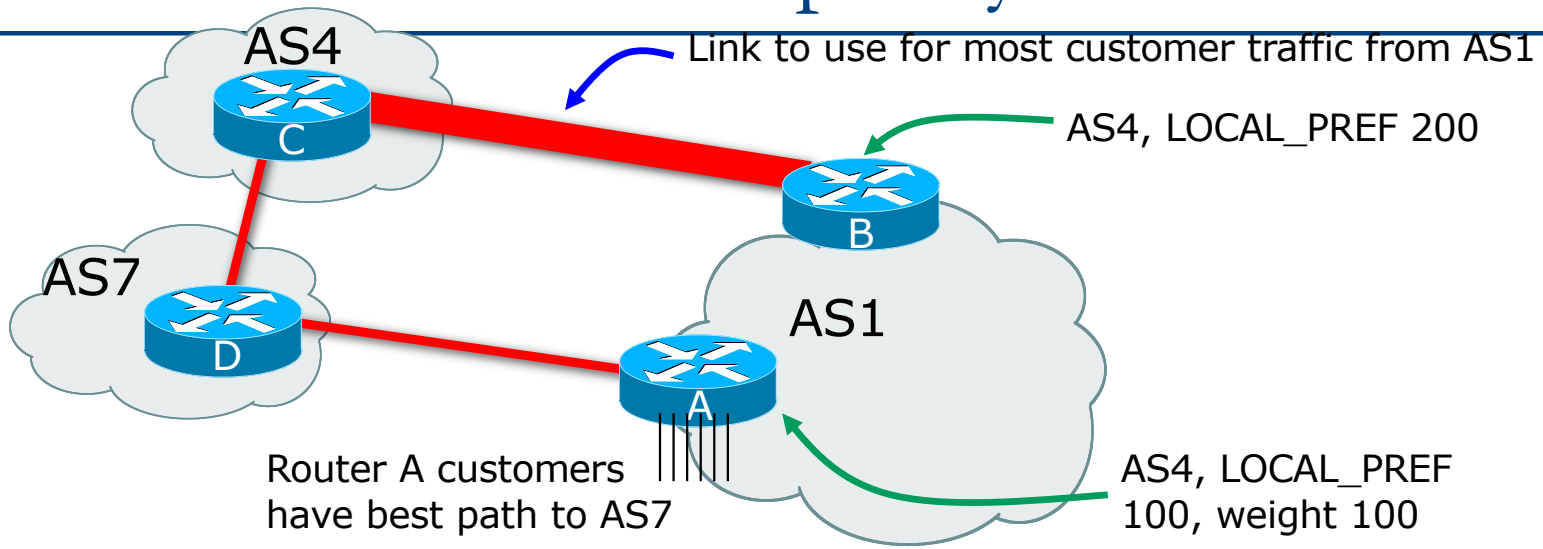
- Router compares source address of incoming packet with FIB entry
  - If FIB entry interface matches incoming interface, the packet is forwarded
  - If FIB entry interface does not match incoming interface, the packet is dropped

Aside: What is uRPF?



