BGP Techniques for Network Operators

Philip Smith
<philip@nsrc.org>
NANOG92
Toronto, Canada

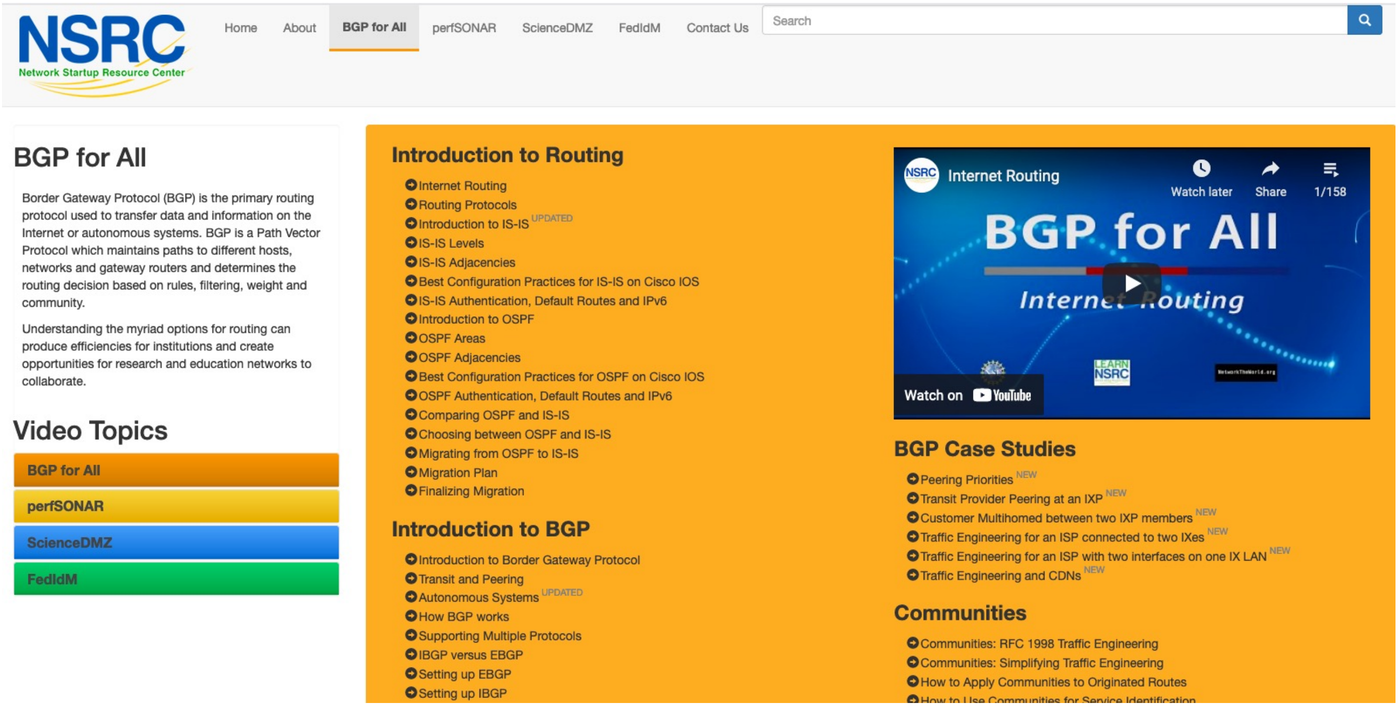
Last updated 7th October 2024

Presentation Slides

- Will be available on:
  - https://bgp4all.com/pfs/conferences/
  - And on the NANOG92 website
- Feel free to ask questions any time

BGP Videos

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



Background

- The hierarchy of Routing Protocols
  1. How do routers inside a network ”find each other”?
     - Interior routing protocol (called IGP)
     - Examples in common use: ISIS & OSPF
  2. How do routers inside a network share globally reachable destinations?
     - By the use of BGP (called IBGP)
  3. How do routers in the networks making up the Internet share globally reachable destinations?
     - By the use of BGP (called EBGP)
- Tutorial focus on the latter two circumstances

BGP Techniques for Network Operators

- BGP Basics
- Scaling BGP
- Using Communities
- Deploying BGP in a Service Provider Network

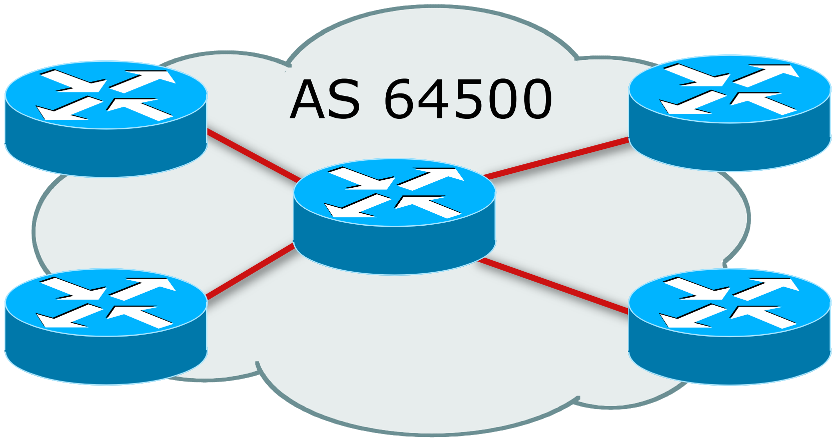
BGP Basics

What is BGP?

Border Gateway Protocol

- A Routing Protocol used to exchange routing information between different networks
  - Exterior gateway protocol
- Described in RFC4271
  - RFC4276 gives an implementation report on BGP
  - RFC4277 describes operational experiences using BGP
- The Autonomous System is the cornerstone of BGP
  - It is used to uniquely identify networks with a common routing policy

Autonomous System (AS)



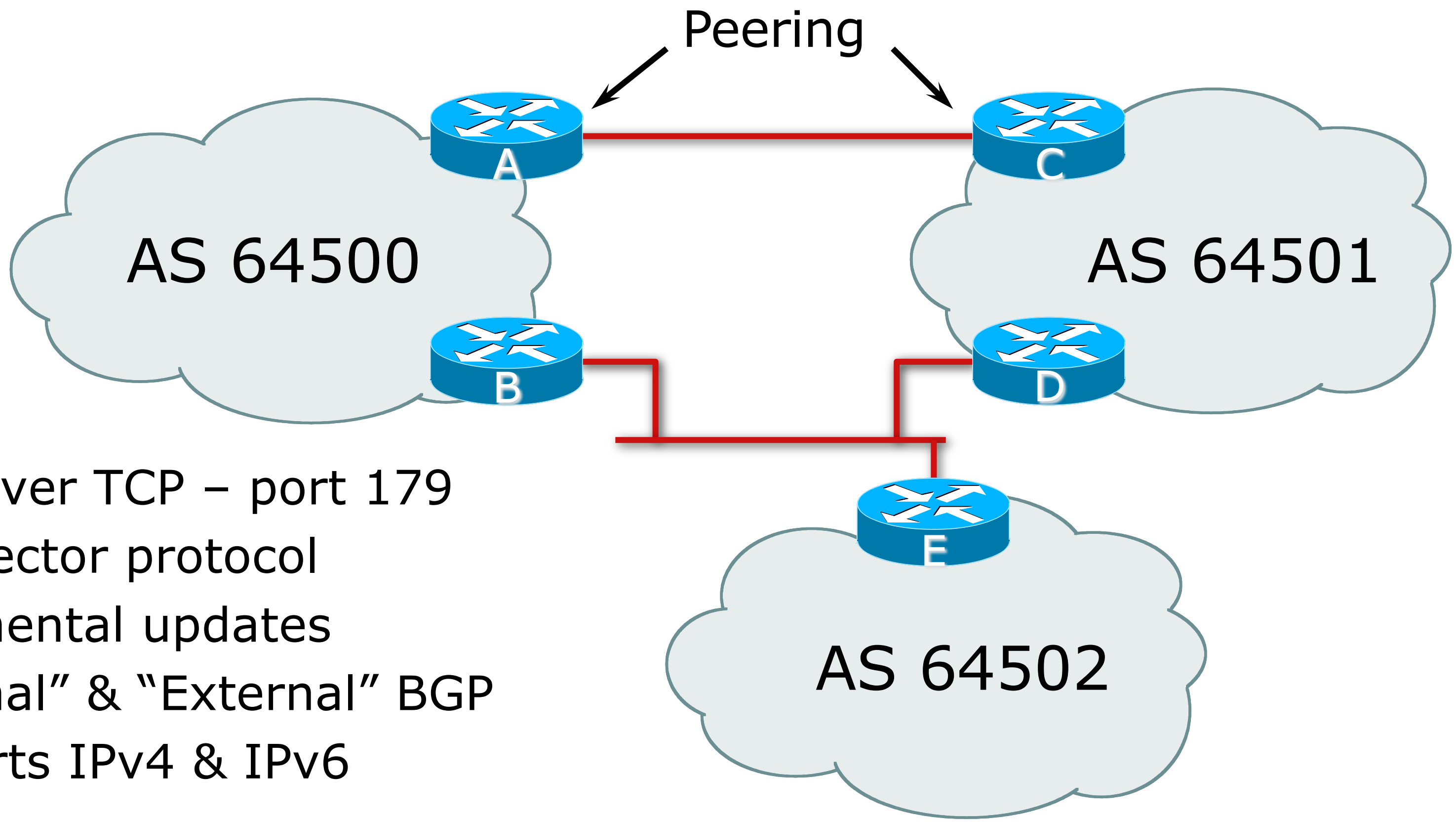
- Collection of networks with same routing policy
- Single routing protocol
- Usually under single ownership, trust and administrative control
- Identified by a unique 32-bit integer (ASN)

Autonomous System Number

Range:	
0-4294967295	(32-bit range – RFC6793)
	(0-65535 was original 16-bit range)
Usage:	
0 and 65535	(IANA Reserved)
1-64495	(public Internet)
64496-64511	(documentation – RFC5398)
64512-65534	(private use only)
23456	(represent 32-bit range in 16-bit world)
65536-65551	(documentation – RFC5398)
65552-131071	(IANA Reserved)
131072-458751	(public Internet)
458752-4199999999	(IANA Reserved/Unallocated)
4200000000-4294967294	(private use only – RFC6996)
4294967295	(IANA Reserved – RFC7300)

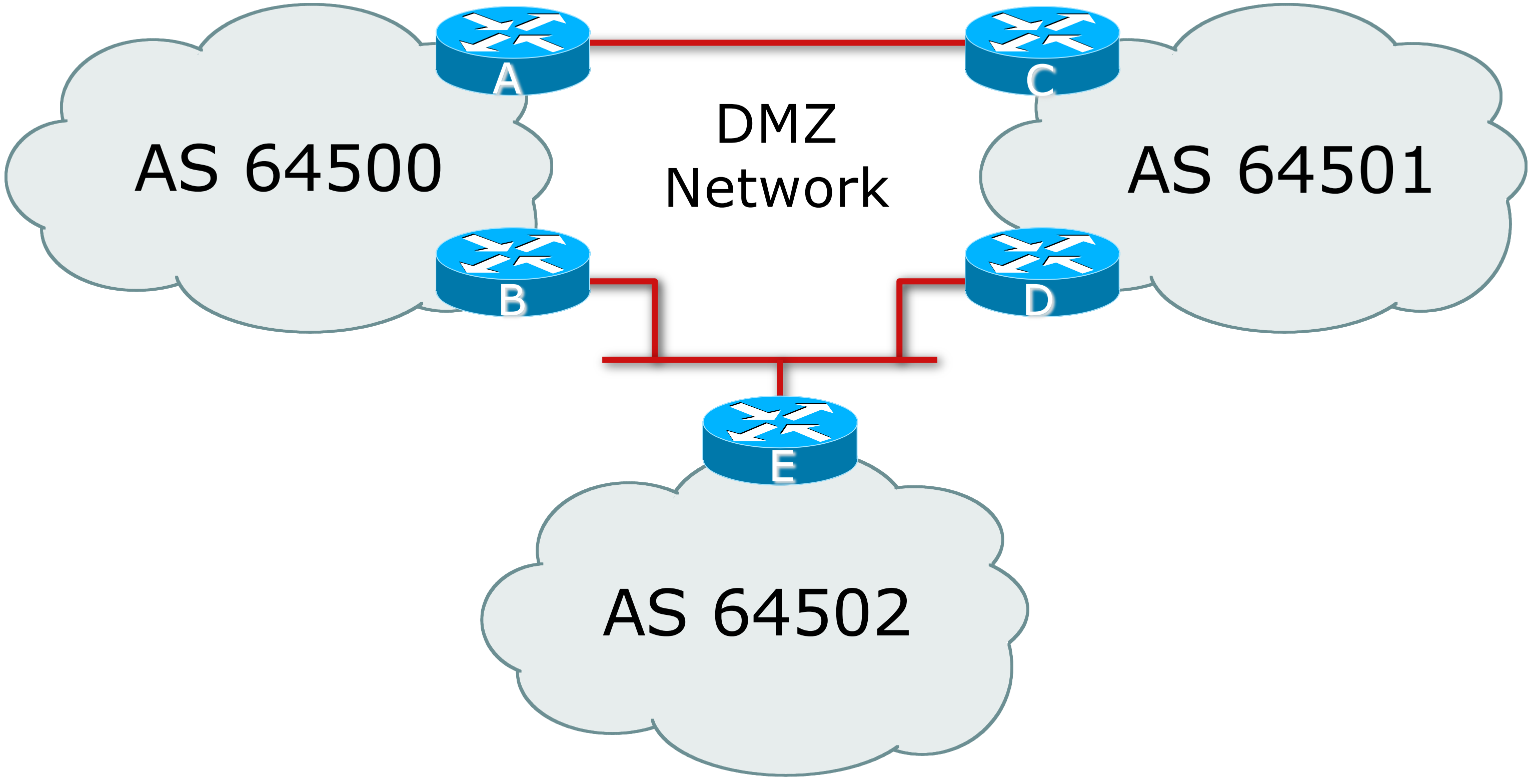
- 32-bit range representation specified in RFC5396
  - Defines "asplain" (traditional format) as standard notation

BGP Basics



- Runs over TCP – port 179
- Path vector protocol
- Incremental updates
- “Internal” & “External” BGP
- Supports IPv4 & IPv6

Demarcation Zone (DMZ)



- DMZ is the link or network shared between ASes

BGP General Operation

- Learns multiple paths via internal and external BGP speakers
- Picks the best path and installs it in the routing table (RIB)
- Best path is sent to external BGP neighbours
- Policies are applied by influencing the best path selection

Supporting Multiple Protocols

- RFC4760
  - Defines Multi-protocol Extensions for BGP4
  - Enables BGP to carry routing information of protocols other than IPv4
    - e.g. MPLS, IPv6, Multicast etc
  - Exchange of multiprotocol NLRI must be negotiated at session startup
- RFC2545
  - Use of BGP Multiprotocol Extensions for IPv6 Inter-Domain Routing
  - Address family for IPv6

Supporting Multiple Protocols

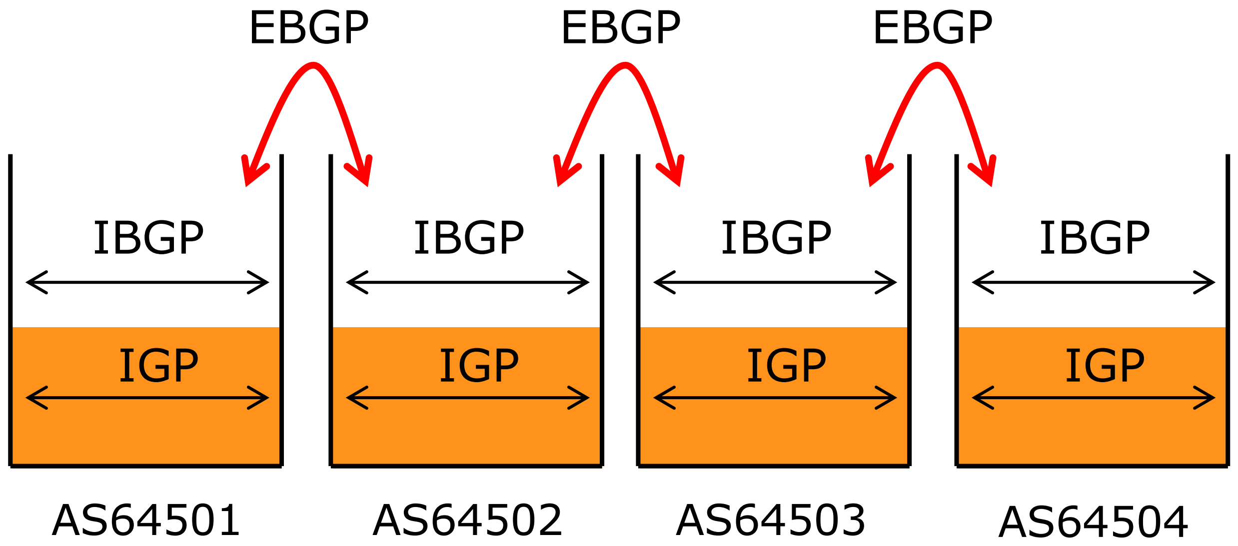
- Independent operation
  - One RIB per protocol
    - IPv6 routes in BGP’s IPv6 RIB
    - IPv4 routes in BGP’s IPv4 RIB
  - Each protocol can have its own policies
- NEXTHOP
  - The IP address of the next router must belong to the same address family as that of the local router

EBGP & IBGP

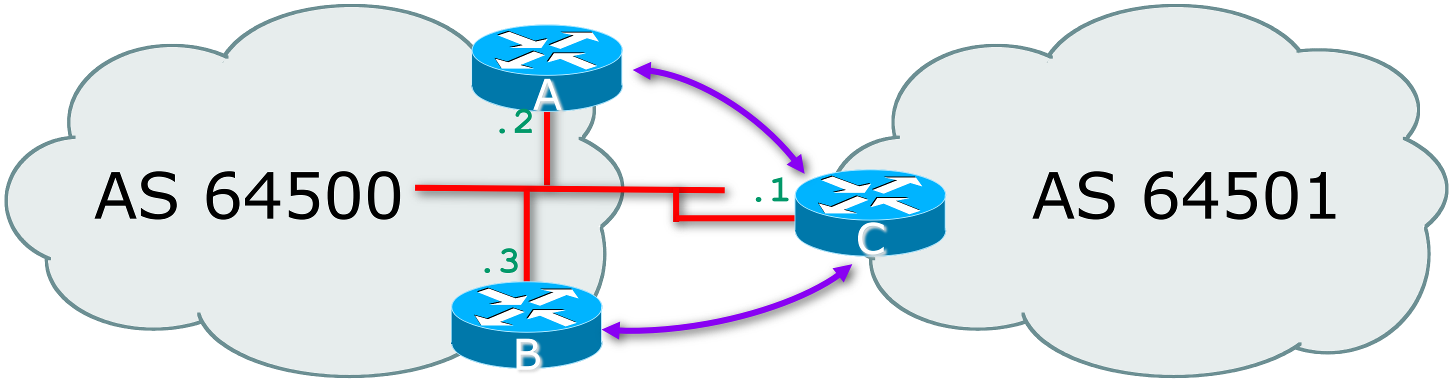
- BGP is used
  - Internally (IBGP)
  - Externally (EBGP)
- IBGP used to carry
  - Some/all Internet prefixes across Service Provider backbone
  - Service Provider’s customer prefixes
- EBGP used to
  - Exchange prefixes with other ASes
  - Implement routing policy

BGP/IGP model used in Service Provider networks

- Model representation



External BGP Peering (EBGP)

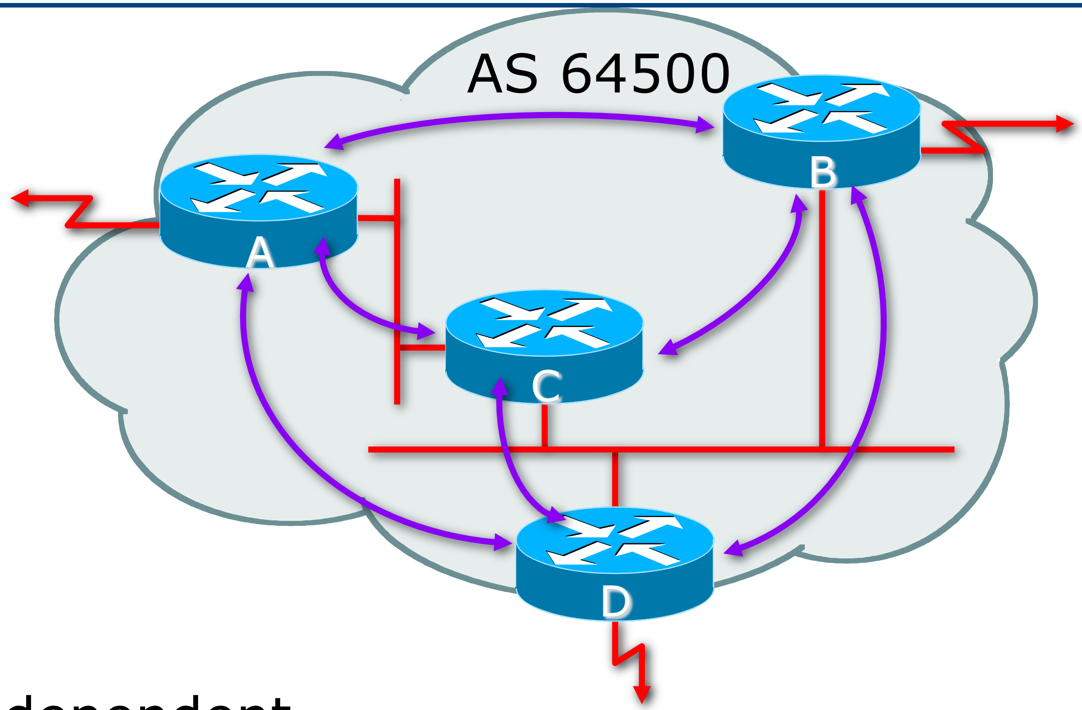


- Between BGP speakers in different AS
- Should be directly connected
- **Never** run an IGP between EBGP peers

Internal BGP (IBGP)

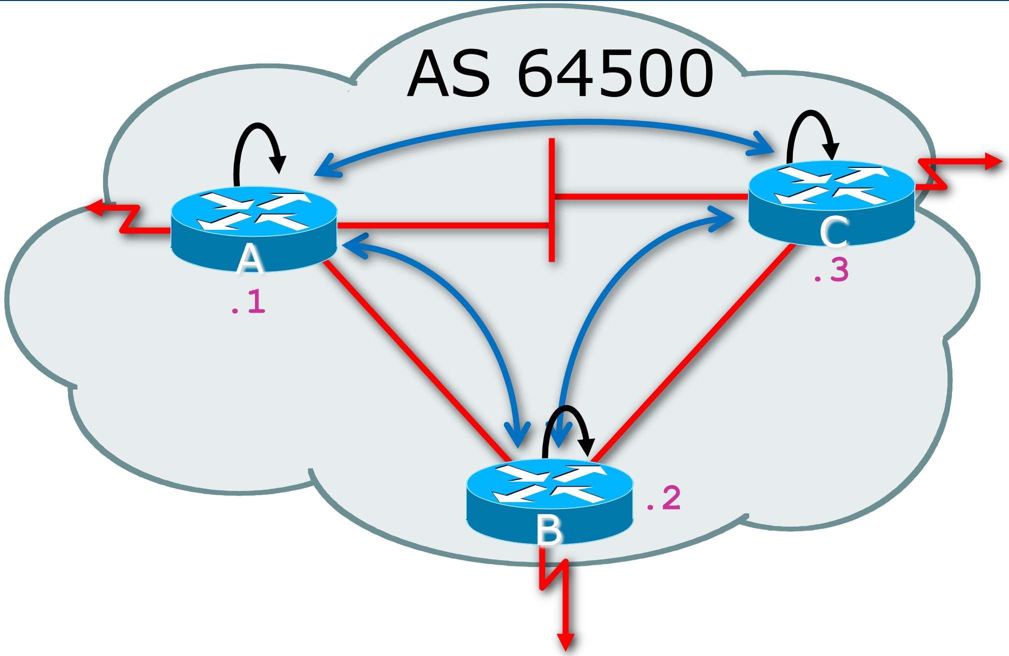
- BGP peer within the same AS
- Not required to be directly connected
  - IGP takes care of inter-BGP speaker connectivity
- IBGP speakers must be fully meshed:
  - They originate connected networks
  - They pass on prefixes learned from outside the AS
  - They do not pass on prefixes learned from other IBGP speakers

Internal BGP Peering (IBGP)



- Topology independent
- Each IBGP speaker must peer with every other IBGP speaker in the AS as per ⟷

Peering between Loopback Interfaces



- Peer with loop-back interface
  - Loop-back interface does not go down – ever!
- Do not want IBGP session to depend on state of a single interface or the physical topology

Summary
BGP neighbour status (Cisco IOS IPv4)

```
Router6>show ip bgp summary
BGP router identifier 10.0.15.246, local AS number 10
BGP table version is 16, main routing table version 16
7 network entries using 819 bytes of memory
14 path entries using 728 bytes of memory
2/1 BGP path/bestpath attribute entries using 248 bytes of memory
0 BGP route-map cache entries using 0 bytes of memory
0 BGP filter-list cache entries using 0 bytes of memory
BGP using 1795 total bytes of memory
BGP activity 7/0 prefixes, 14/0 paths, scan interval 60 secs

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.0.15.241     4    10       9       8       16    0    0 00:04:47        2
10.0.15.242     4    10       6       5       16    0    0 00:01:43        2
10.0.15.243     4    10       9       8       16    0    0 00:04:49        2
...
```

BGP Version

Updates sent and received

Updates waiting

Summary
BGP neighbour status (Cisco IOS IPv6)

```
Router1>sh bgp ipv6 unicast summary
BGP router identifier 10.10.15.224, local AS number 10
BGP table version is 28, main routing table version 28
18 network entries using 2880 bytes of memory
38 path entries using 3040 bytes of memory
9/6 BGP path/bestpath attribute entries using 1152 bytes of memory
4 BGP AS-PATH entries using 96 bytes of memory
0 BGP route-map cache entries using 0 bytes of memory
0 BGP filter-list cache entries using 0 bytes of memory
BGP using 7168 total bytes of memory
BGP activity 37/1 prefixes, 95/19 paths, scan interval 60 secs

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
2001:DB8::2     4    10     185     182       28    0    0 02:36:11       16
2001:DB8::3     4    10     180     181       28    0    0 02:36:08       11
2001:DB8:0:4::1 4    40     153     152       28    0    0 02:05:39        9
```

Neighbour Information   BGP Messages Activity

Summary
BGP neighbour status (JunOS)

```
philip@R6> show bgp summary
Groups: 1 Peers: 14 Down peers: 0
Table          Tot Paths  Act Paths Suppressed    History Damp State    Pending
inet.0                20         20          0          0          0          0
inet6.0               20         20          0          0          0          0

Peer            AS      InPkt     OutPkt    OutQ   Flaps Last Up/Dwn  State|#Active/Received/Accepted/Damped..
10.0.15.241     10    1067980     202487       0       0 9w1d 4:32:05 Establ   inet.0: 10/10/10/0
10.0.15.242     10     204577    1001705       0       0 9w1d 4:32:09 Establ   inet.0: 3/3/3/0
10.0.15.243     10     277630    1886656       0       0 9w1d 4:32:06 Establ   inet.0: 4/4/4/0
...
2001:DB8::1     10     416832     202568       0       0 9w1d 4:30:46 Establ   inet6.0: 10/10/10/0
2001:DB8::2     10     204605     411166       0       0 9w1d 4:34:47 Establ   inet6.0: 3/3/3/0
2001:DB8::3     10     277568     729073       0       0 9w1d 1:03:31 Establ   inet6.0: 2/2/2/0
...
```

AS Number

Updates sent and received

Updates waiting

Address Family

Summary
BGP Table (Cisco IOS IPv4)

```
Router6>sh ip bgp
BGP table version is 18, 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
 *>i 10.0.2.128/26    10.0.15.251              0    100      0 i
 *>i 10.0.2.192/26    10.0.15.252              0    100      0 i
 *>i 10.0.3.0/26      10.0.15.253              0    100      0 i
 *>i 10.0.3.64/26     10.0.15.254              0    100      0 i
```

Summary
BGP Table (Cisco IOS IPv6)

```
Router6>sh bgp ipv6 unicast
BGP table version is 18, 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 2001:DB8:1::/48  2001:DB8::1              0    100      0 i
 *>i 2001:DB8:2::/48  2001:DB8::2              0    100      0 i
 *>i 2001:DB8:3::/48  2001:DB8::3              0    100      0 i
 *>i 2001:DB8:4::/48  2001:DB8::4              0    100      0 i
 *>i 2001:DB8:5::/48  2001:DB8::5              0    100      0 i
 *>  2001:DB8:6::/48  ::                       0         32768 i
 *>i 2001:DB8:7::/48  2001:DB8::7              0    100      0 i
 *>i 2001:DB8:8::/48  2001:DB8::8              0    100      0 i
 *>i 2001:DB8:9::/48  2001:DB8::9              0    100      0 i
 *>i 2001:DB8:A::/48  2001:DB8::A              0    100      0 i
 *>i 2001:DB8:B::/48  2001:DB8::B              0    100      0 i
 *>i 2001:DB8:C::/48  2001:DB8::C              0    100      0 i
 *>i 2001:DB8:D::/48  2001:DB8::D              0    100      0 i
 *>i 2001:DB8:E::/48  2001:DB8::E              0    100      0 i
```

Summary
BGP Table (JunOS)

```
philip@R6> show route protocol bgp terse

inet.0: 14 destinations, 14 routes (14 active, 0 holddown, 0 hidden)
+ = Active Route, - = Last Active, * = Both

A V Destination         P Prf   Metric 1   Metric 2  Next hop        AS path
  ? 10.0.0.0/26         B 100                                        I
  unverified                                        >10.0.15.241
  ? 10.0.0.64/26        B 100                                        I
  unverified                                        >10.0.15.241
...
  ? 10.1.0.0/24         B 100                                        20 I
  unverified                                        >10.0.15.242
  ? 10.4.0.0/24         B 100                                        20 I
  unverified                                        >10.0.15.241
...
inet6.0: 14 destinations, 14 routes (14 active, 0 holddown, 0 hidden)
+ = Active Route, - = Last Active, * = Both

A V Destination         P Prf   Metric 1   Metric 2  Next hop        AS path
  ? 2001:DB8:1::/48     B 100                                        I
  unverified                                        >fe80::82ac:acff:fed2:ea88
  ? 2001:DB8:2::/48     B 100                                        I
  unverified                                        >fe80::82ac:acff:fed2:ea88
...
  ? 2001:DB9::/32       B 100                                        20 I
  unverified                                        >fe80::224e:71ff:fe90:2500
  ? 2001:DB9::/32       B 100                                        20 I
  unverified                                        >fe80::82ac:acff:fed2:ea88
...
```

BGP Attributes

BGP’s policy tool kit

What is an Attribute?

…	Origin	AS Path	Next Hop	MED	…

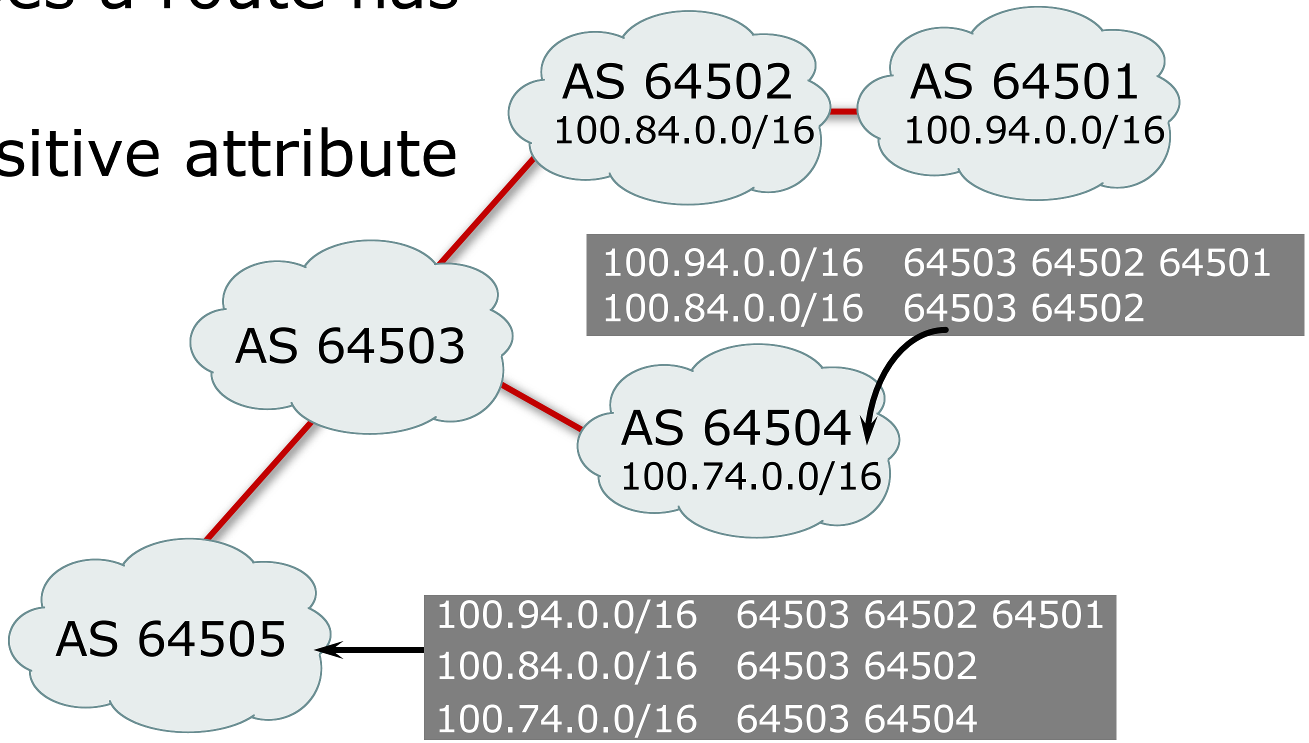
- Part of a BGP Update
- Describes the characteristics of prefix
- Can either be transitive or non-transitive
- Some are mandatory

BGP Attributes

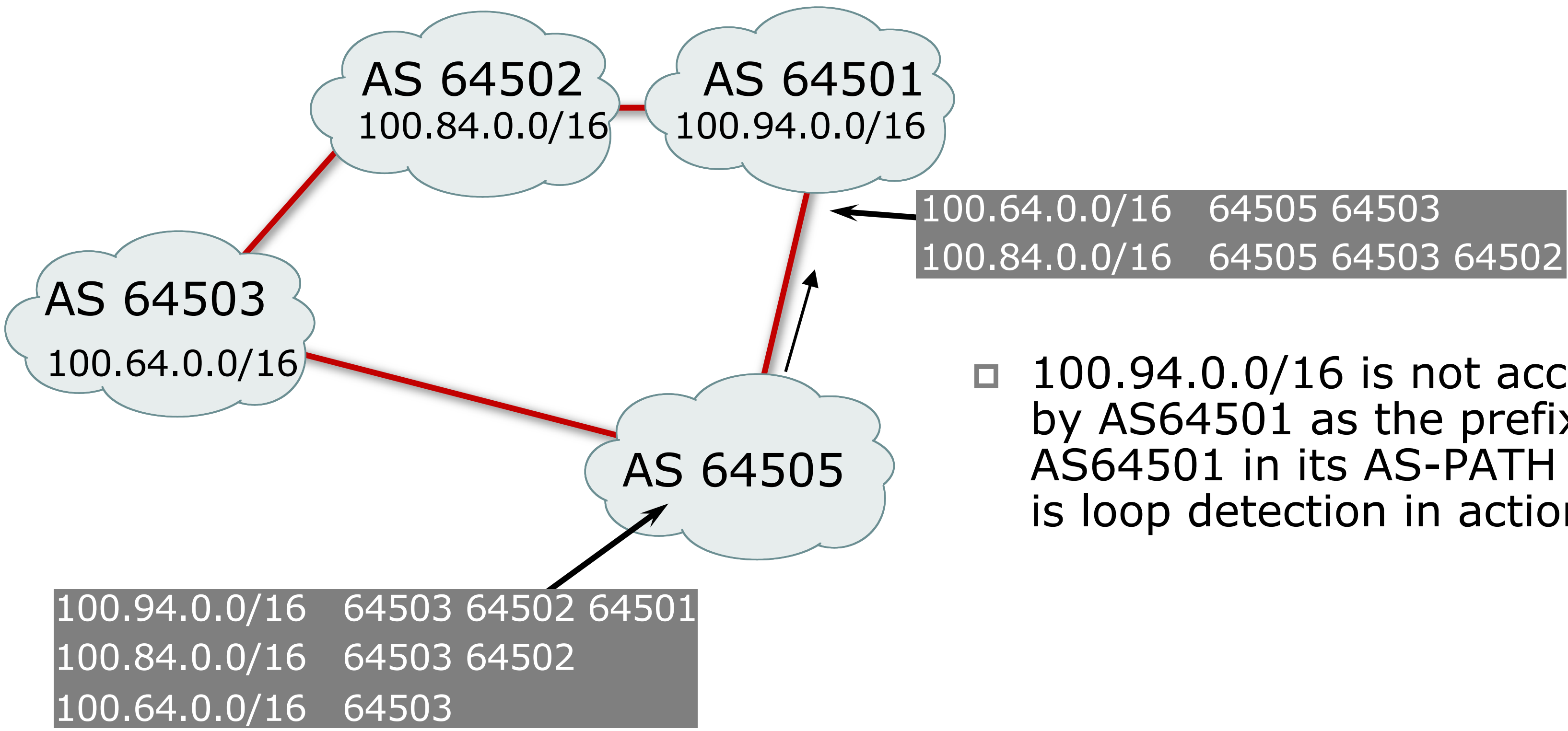
- Carry various information about or characteristics of the prefix being propagated
  - AS-PATH
  - NEXT-HOP
  - ORIGIN
  - AGGREGATOR
  - LOCAL_PREFERENCE
  - Multi-Exit Discriminator
  - (Weight)
  - COMMUNITY