- 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

Weight – Used for traffic policy

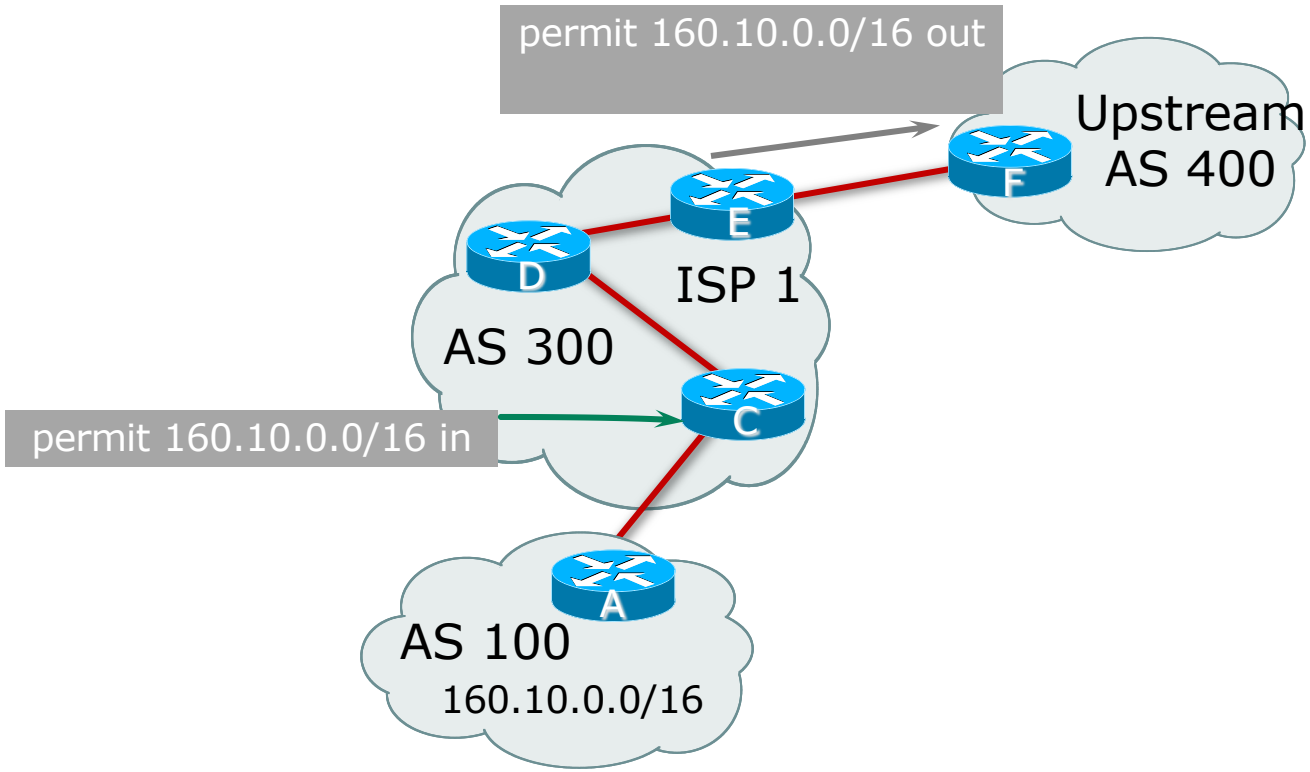


- Best path to AS4 from AS1 is always via B due to local-pref
- But customers connected directly to Router A use the link to AS7 as best outbound path because of the high weight applied to routes heard from AS7
  - If the A to D link goes down, then the Router A customers see best path via Router B and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