AS-Path

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

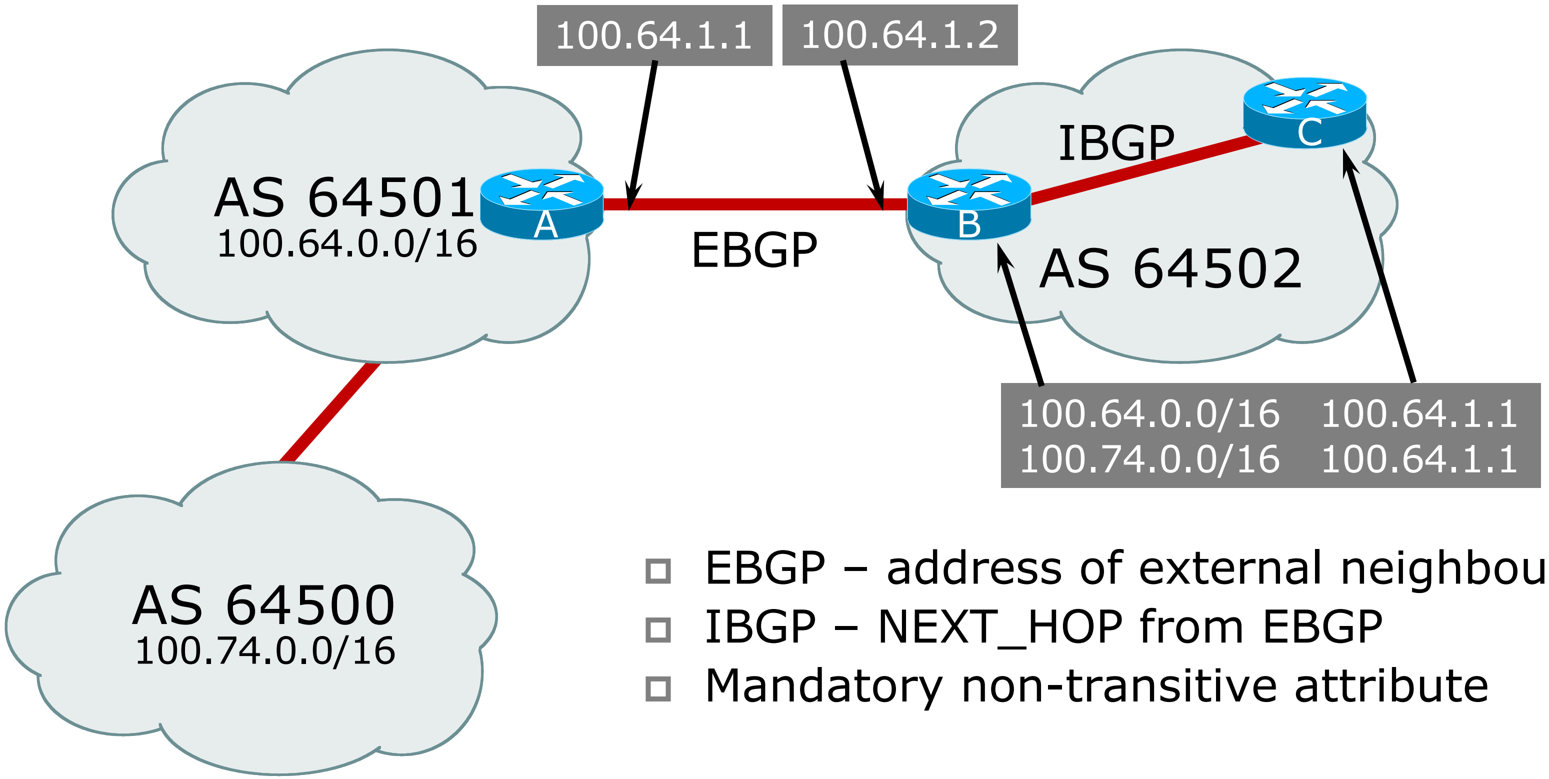


AS-Path loop detection



- 100.94.0.0/16 is not accepted by AS64501 as the prefix has AS64501 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

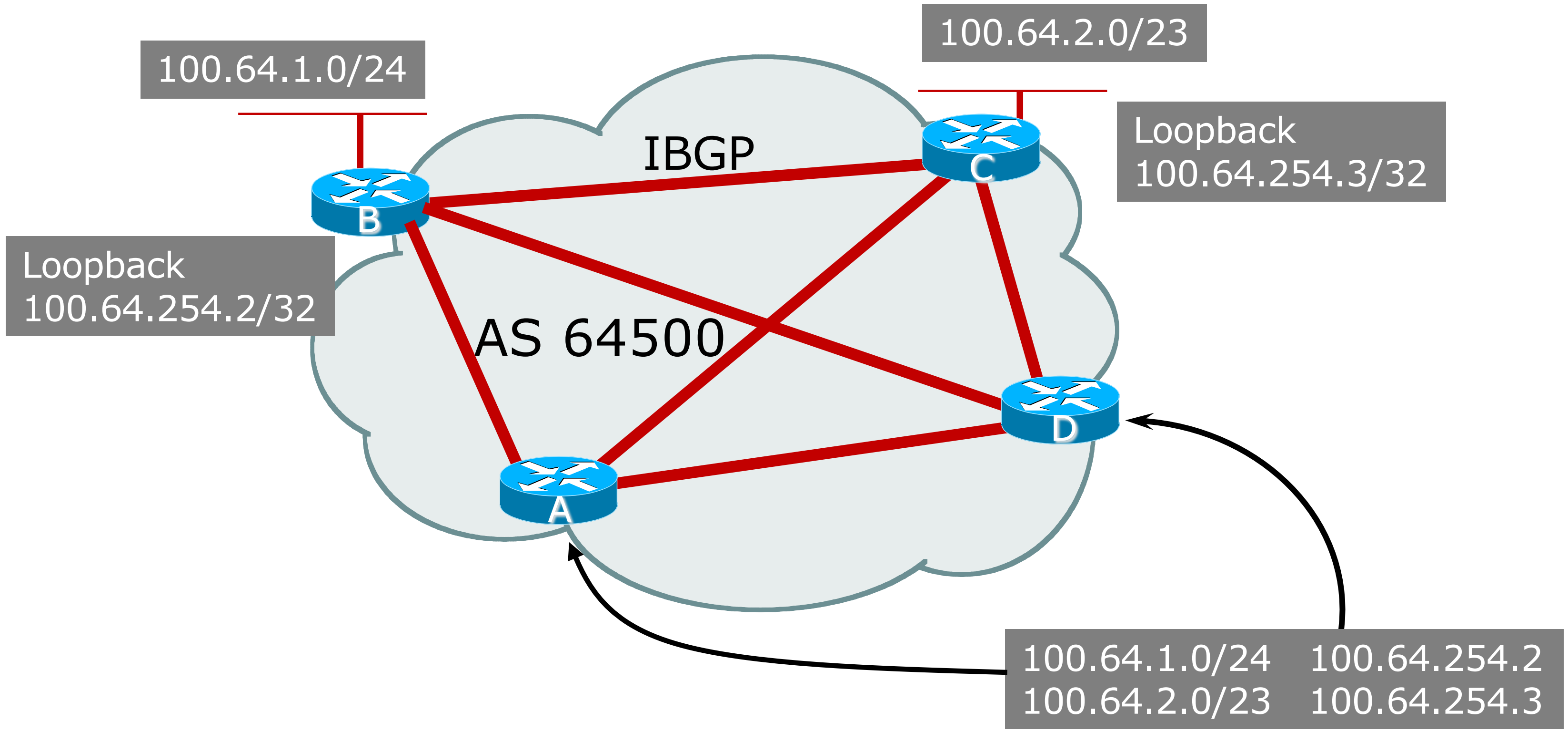
- The default behaviour is for external next-hop to be propagated unchanged to IBGP peers
  - This means that IGP has to carry external next-hops
  - Forgetting means external network is invisible
  - With many EBGP peers, it is unnecessary extra load on IGP
- Network operator Best Practice is to change external next-hop to be that of the local router
  - Cisco IOS:

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

  - JunOS:

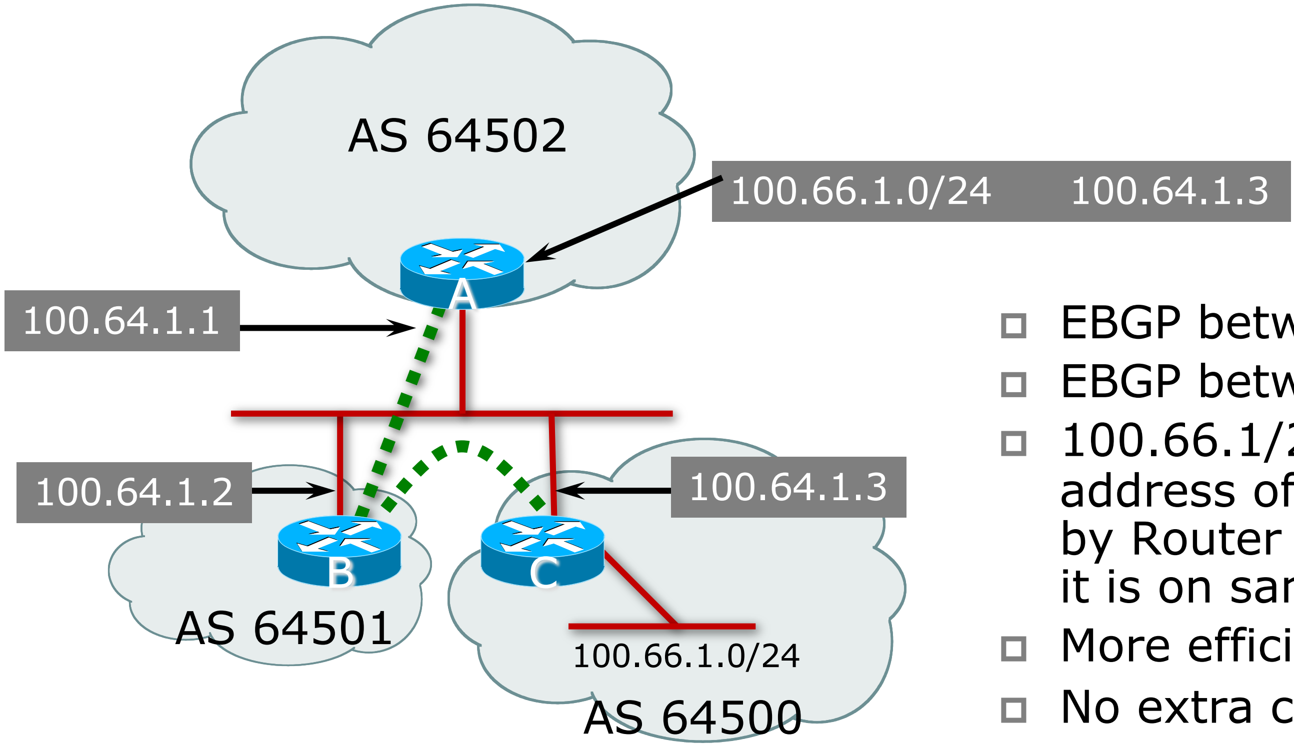
```
set policy-options
   policy-statement <name> term <name> then next-hop self
```

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
- 100.66.1/24 prefix has next hop address of 100.64.1.3 – this is used by Router A instead of 100.64.1.2 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

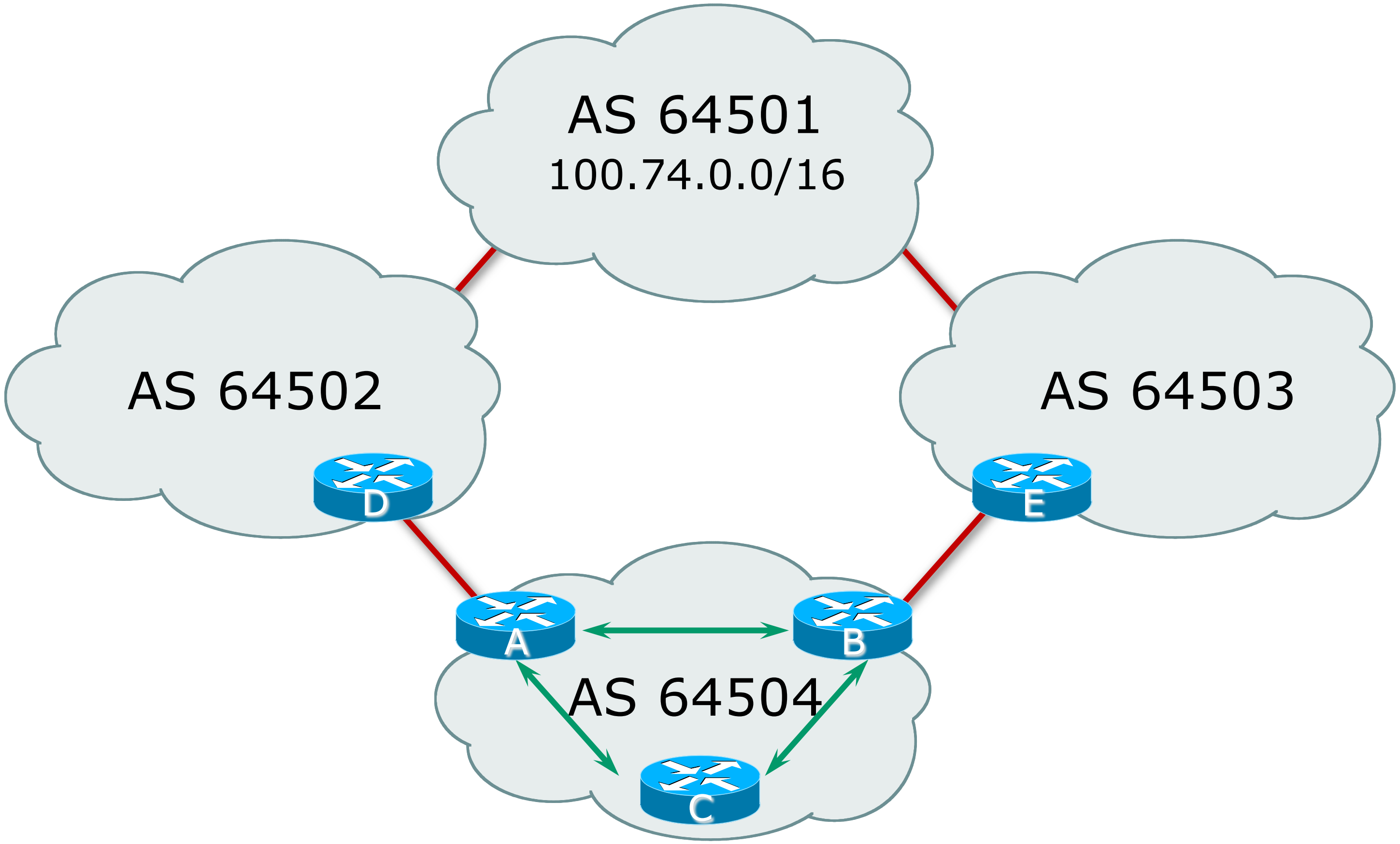
Origin

- 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

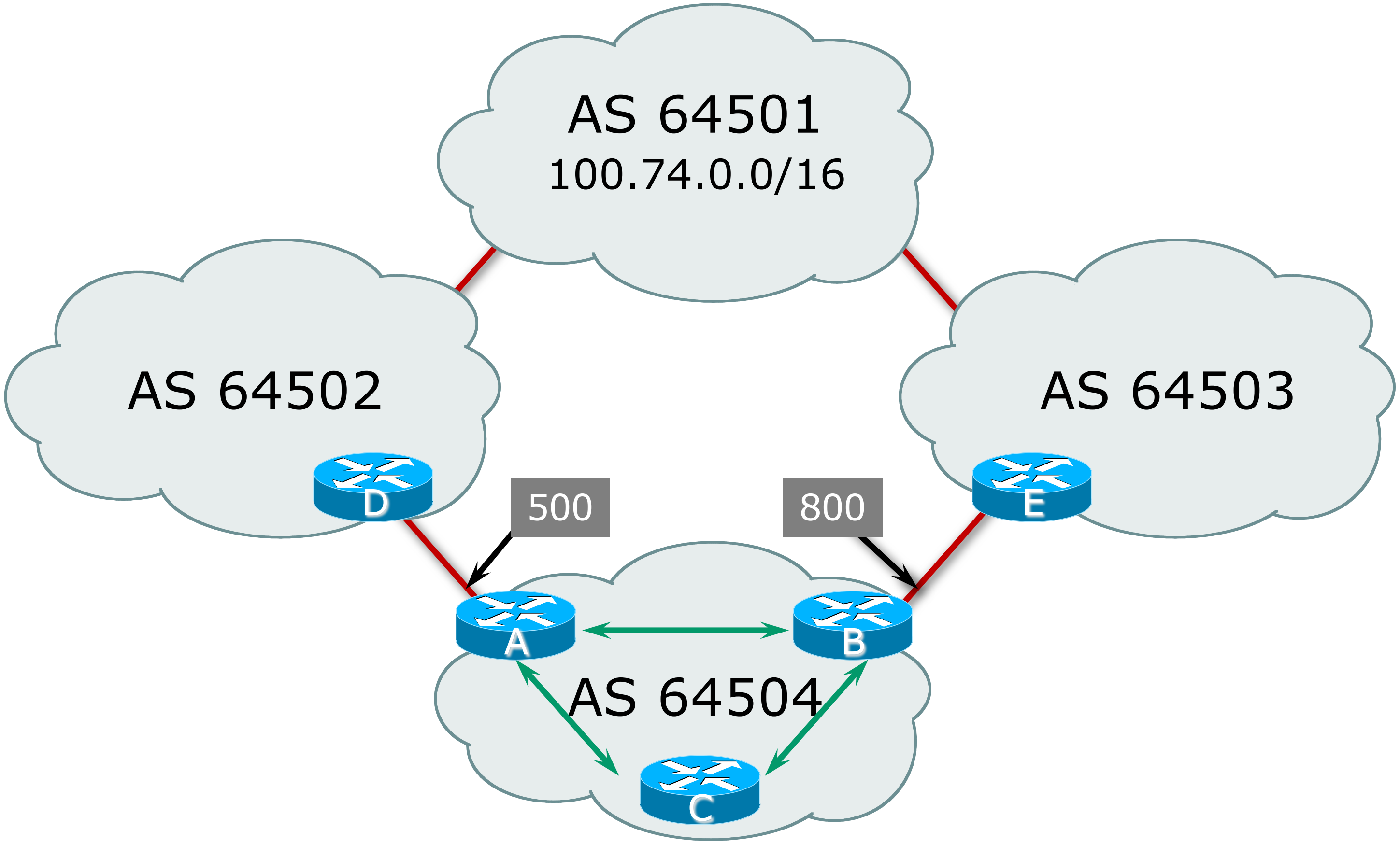
Aggregator

- 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

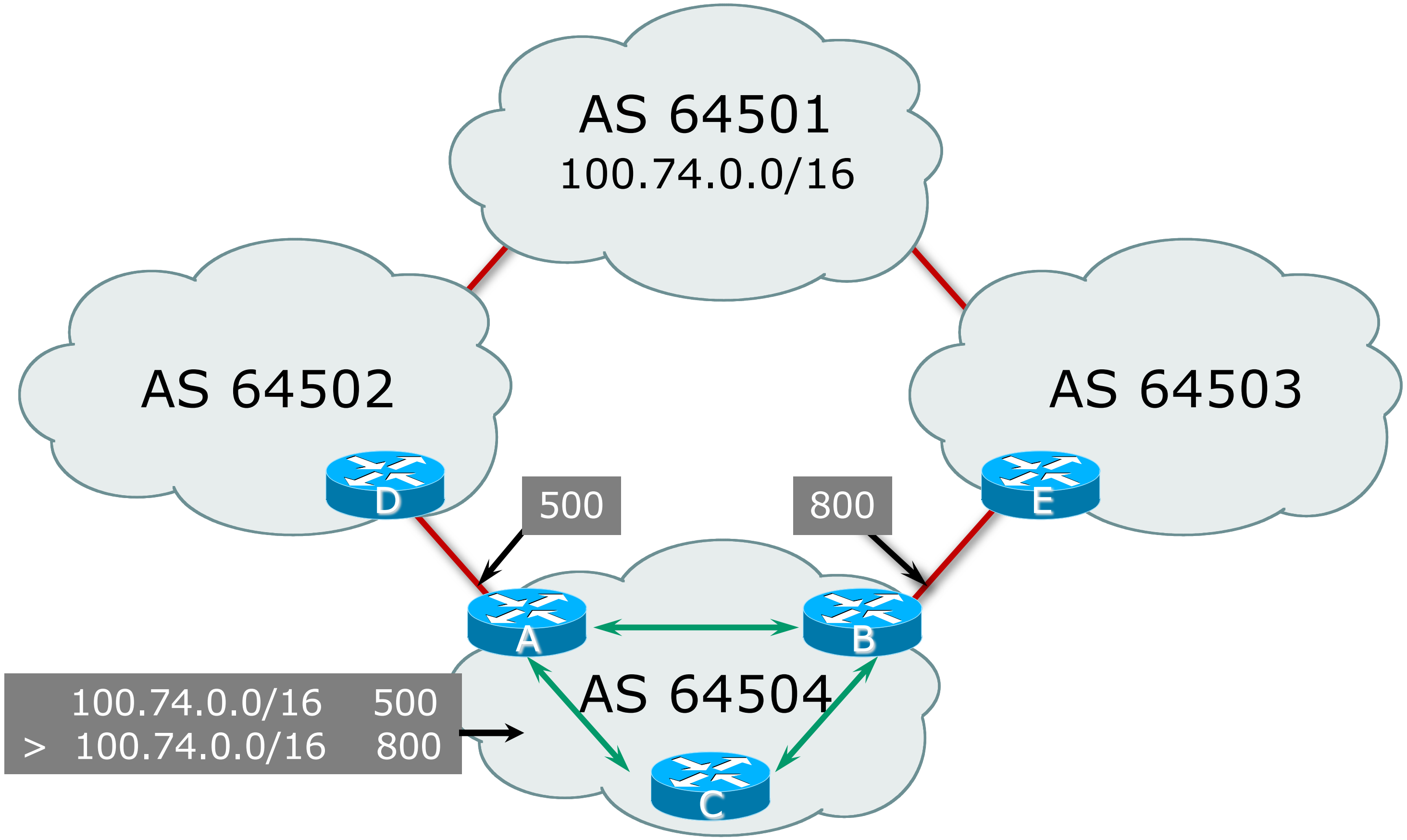
Local Preference



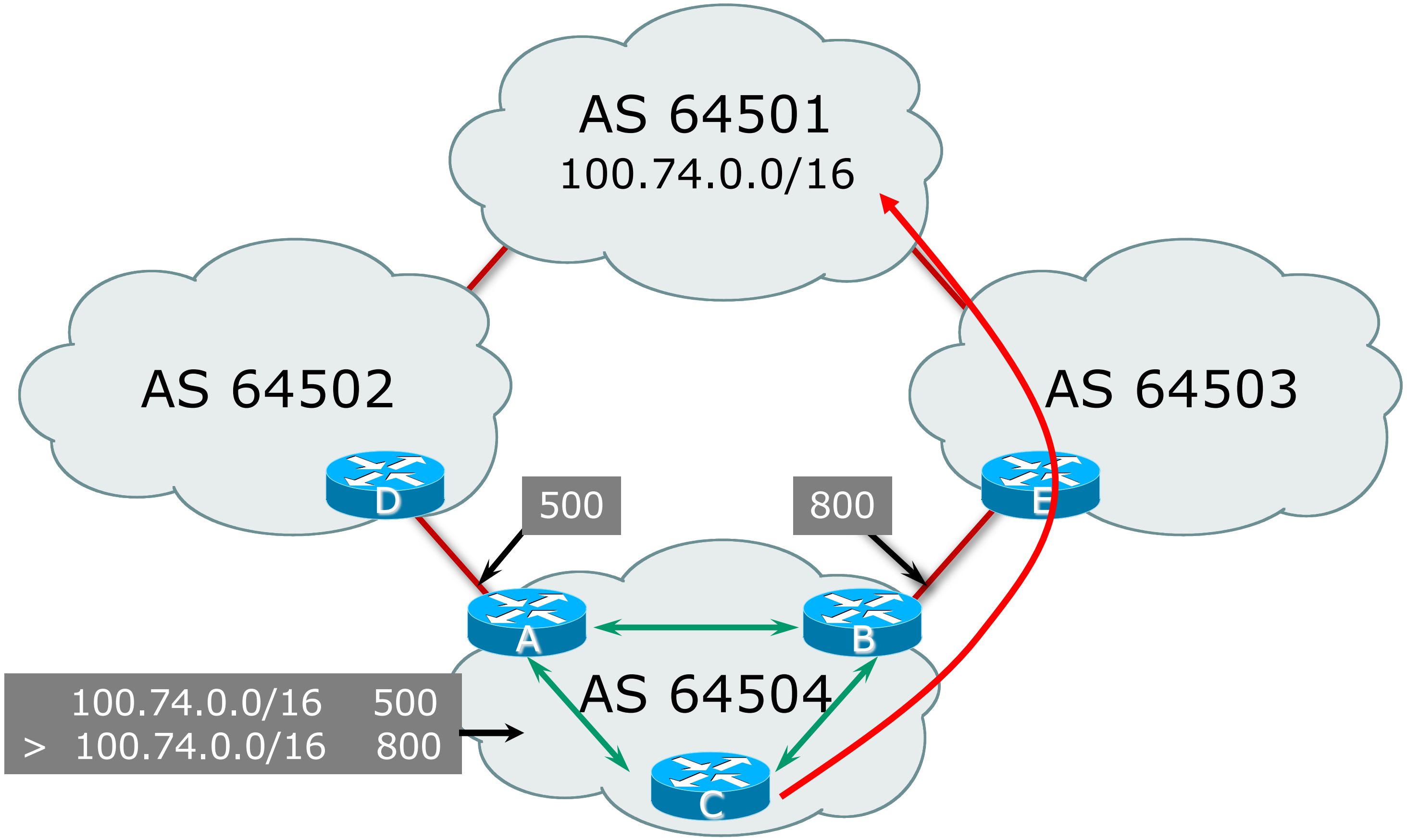
Local Preference



Local Preference



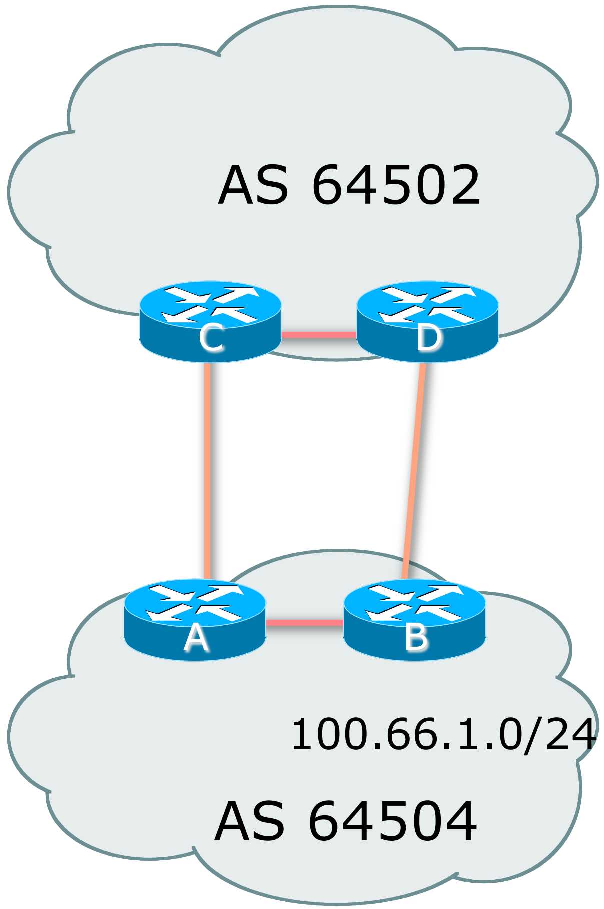
Local Preference



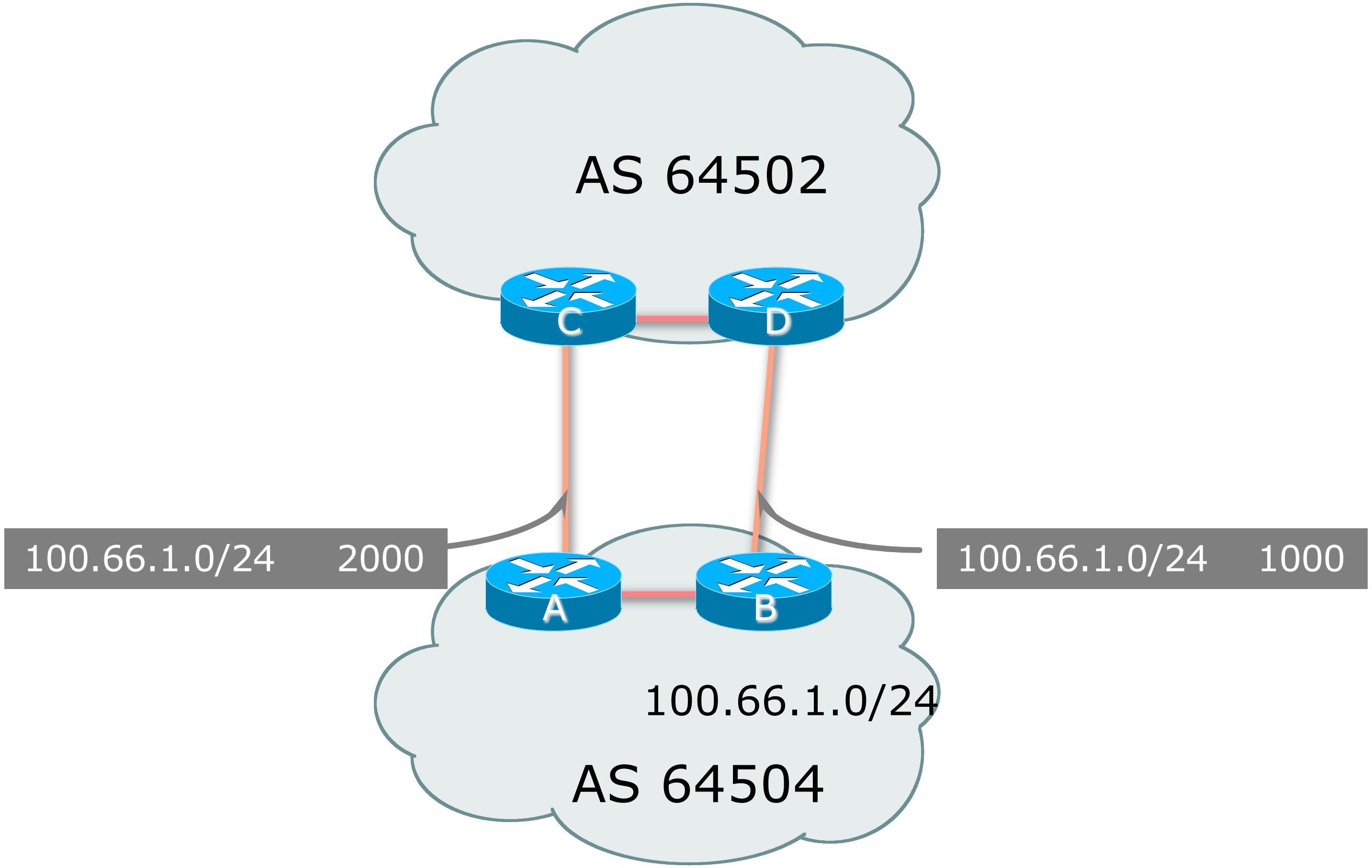
Local Preference

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

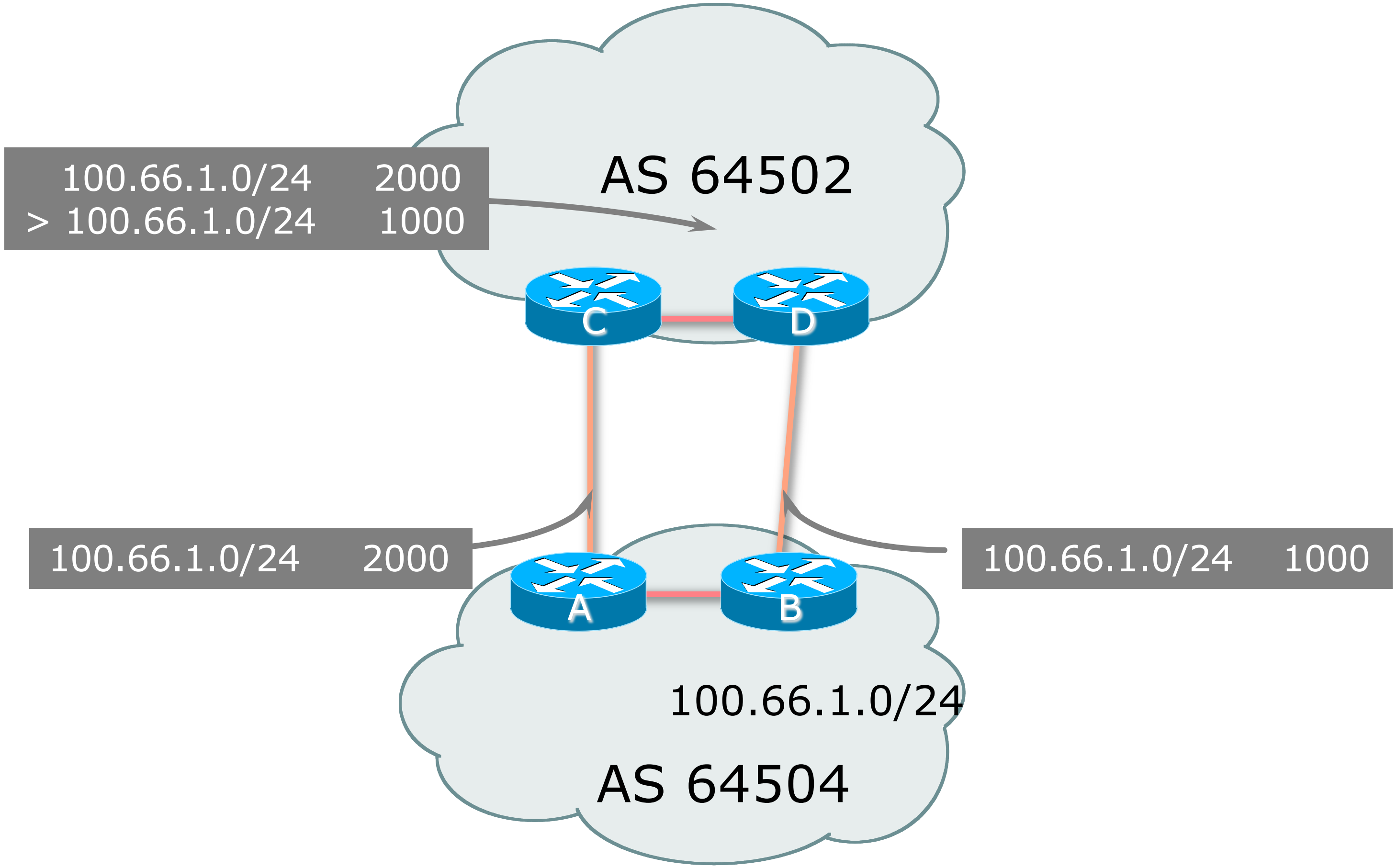
Multi-Exit Discriminator (MED)