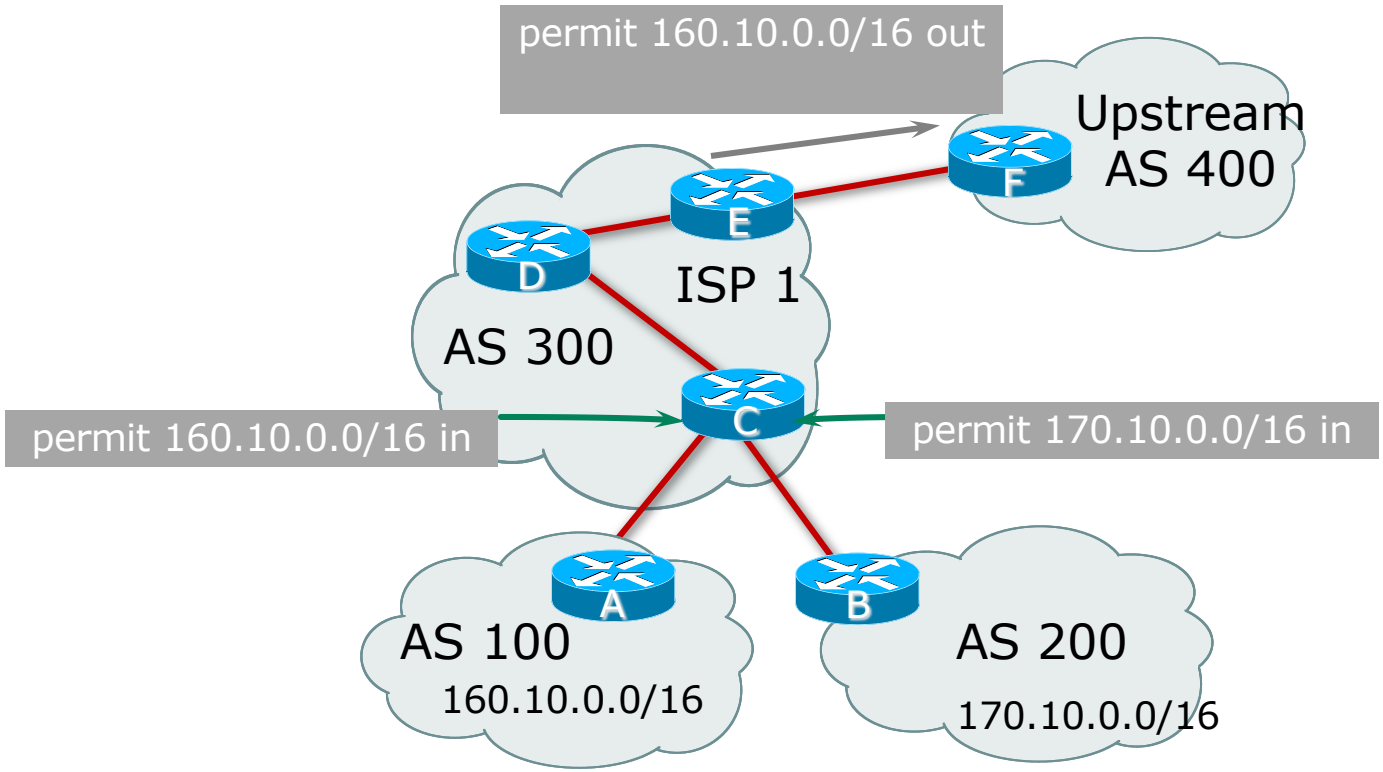
Community

- 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

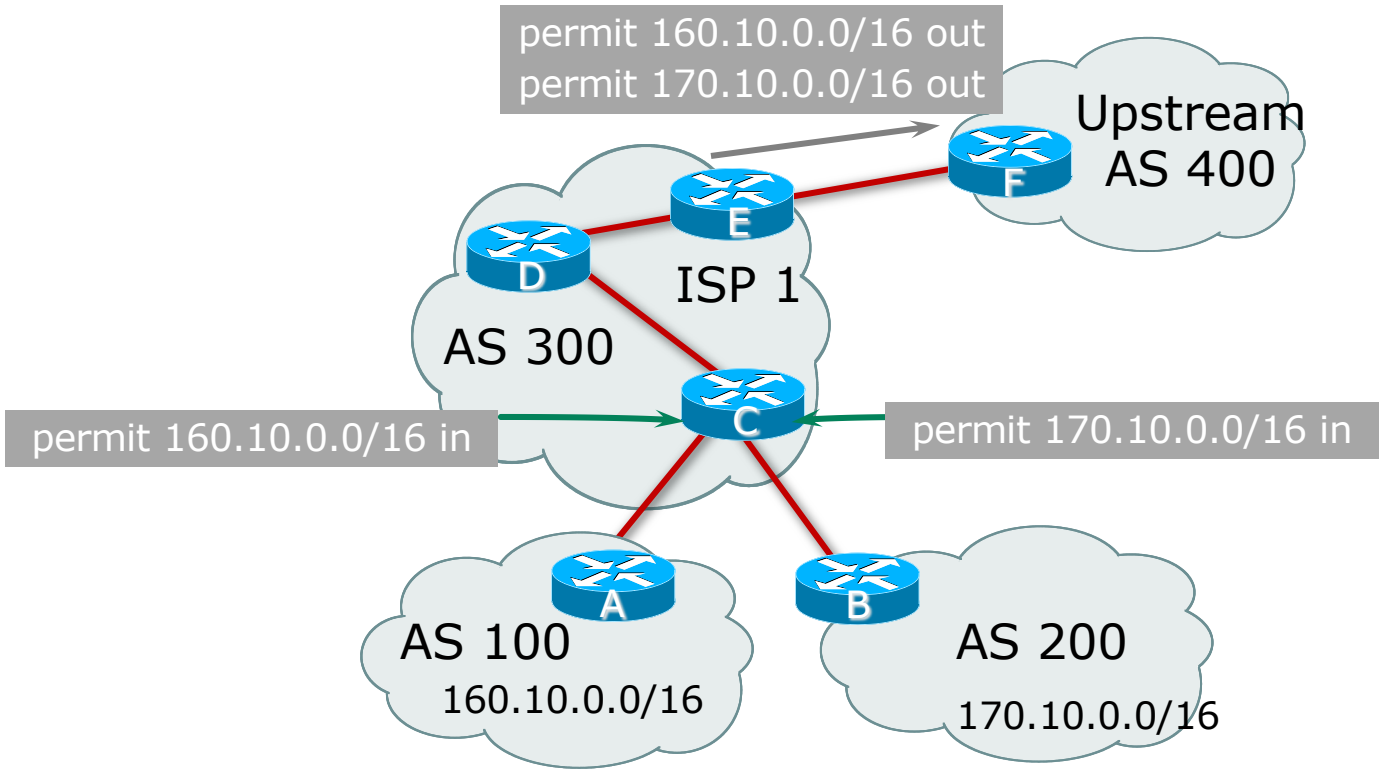
Community Example (before)



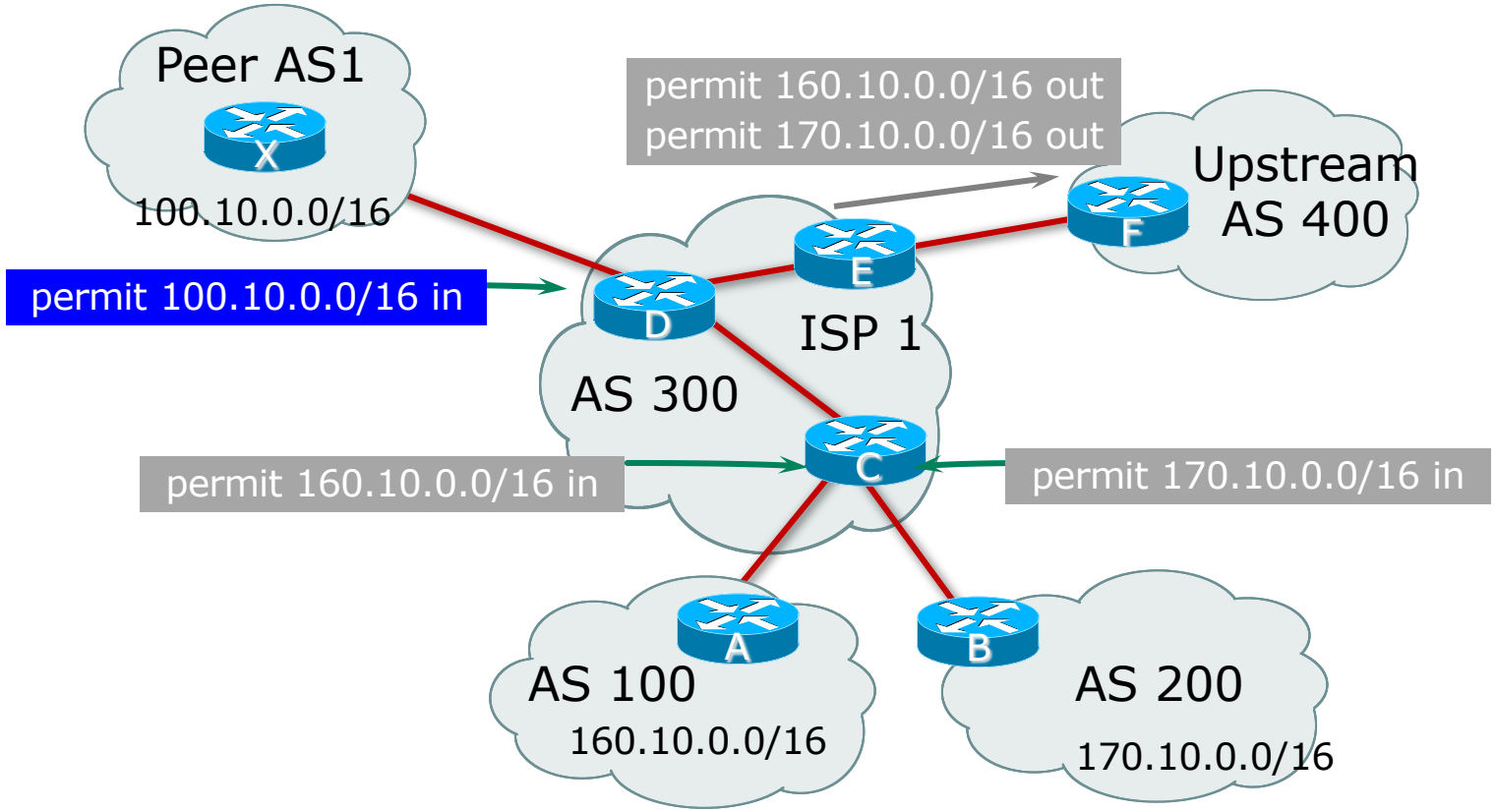
Community Example (before)