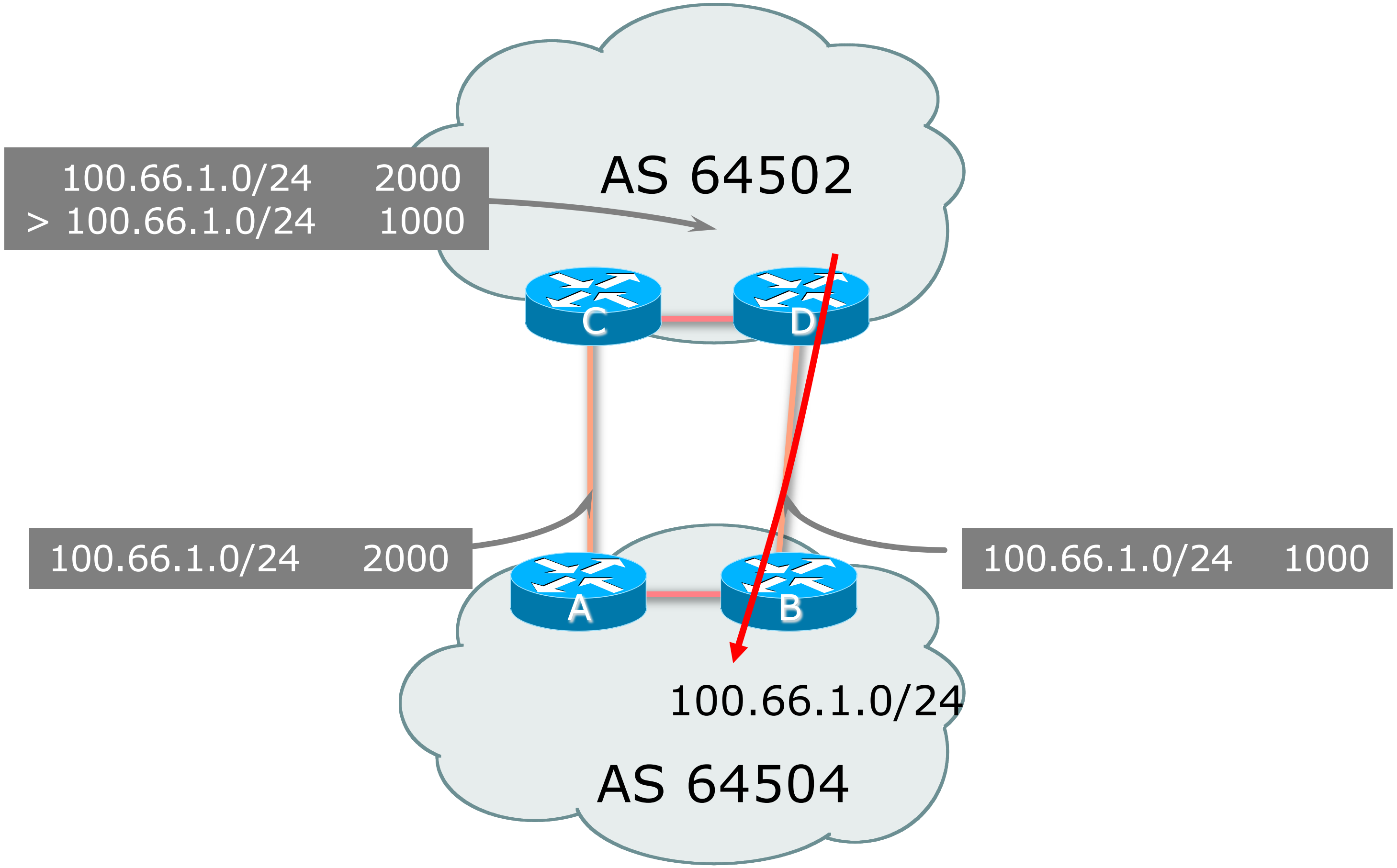
Multi-Exit Discriminator (MED)



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
  - Cisco's `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)

Deterministic MED

- Cisco 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:

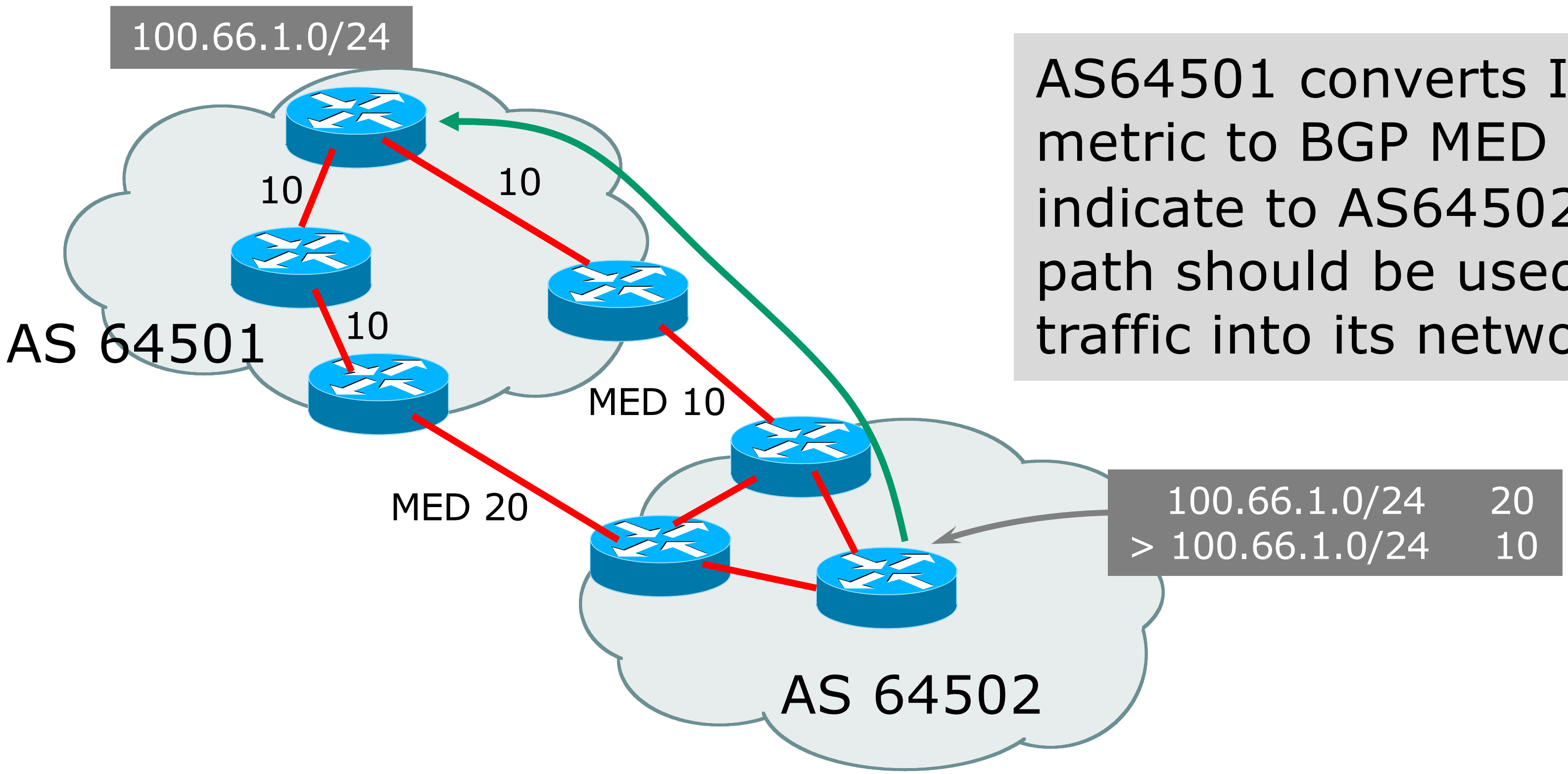
      ```
      bgp dynamic-med-interval <secs>
      ```

  - Also available in JunOS:

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

MED & IGP Metric

- Example: IGP metric conveyed as MED

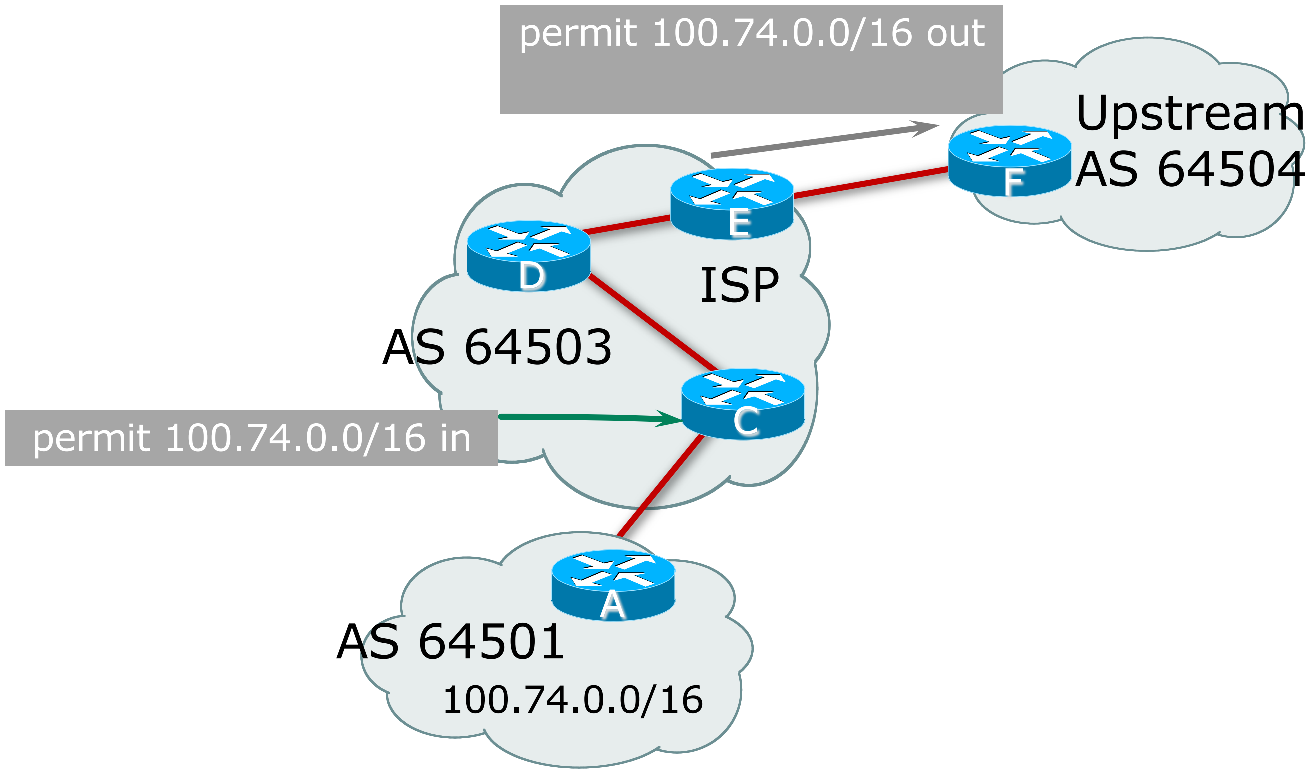


AS64501 converts IGP metric to BGP MED to indicate to AS64502 which path should be used for traffic into its network

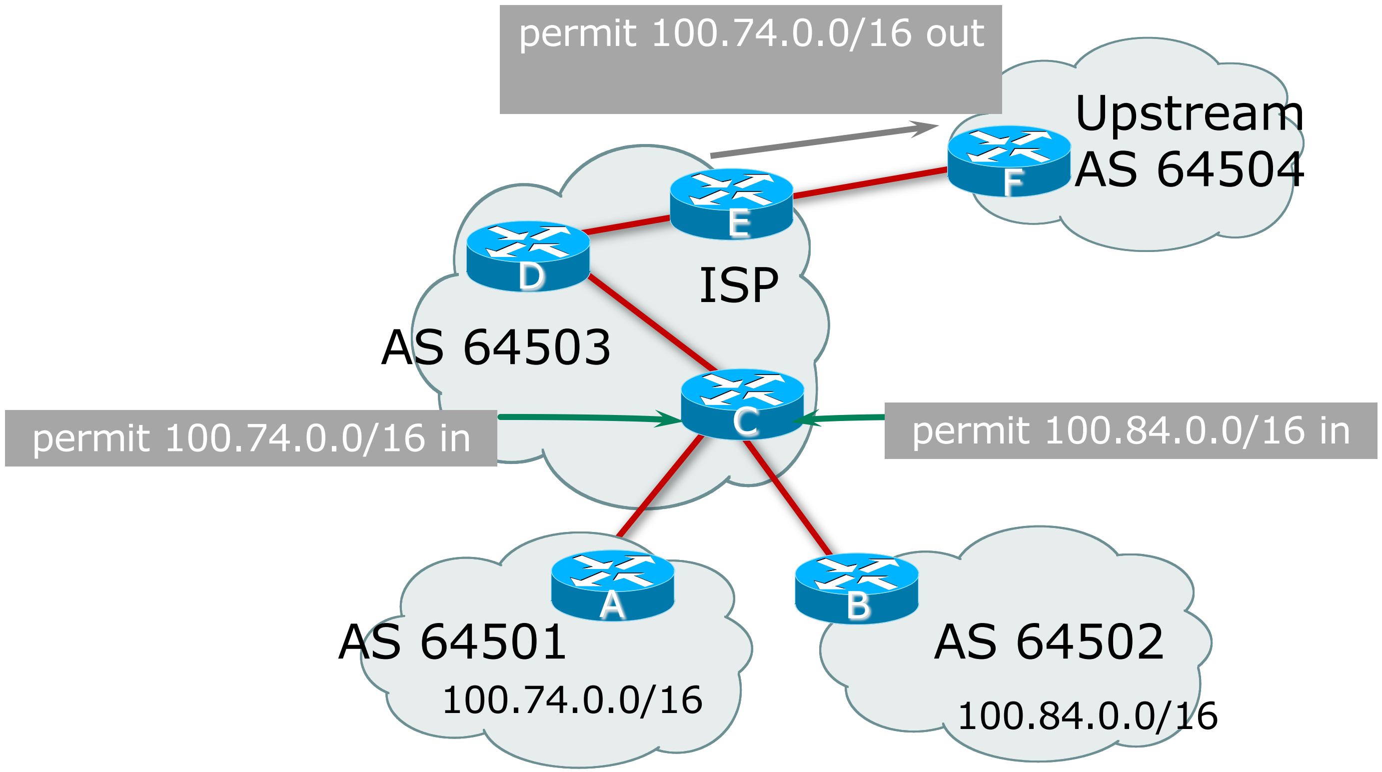
Community

- 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

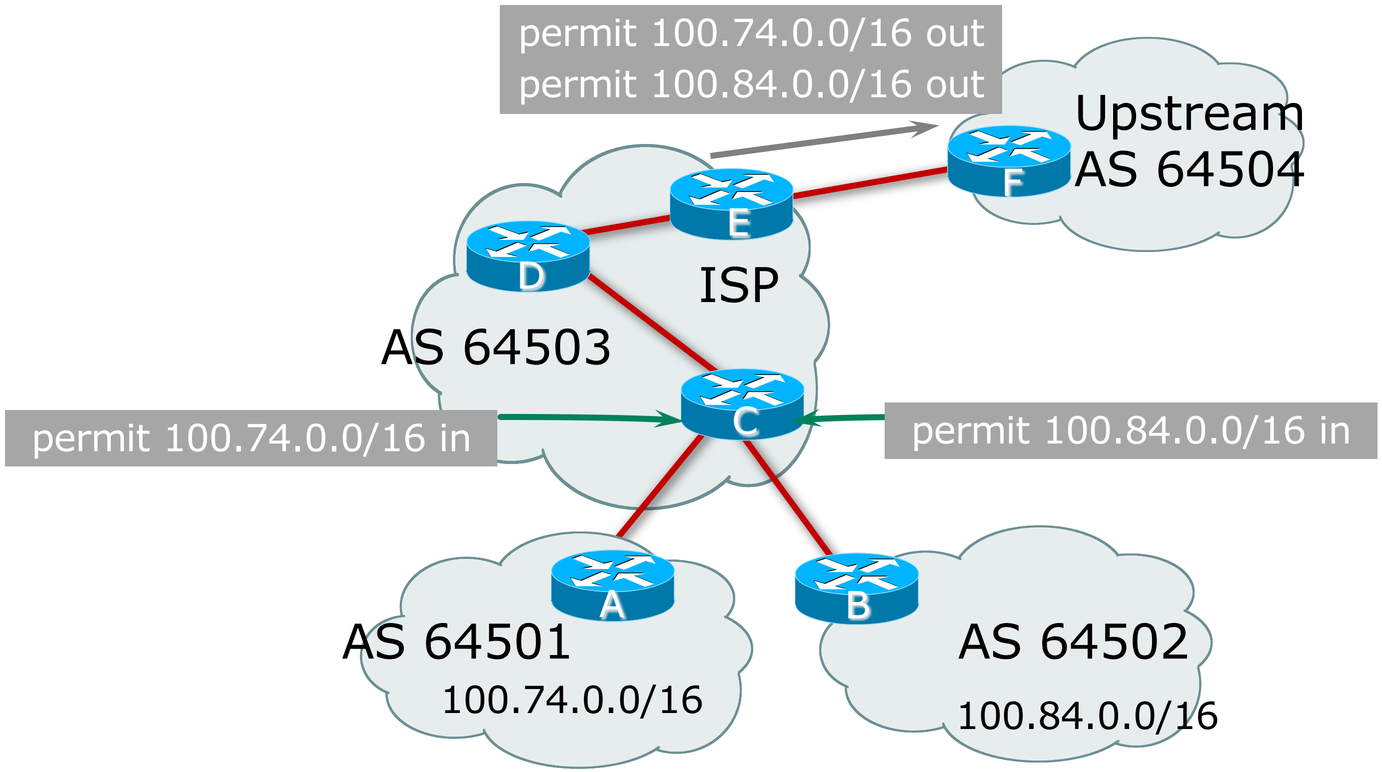
Community Example (before)