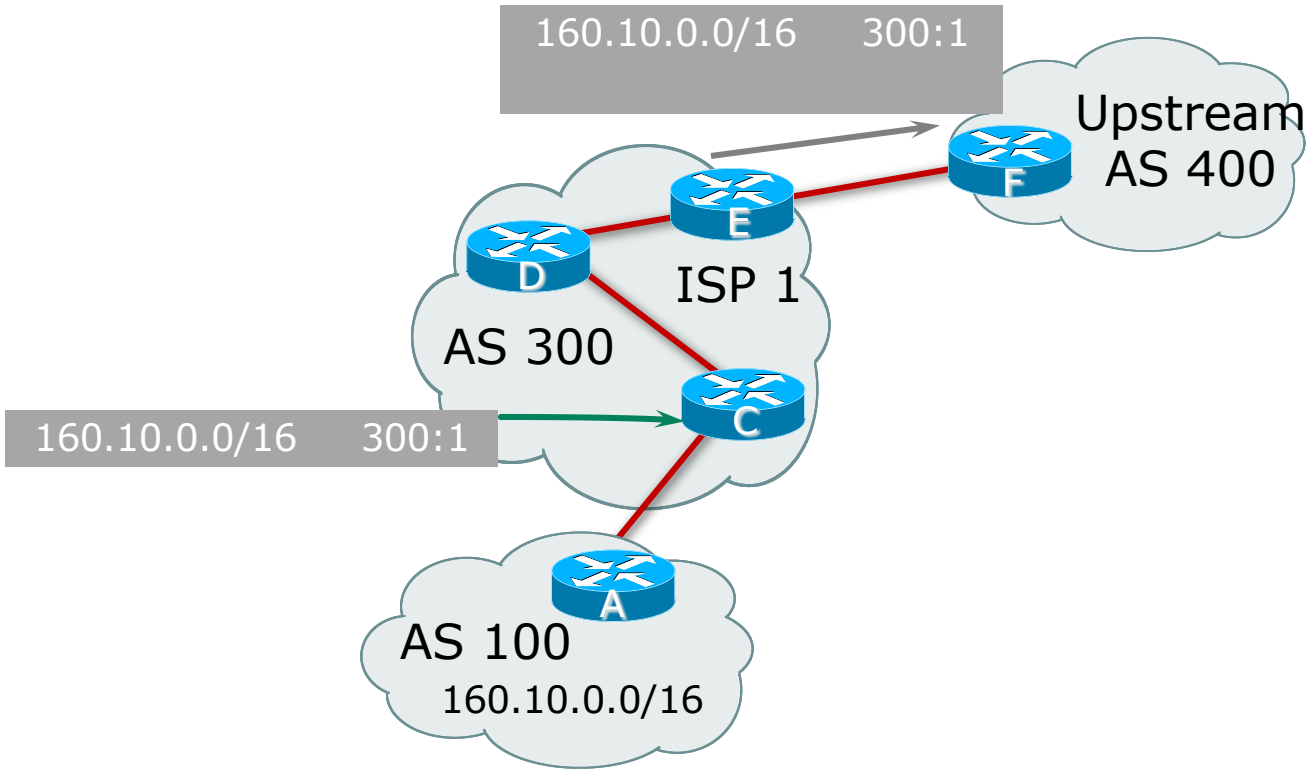
Community Example (before)



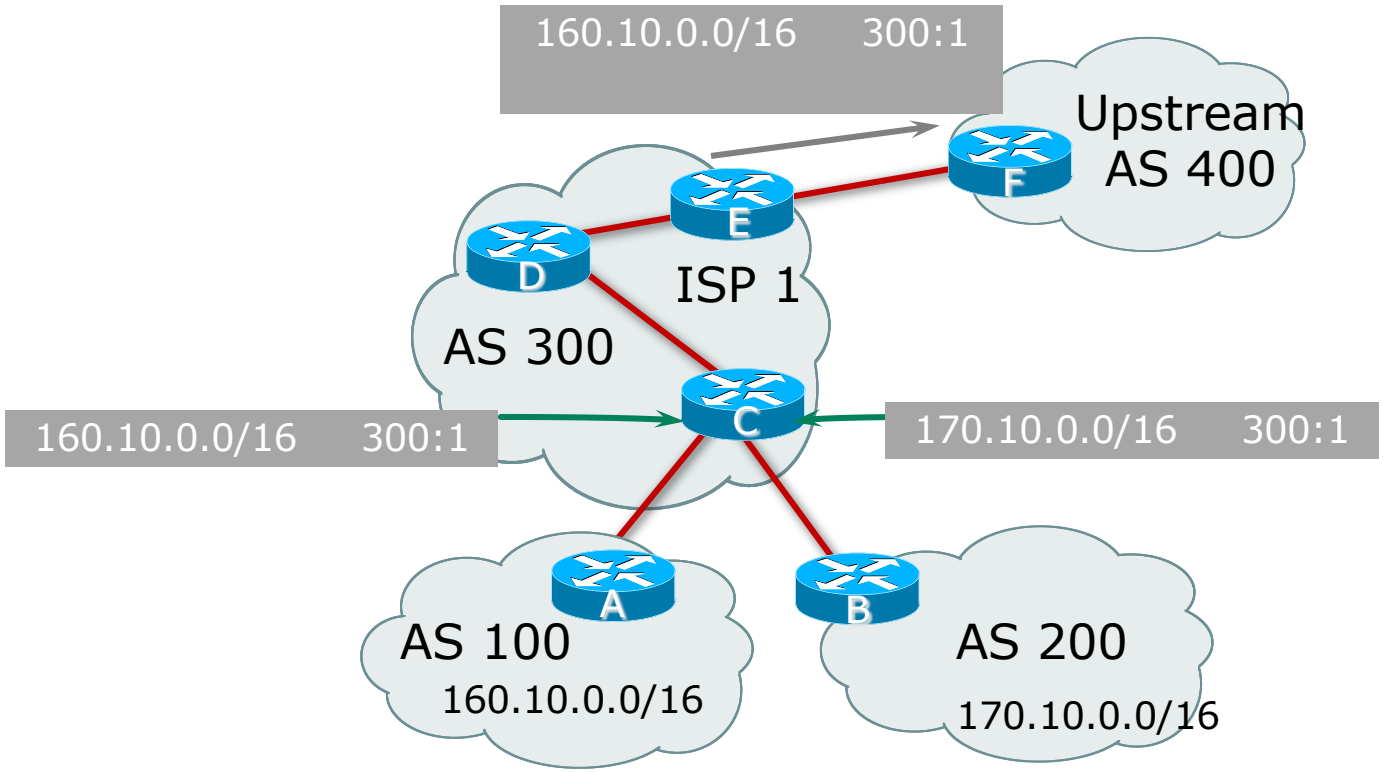
Community Example (before)