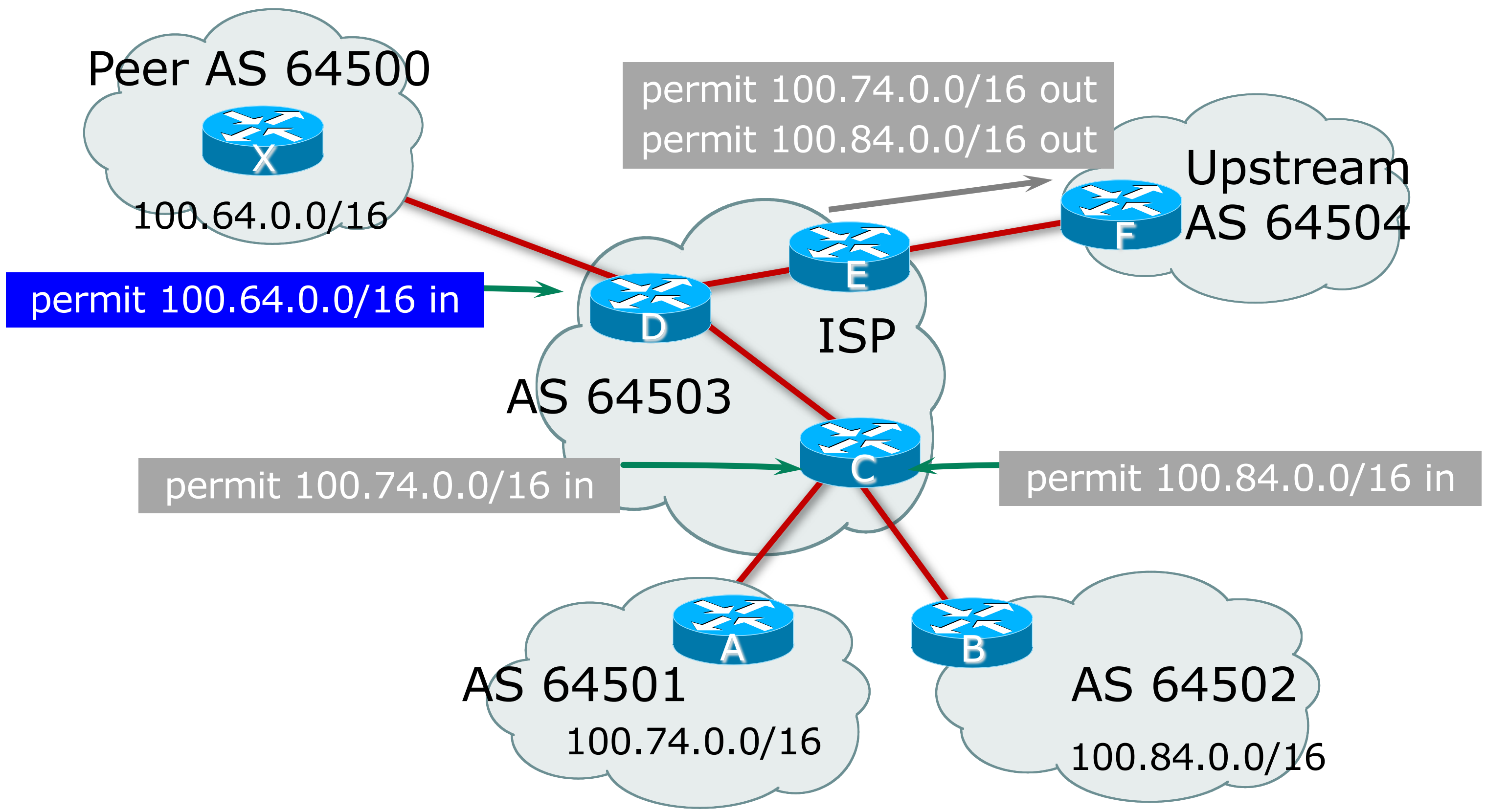
Community Example (before)



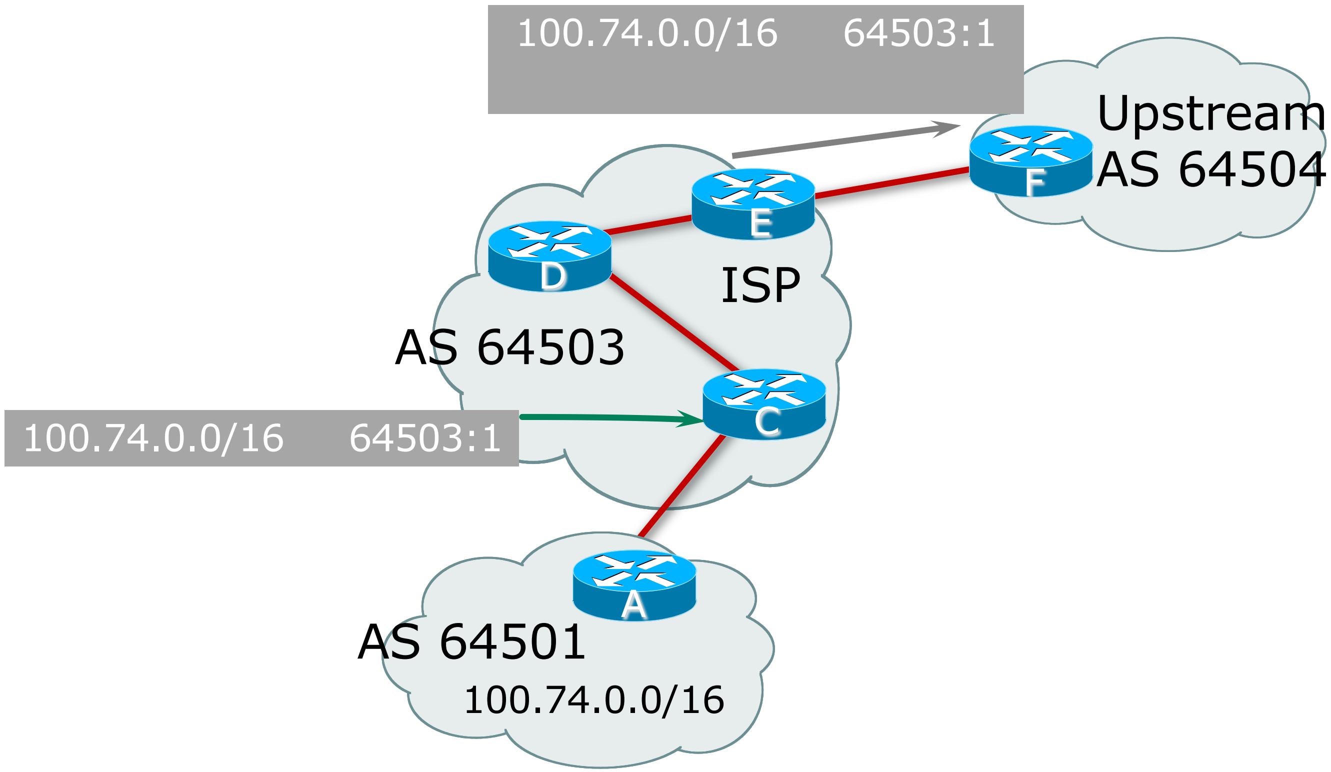
Community Example (before)



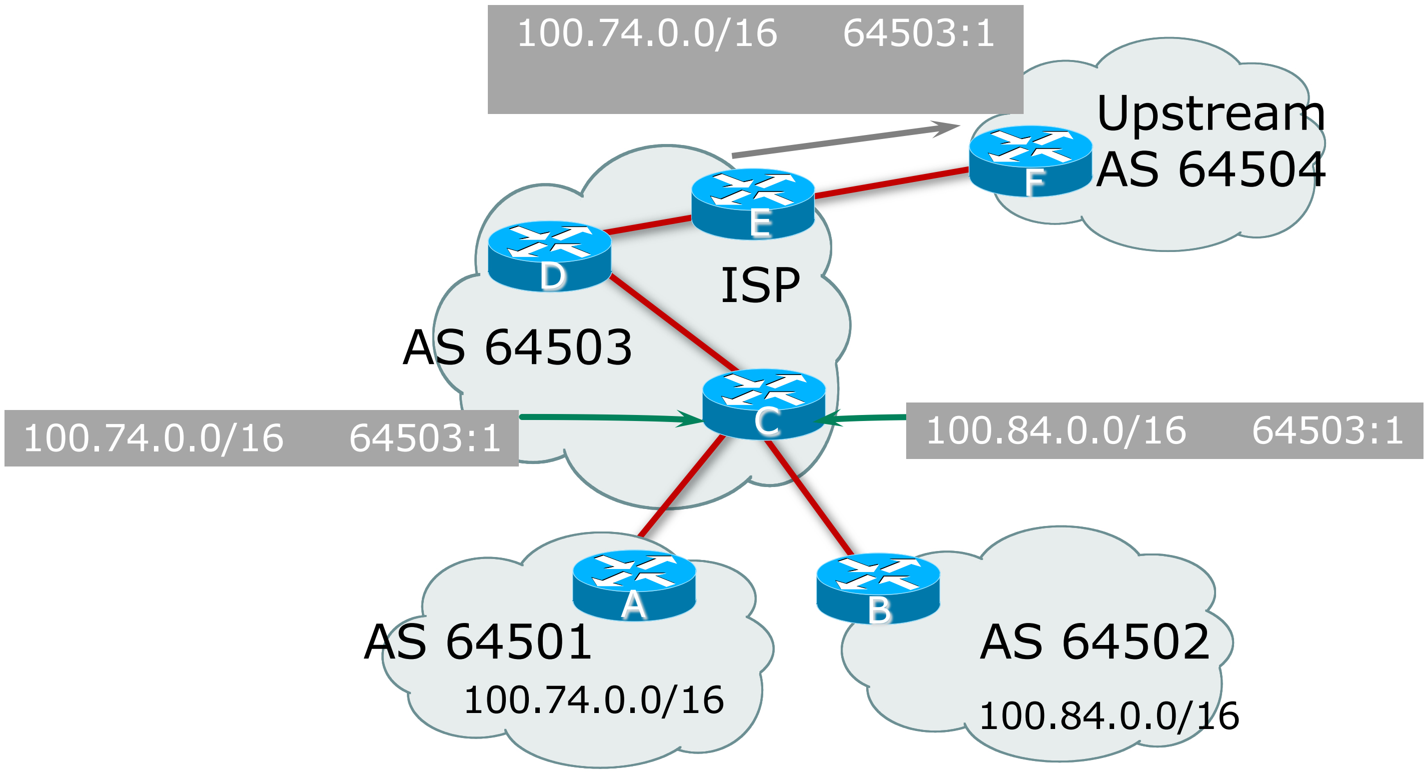
Community Example (before)



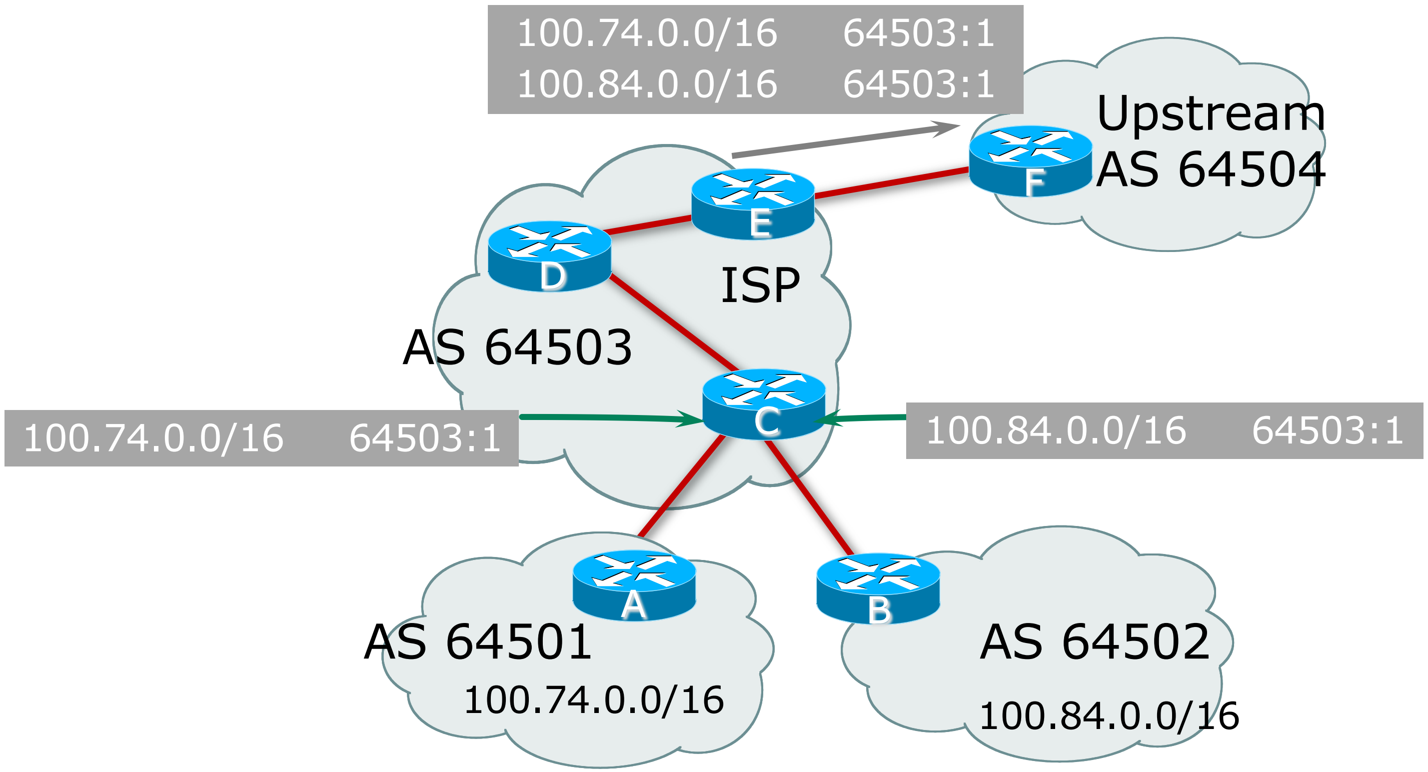
Community Example (after)



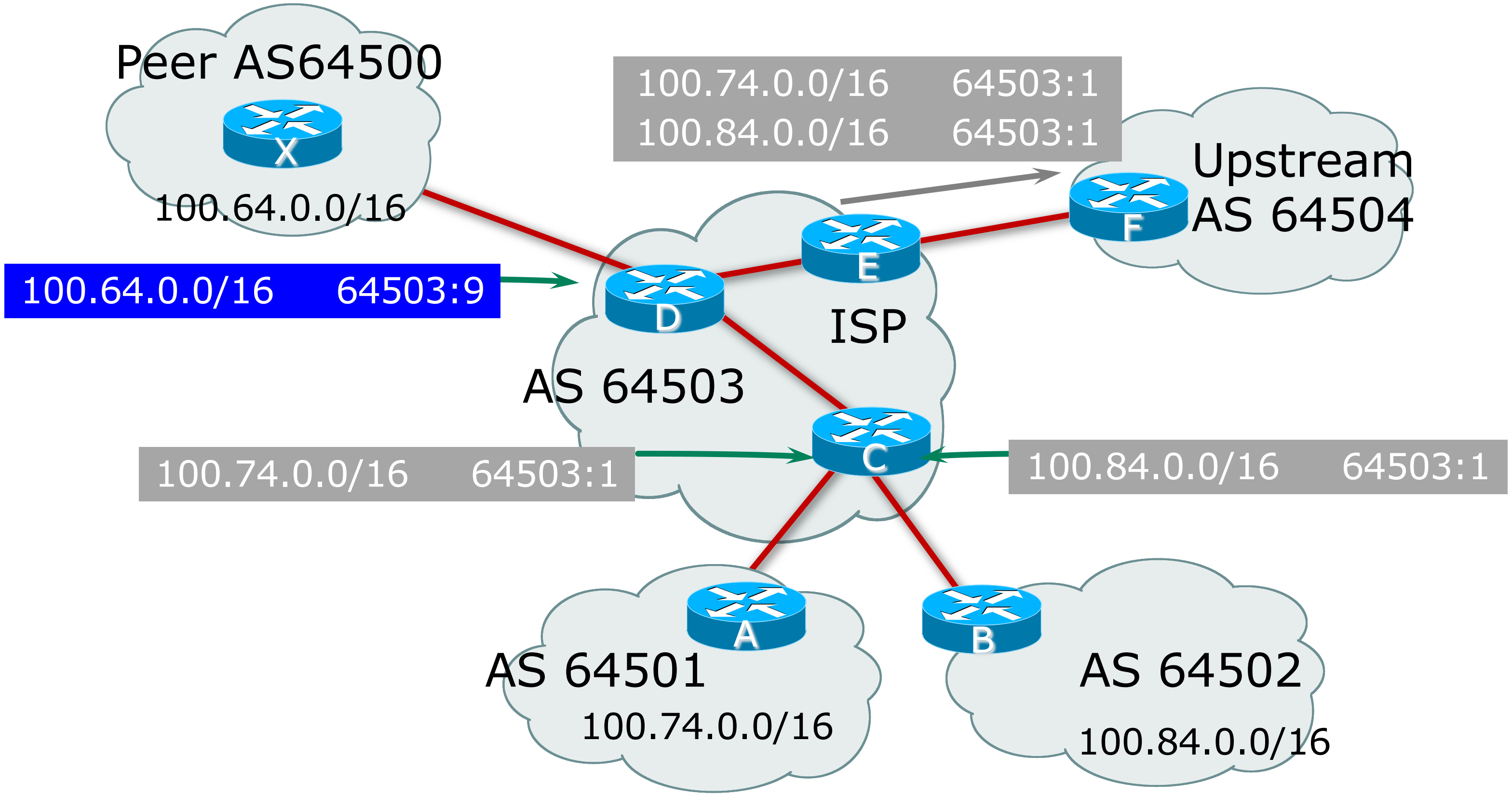
Community Example (after)



Community Example (after)



Community Example (after)



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/rfc8642.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 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 section for more detailed examples of typical network operator 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

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 (as per RFC)
    - If router-id is not the same, select the oldest path (non-RFC)
        - To turn off @ Cisco: bgp bestpath compare-routerid
        - To turn off @ Juniper: path-selection external-router-id
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 Path Selection Algorithm

- In multi-vendor environments:
  - Make sure the path selection processes are understood for each brand of equipment
  - All must follow the RFC, but because of “customer demand”, each vendor has:
    - Slightly different implementations
    - Extra steps
    - Extra features
  - Watch out for possible MED confusion

Applying Policy with BGP

Controlling Traffic Flow & Traffic Engineering

Applying Policy in BGP: Why?