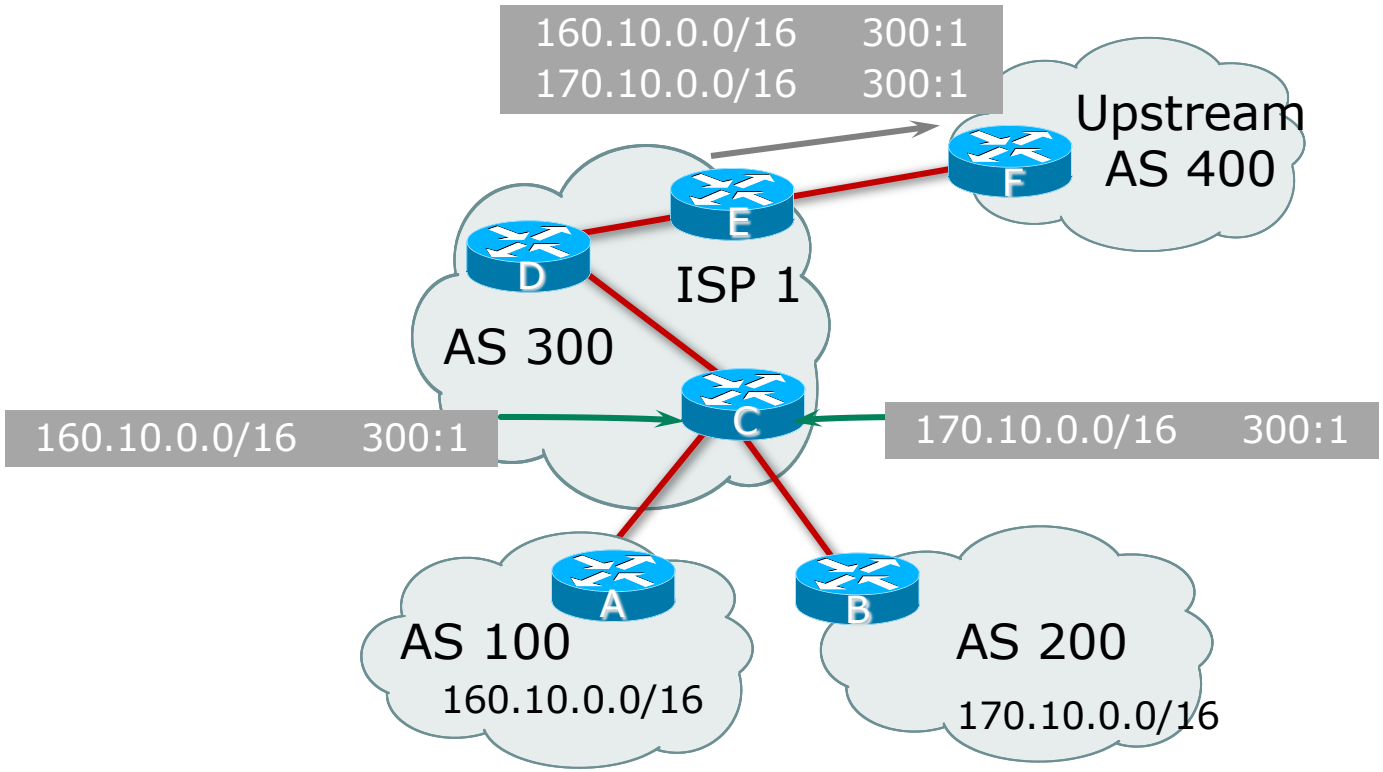
Community Example (after)



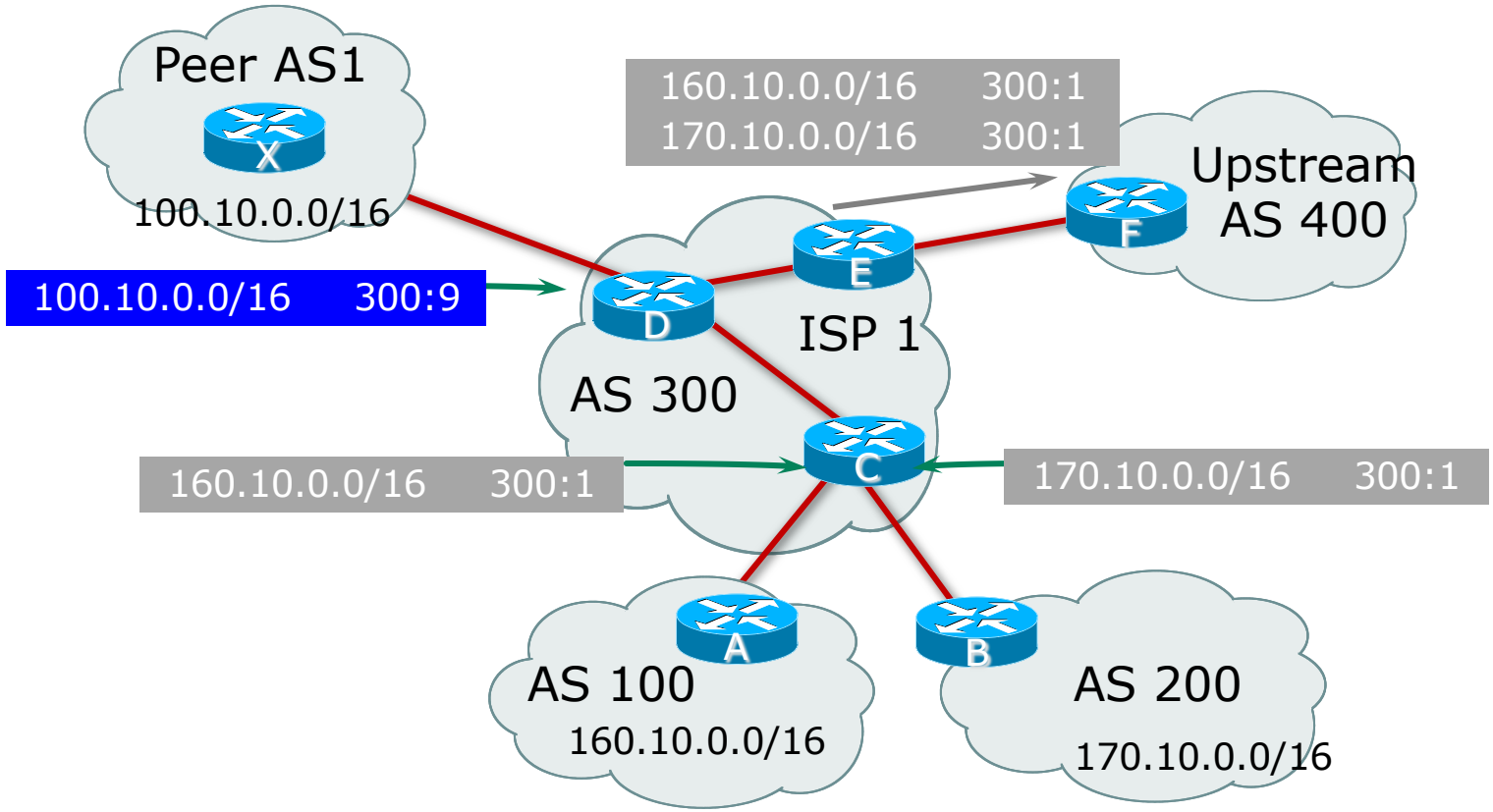
Community Example (after)