- Network operators rarely “plug in routers and go”
- External relationships:
  - Control who they peer with
  - Control who they give transit to
  - Control who they get transit from
- Traffic flow control:
  - Efficiently use the scarce infrastructure resources (external link load balancing)
  - Congestion avoidance
  - Terminology: Traffic Engineering

Applying Policy in BGP: How?

- Policies are applied by:
  - Setting BGP attributes (local-pref, MED, AS-PATH, community), thereby influencing the path selection process
  - Advertising or Filtering prefixes
  - Advertising or Filtering prefixes according to ASN and AS-PATHs
  - Advertising or Filtering prefixes according to Community membership

Applying Policy with BGP: Tools

- Most implementations have tools to apply policies to BGP:
  - Prefix manipulation/filtering
  - AS-PATH manipulation/filtering
  - Community Attribute setting and matching
- Implementations also have policy language which can do various match/set constructs on the attributes of chosen BGP routes

Applying Policy with BGP: Tools

- Cisco and Cisco-like CLI (eg FRR):
  - Makes use of `route-map`s for policy, `prefix-list`s for filtering prefixes, and `as-path` filters for handling filtering by ASNs
- Juniper and similar CLI:
  - `policy-options` subsystem
    - `prefix-list` for filtering prefixes
    - `policy-statement` with different TERMs for policy actions
      - `route-filter` for specific filtering within policy-statements
    - `as-path` statements for handling ASN filtering
- Recommendation:
  - In a mixed-vendor/implementation environment, spend effort to ensuring policy language constructs do the "same thing"

BGP Capabilities

Extending BGP

BGP Capabilities

- Documented in RFC5492 and RFC8810
- 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 238 are assigned by IANA "first come first served"
  - 239 to 254 are "Experimental Use"

BGP Capabilities

- Current capabilities are listed opposite
- Most implementations support:
  - Multiprotocol extensions
  - Route Refresh
  - BGP ORF
  - Graceful Restart
  - 4-byte ASNs

```
  0 Reserved                                        [RFC3392]
  1 Multiprotocol Extensions for BGP-4              [RFC4760]
  2 Route Refresh Capability for BGP-4              [RFC2918]
  3 Outbound Route Filtering Capability             [RFC5291]
  4 Multiple routes to a destination capability     [RFC3107]
  5 Extended Next Hop Encoding                      [RFC5549]
  6 BGP Extended Message                            [RFC8654]
  7 BGPsec Capability                               [RFC8205]
  8 Multiple Labels Capability                      [RFC8277]
  9 BGP Role                                        [RFC9234]
 64 Graceful Restart Capability                     [RFC4724]
 65 Support for 4 octet ASNs                        [RFC6793]
 66 Deprecated
 67 Support for Dynamic Capability                  [ID]
 68 Multisession BGP                                [ID]
 69 Add Path Capability                             [RFC7911]
 70 Enhanced Route Refresh Capability               [RFC7313]
 71 Long Lived Graceful Restart                     [ID]
 72 Routing Policy Distribution                     [ID]
 73 FQDN Capability                                 [ID]
 74 BFD Capability                                  [ID]
 75 Software Version Capability                     [ID]
 76 PATHS-LIMIT Capability                          [ID]
128-131 & 184-185 Deprecated                        [RFC8810]
```

https://www.iana.org/assignments/capability-codes

BGP Techniques for Network Operators

- BGP Basics
- Scaling BGP
- Using Communities
- Deploying BGP in a Service Provider Network

BGP Scaling Techniques

Scaling BGP

BGP Scaling Techniques

- Original BGP specification and implementation was fine for the Internet of the early 1990s
  - But didn't scale
- Issues as the Internet grew included:
  - Scaling the IBGP mesh beyond a few peers?
  - Implement new policy without causing flaps and route churning?
  - Keep the network stable, scalable, as well as simple?

BGP Scaling Techniques

- BGP Configuration Scaling
  - Grouping BGP peers
- Industry Best Practice Scaling Techniques
  - Route Refresh
  - Route Reflectors
- Historical Scaling Techniques
  - Soft Reconfiguration
  - Confederations (not covered)
  - Route Flap Damping (not covered)

BGP Configuration Scaling

Cisco’s peer-groups

&

Juniper’s BGP groups

Grouping similar BGP peers

- What are they for?
  - Allows operators to group peers with the same outbound policy
  - Makes configuration easier
  - Makes configuration less prone to error
  - Makes configuration more readable
  - Members can have different inbound policy
  - Can be used for EBGP neighbours too!

Grouping similar BGP peers

- Cisco:
  - peer-groups
    - Originally designed to speed IBGP convergence – now for scaling BGP configuration management
  - Internal code optimisation called *update-groups*
    - Speeds IBGP convergence; update only calculated once for neighbours with the same outbound policy

- Juniper:
  - BGP groups

Configuring a Peer Group in IOS

```
router bgp 64500
 address-family ipv4
  neighbor IBGP peer-group
  neighbor IBGP remote-as 64500
  neighbor IBGP update-source loopback 0
  neighbor IBGP send-community
  neighbor IBGP route-map outfilter out
  neighbor 100.64.0.1 peer-group IBGP
  neighbor 100.64.0.2 peer-group IBGP
  neighbor 100.64.0.2 route-map  infilter in
  neighbor 100.64.0.3 peer-group IBGP
!
```

- Note how 100.64.0.2 has an additional inbound filter over the peer-group

Configuring a Peer Group in IOS

```
router bgp 64500
 address-family ipv4
  neighbor EBGP peer-group
  neighbor EBGP send-community
  neighbor EBGP route-map set-metric out
  neighbor 100.89.1.2 remote-as 64502
  neighbor 100.89.1.2 peer-group EBGP
  neighbor 100.89.1.4 remote-as 64503
  neighbor 100.89.1.4 peer-group EBGP
  neighbor 100.89.1.6 remote-as 64504
  neighbor 100.89.1.6 peer-group EBGP
  neighbor 100.89.1.6 filter-list infilter in
!
```

- Can be used for EBGP as well

Juniper BGP groups

- JunOS has very similar configuration concept
  - Simply known as bgp groups, for example:

```
protocols {
    bgp {
        group ibgp {
            type internal;
            local-address 10.0.15.241;
            family inet {
                unicast;
            }
            export export-ibgp;
            peer-as 10;
            neighbor 10.0.15.242 {
                description "Router 2";
            }
            neighbor 10.0.15.243 {
                description "Router 3";
            }
    ...etc...
        }
    }
}
```

Grouping similar BGP peers

- Always configure peer-groups or BGP groups for IBGP
  - Even if there are only a few IBGP peers
  - Easier to scale network in the future
  - Makes configuration easier to read
- Consider using peer-groups for EBGP
  - Especially useful for multiple BGP customers using same AS (RFC2270)
  - Also useful at Exchange Points:
    - Where network operator policy is generally the same to each peer
    - For Route Server where all peers receive the same routing updates

Dynamic Reconfiguration

Non-destructive policy changes

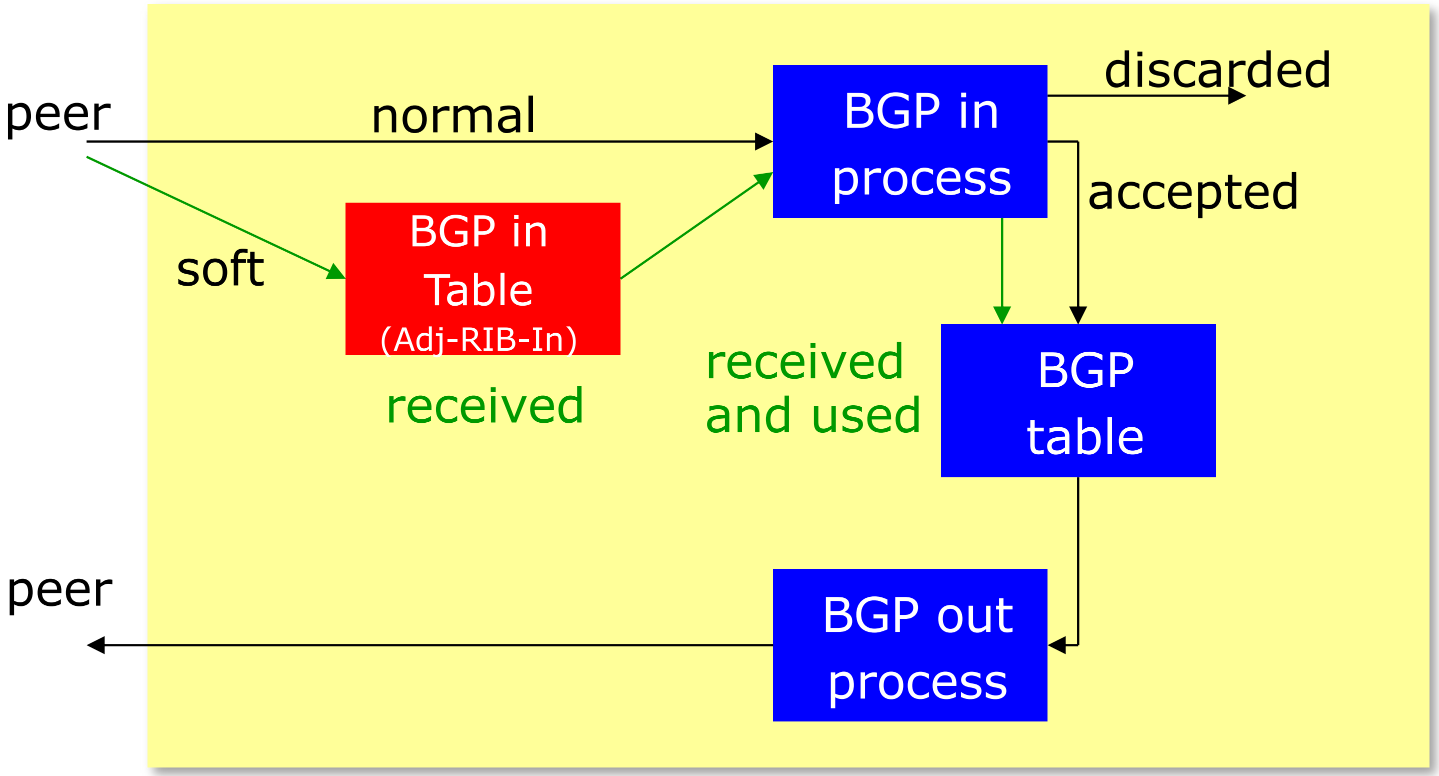
Route Refresh: History

- Historically, routers only stored prefixes which were accepted by incoming policy
  - Those rejected by policy were discarded
  - No storage of discarded prefixes
- If a change of incoming policy was required:
  - The EBGP session had to be shutdown, and then brought up again
  - Destructive change: EBGP session down means lost connectivity to that peer, and potentially the rest of the Internet (outage of many minutes!)
- Changes in BGP policy usually had to be carried out during published scheduled maintenance timeslots
  - To minimise impact on end-users

Route Refresh: Step One

- First step at solving this problem was by Cisco with the "soft reconfiguration" concept
  - Router keeps a record of all prefixes received **before** any policy applied (known as Adj-RIB-In)
  - Needed extra memory (highly problematic in early routers and modern routers with limited memory)
    - Full BGP table with policy change could require double the control plane memory for BGP
  - Policy changes applied to the stored received prefixes
  - No shutdown and restart of the BGP session needed when implementing policy changes

Cisco’s Soft Reconfiguration



Route Refresh: Step Two

- Second step at solving this problem was the introduction of "route refresh"
  - A BGP Capability: RFC2918
  - Peering remains active
  - Impacts only those prefixes affected by the policy change
  - No configuration needed
    - Automatically negotiated at peer establishment
    - No extra memory needed (no need for Adj-Rib-In)
- Today most implementations do an automatic route-refresh after BGP Policy changes
  - Beware: not all vendor implementations do an automatic route-refresh – know your software implementation!

Route Refresh

- Use Route Refresh capability, *not* hard reset
  - Supported on virtually all BGP implementations
  - Find out from the detailed BGP neighbour status
  - Non-disruptive, "Good For the Internet"

- Only hard-reset a BGP peering as a last resort

Consider the impact of a hard-reset of BGP to be equivalent to a router reboot

Route Refresh: Route Origin Validation

- Route Origin Validation means checking if the prefix received has a valid ROA
  - Route Origination Authorisation – digital object indicating the origin AS for the prefix (and subnet size) using RPKI
  - Valid ROA means that the prefix (and subnet) is being originated from the correct origin AS
  - See the "BGP Origin Validation" presentation for more in-depth content
- Routers implementing ROV apply the validation results via the existing policy language & process
  - Valid – allow; Invalid – drop; NotFound – allow (at lower preference?)
- Problem: how is incoming policy applied on routers today?

Route Refresh: Route Origin Validation

- Routers which maintain the Adj-RIB-In:
  - Apply the ROV policy to the stored received BGP table
  - Updates are applied "automatically" to the BGP table and therefore the FIB
  - No impact on any BGP peers (Route Refresh not needed)

Route Refresh: Route Origin Validation

- Routers which do NOT maintain the Adj-RIB-In:
  - Apply the ROV policy by sending a Route Refresh to peers
  - When there are a large number of ROAs (May 2024 saw over 438k IPv4 and 106k IPv6), and frequent changes or updates of ROAs:
    - Routers are sending frequent Route Refresh requests to peers (typically every few minutes)
    - Peers are being "bombarded" by Route Refresh requests: significant resource burden when they send the full or a large portion of the BGP table
    - Severe control plane CPU impact on the peer router (effectively a Denial of Service on the peer router)
  - As more and more ROAs are created and altered globally, this problem becomes significantly more serious!

Route Refresh: Route Origin Validation

- JunOS implements Adj-RIB-In by default
  - ROA updates do not cause a problem when operating ROV

- Cisco does not implement Adj-RIB-In by default:
  - Applies to all Cisco IOS/IOS-XE/IOS-XR apart from the most recent releases
  - **MUST turn on soft-reconfiguration if running ROV on the router**
  - Soft-reconfiguration is similar concept to Adj-RIB-In

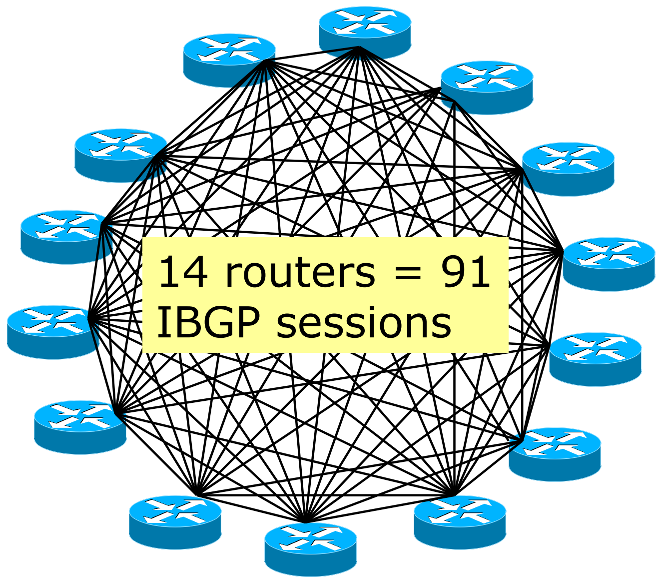
Route Reflectors

Scaling the IBGP mesh

Scaling the IBGP mesh

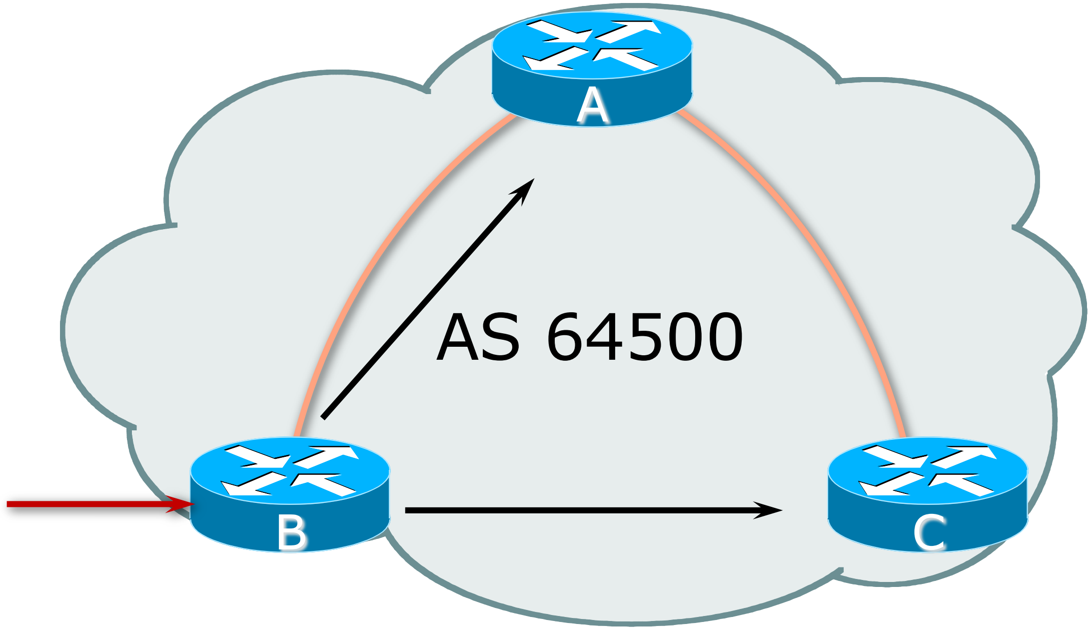
- Avoid ½n(n-1) IBGP mesh

n=1000 ⇒ nearly
half a million
IBGP sessions!