Community Example (after)



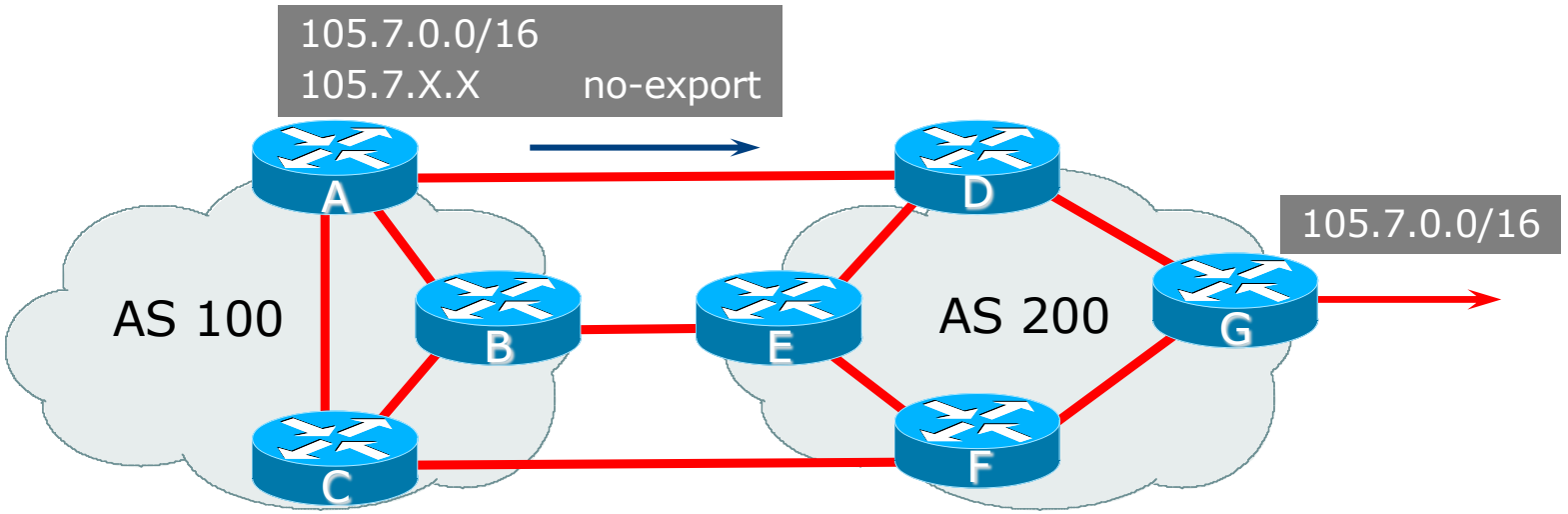
Community Example (after)



Well-Known Communities

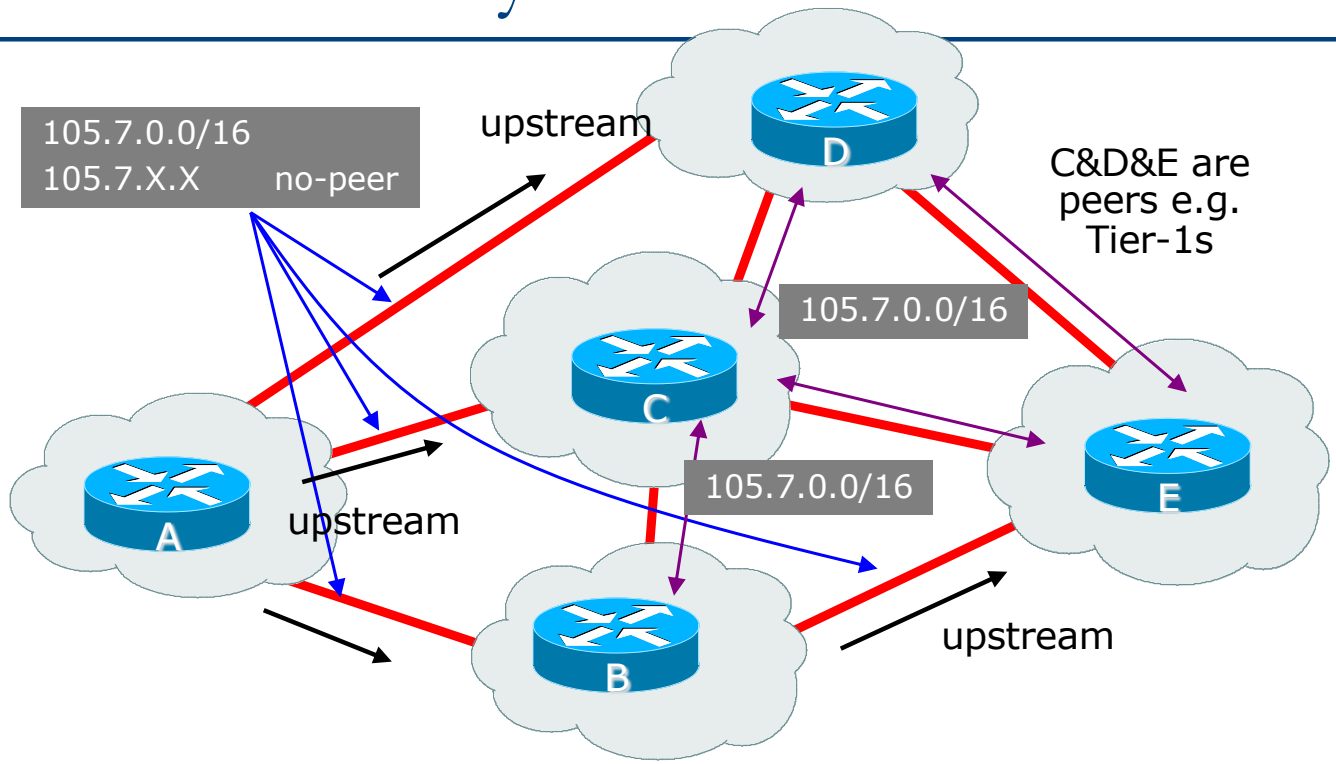
- Several well-known communities
  - www.iana.org/assignments/bgp-well-known-communities
- Six most common:
  - *no-export* 65535:65281
    - Do not advertise to any EBGP peers
  - *no-advertise* 65535:65282
    - Do not advertise to any BGP peer
  - *no-export-subconfed* 65535:65283
    - Do not advertise outside local AS (BGP confederations)
  - *no-peer* 65535:65284
    - Do not advertise to bi-lateral peers (RFC3765)
  - *blackhole* 65535:666
    - Null route the prefix (RFC7999)
  - *graceful-shutdown* 65535:0
    - Indicate imminent graceful shutdown (RFC8326)

No-Export Community



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

No-Peer Community



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

Vendor Policy implementation

- Be aware that each vendor has differing policy language behaviours for:
  - Treatment of well known communities
  - Setting communities
  - Removing communities
  - Replacing communities
- Consult:
  - Vendor documentation
  - https://www.rfc-editor.org/rfc/rfc8651.txt for discussion of some of the issues operators need to be aware of

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
  - RFC8092 – "BGP Large Communities"

BGP 'Large Community' Attribute

- New attribute designed to accommodate:
  - Local 32-bit ASN
  - Local Operator Defined Action (32-bits)
  - Remote Operator Defined Action (32-bits)
- This allows operators using 32-bit ASNs to peer with others using 32-bit ASNs and define policy actions
  - Compare with standard Communities which only accommodated 16-bit ASNs and 16-bits of action

BGP 'Large Community' Examples

- Some examples using common community conventions
  - (see BGP Community presentation for more detailed examples of typical ISP BGP Community policy)
  - 131072:3:131074
    - AS 131072 requests AS 131074 to do a three times prepend of this prefix on AS 131074's peerings
  - 131072:0:131074
    - AS 131072 requests AS 131074 not to announce this prefix

Summary
Attributes in Action

```
Router1>sh ip bgp
BGP table version is 16, local router ID is 10.10.15.241
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
 *>  10.10.0.0/26     0.0.0.0                  0         32768 i
 * i 10.10.0.0/20     10.10.15.226             0    100      0 i
 * i                  10.10.15.225             0    100      0 i
 *>                   0.0.0.0                  0         32768 i
 *>i 10.10.0.64/26    10.10.15.225             0    100      0 i
 *>i 10.10.0.128/26   10.10.15.226             0    100      0 i
 * i 10.20.0.0/26     10.10.15.226             0    100      0 20 i
 *>i                  10.10.15.225             0    100      0 20 i
 * i 10.20.0.0/20     10.10.15.226             0    100      0 20 i
 *>i                  10.10.15.225             0    100      0 20 i
```

BGP Path Selection Algorithm

Why is this the best path?

BGP Path Selection Algorithm: Part One

1. Do not consider path if no route to next hop
2. Do not consider IBGP path if not synchronised (historical)
3. Highest weight (local to router)
4. Highest local preference (global within AS)
5. Prefer locally originated route
6. Shortest AS path
7. Lowest origin code
   - IGP < EGP < incomplete

BGP Path Selection Algorithm: Part Two

8. Lowest Multi-Exit Discriminator (MED)
   - Cisco IOS: if bgp deterministic-med, order the paths by AS number before comparing
   - Cisco IOS: if bgp always-compare-med, then compare for all paths
   - Otherwise only consider MEDs if paths are from the same neighbouring AS
9. Prefer EBGP path over IBGP path
10. Path with lowest IGP metric to next-hop

BGP Path Selection Algorithm: Part Three

11. For EBGP paths:
    - Cisco IOS: if multipath is enabled, install N parallel paths in forwarding table
    - If router-id is the same, go to next step
    - Cisco IOS: if router-id is not the same, select the oldest path
        - To turn off (to follow RFC): bgp bestpath compare-routerid
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

ISP Workshops