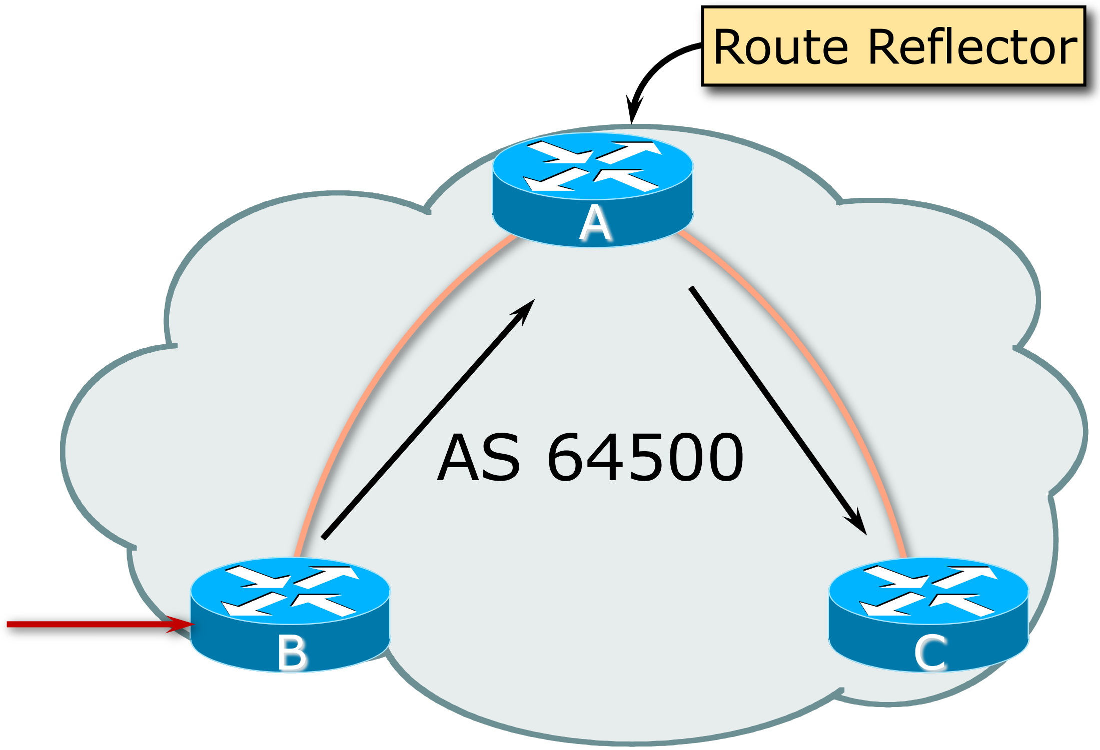


- Two solutions
  - Route reflector: simpler to deploy and run
  - BGP Confederation: more complex, has corner case advantages

Route Reflector: Principle

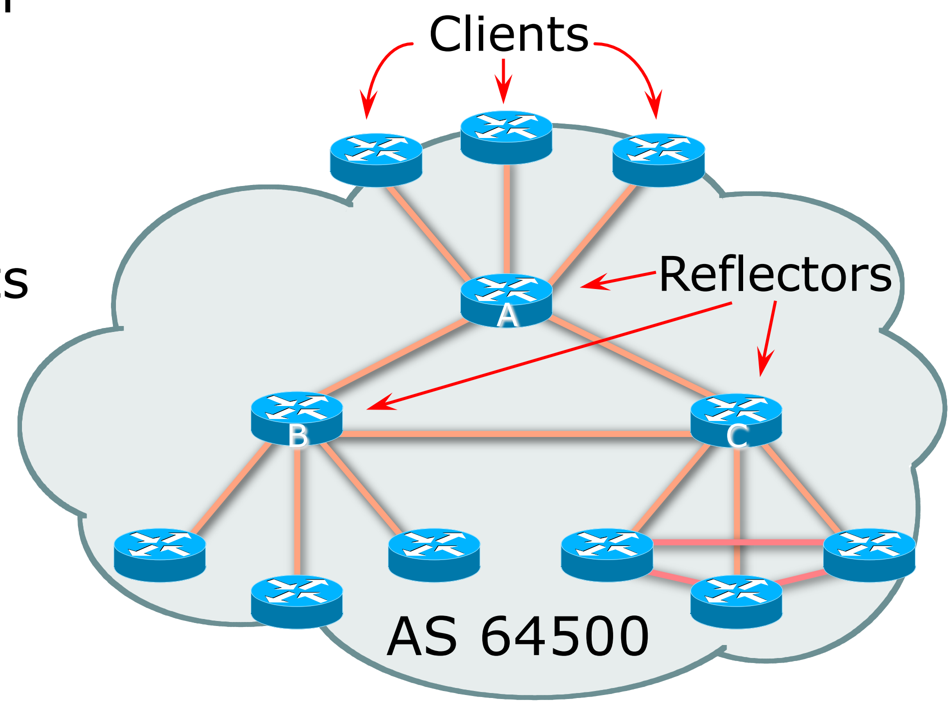


Route Reflector: Principle



Route Reflector: Rules

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



Route Reflector: Topology

- Divide the backbone into multiple clusters
- Provision at least one Route Reflector (RR) and few clients per cluster
- Route reflectors are fully meshed
- Clients in a cluster could be fully meshed
- Single IGP still carries next-hop and any local routes

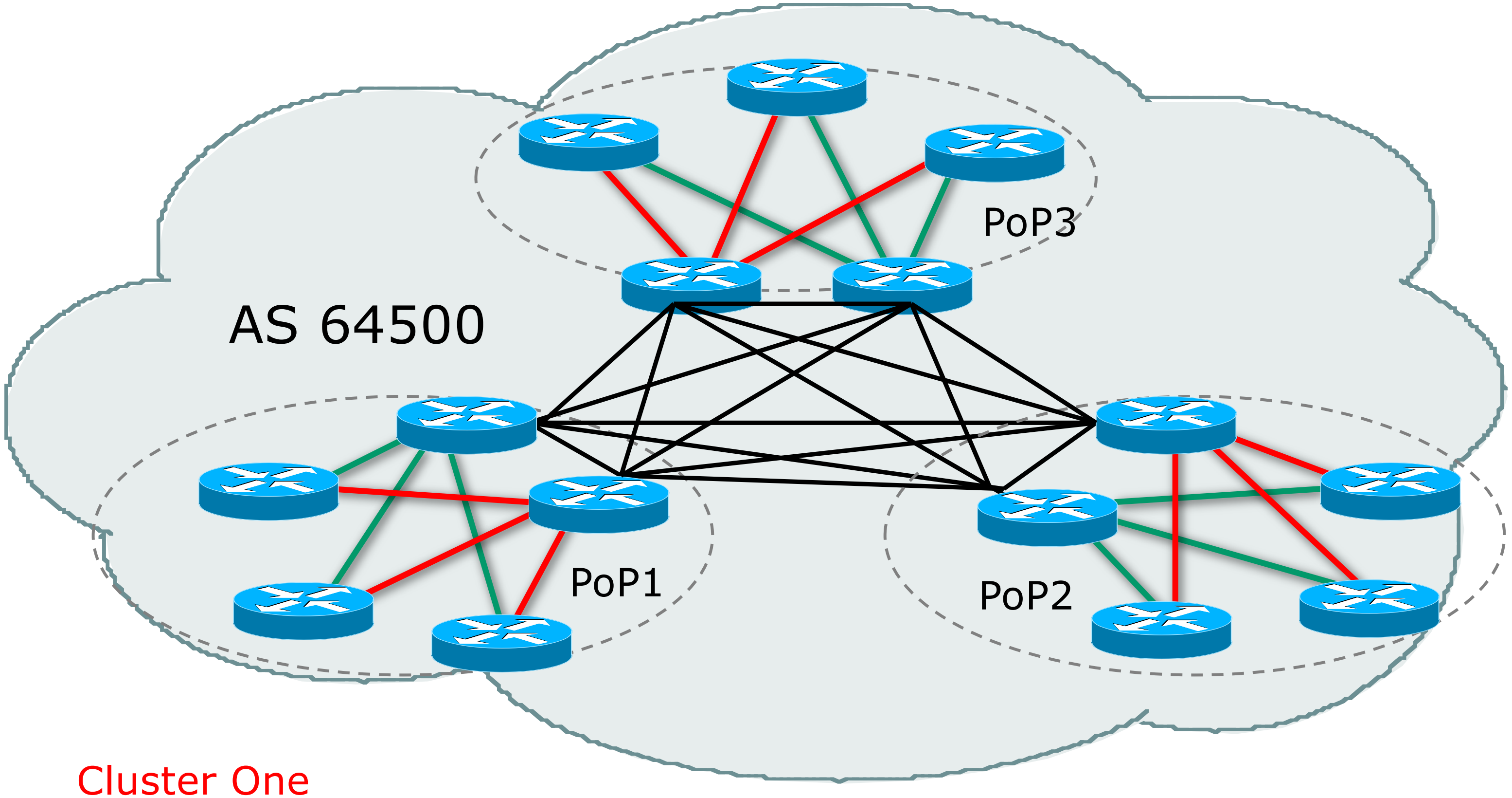
Route Reflector: Loop Avoidance

- Originator_ID attribute
  - Carries the RID of the originator of the route in the local AS (created by the RR)
- Cluster_list attribute
  - The local cluster-id is added when the update is sent by the RR
  - Best to set cluster-id from router-id by (address of loopback interface)
  - (Some Network Operators use their own cluster-id assignment strategy – but needs to be well documented!)

Route Reflector: 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 service provider networks to overlay two clusters – redundancy achieved that way
  - → Each client has two RRs = redundancy

Route Reflector: Redundancy



Cluster One

Cluster Two

Route Reflector: Benefits

- Solves IBGP mesh problem
- Packet forwarding is not affected
- Normal BGP speakers co-exist
- Multiple reflectors for redundancy
- Easy migration
- Multiple levels of route reflectors

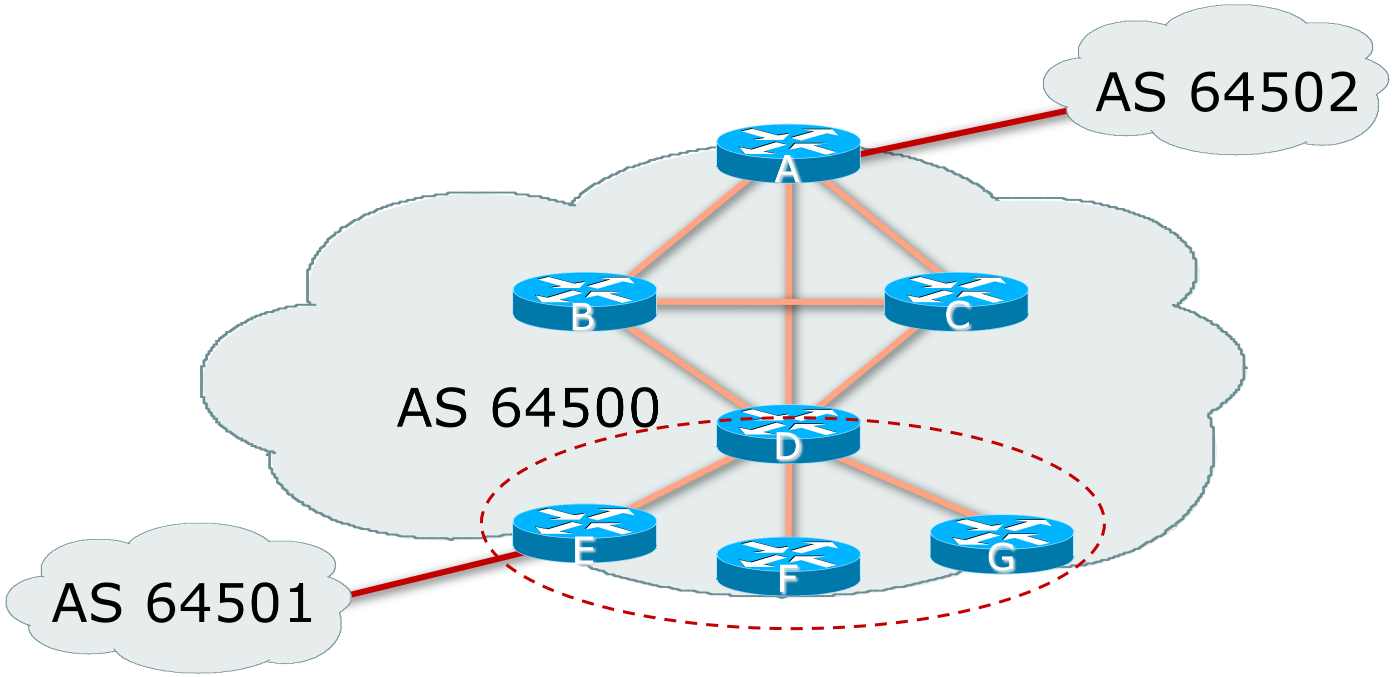
Route Reflector: Deployment

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

Route Reflector: Migration

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

Route Reflector: Migration



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

BGP Scaling Techniques

- These two standards-based techniques must be designed in from the beginning for all network operator infrastructure
  1. Route Refresh
  2. Route Reflectors

BGP Techniques for Network Operators

- BGP Basics
- Scaling BGP
- Using Communities
- Deploying BGP in a Service Provider Network

BGP Communities

Scaling BGP Policies

Multihoming and Communities

- The BGP community attribute is a very powerful tool for assisting and scaling BGP Policies and BGP Multihoming
  - BGP Communities were introduced earlier – now a more detailed look at this power BGP tool is required
- Most major Network Operators make extensive use of BGP communities:
  - Internal policies (IBGP)
  - Inter-provider relationships (MED replacement)
  - Customer traffic engineering

Well-known BGP Communities

How the “well-known” BGP communities
are used

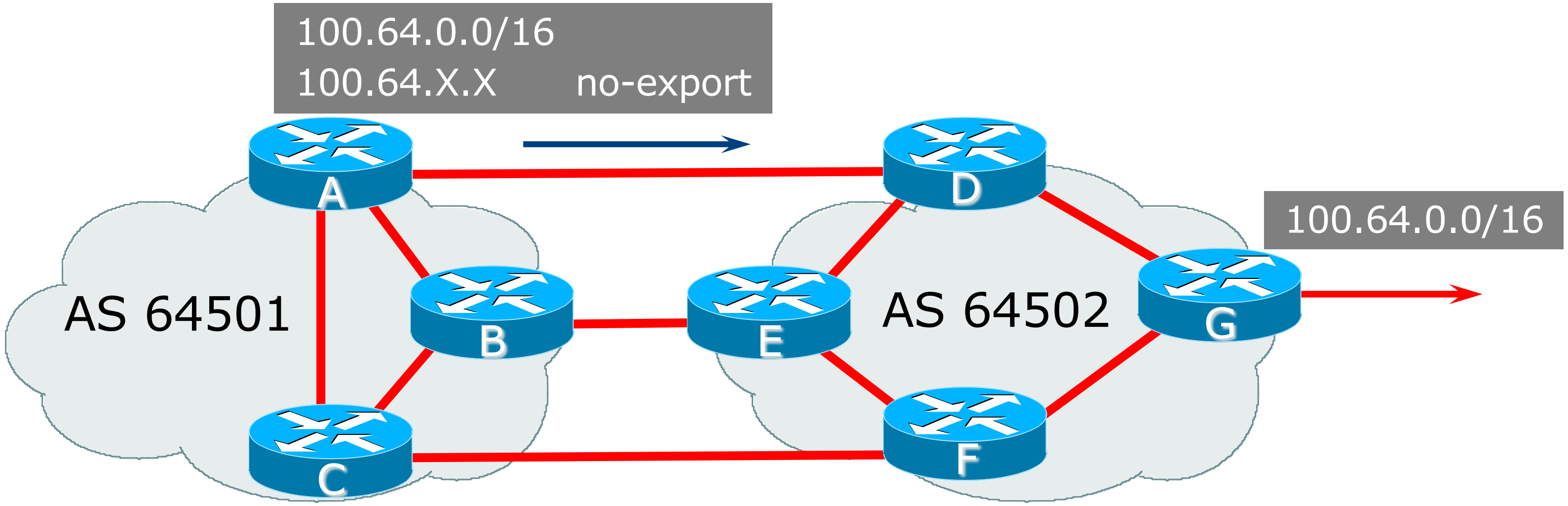
Well-Known Communities

- Several well-known communities
  - www.iana.org/assignments/bgp-well-known-communities
- Five 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-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)

Well-Known Communities: Notes

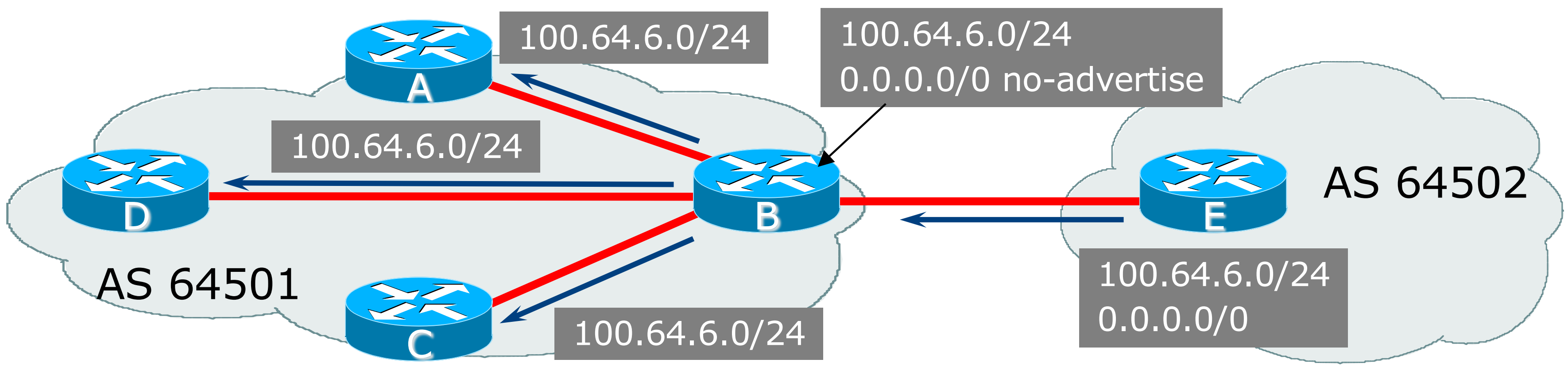
- Even though there are several well-known communities there are variations in implementation support
  - Not all vendors will create configuration key-words to support them
  - Not all vendors will automatically implement their behaviours
  - Not all vendors will allow them to be overwritten
  - *And so on*
- Check vendor documentation for implementation details
  - RFC8642 will give some idea as to the issues to be aware of
- Advice:
  - If the key-word does not exist, create a community declaration that implements the key-word (for configuration clarity & simplicity)

No-Export Community



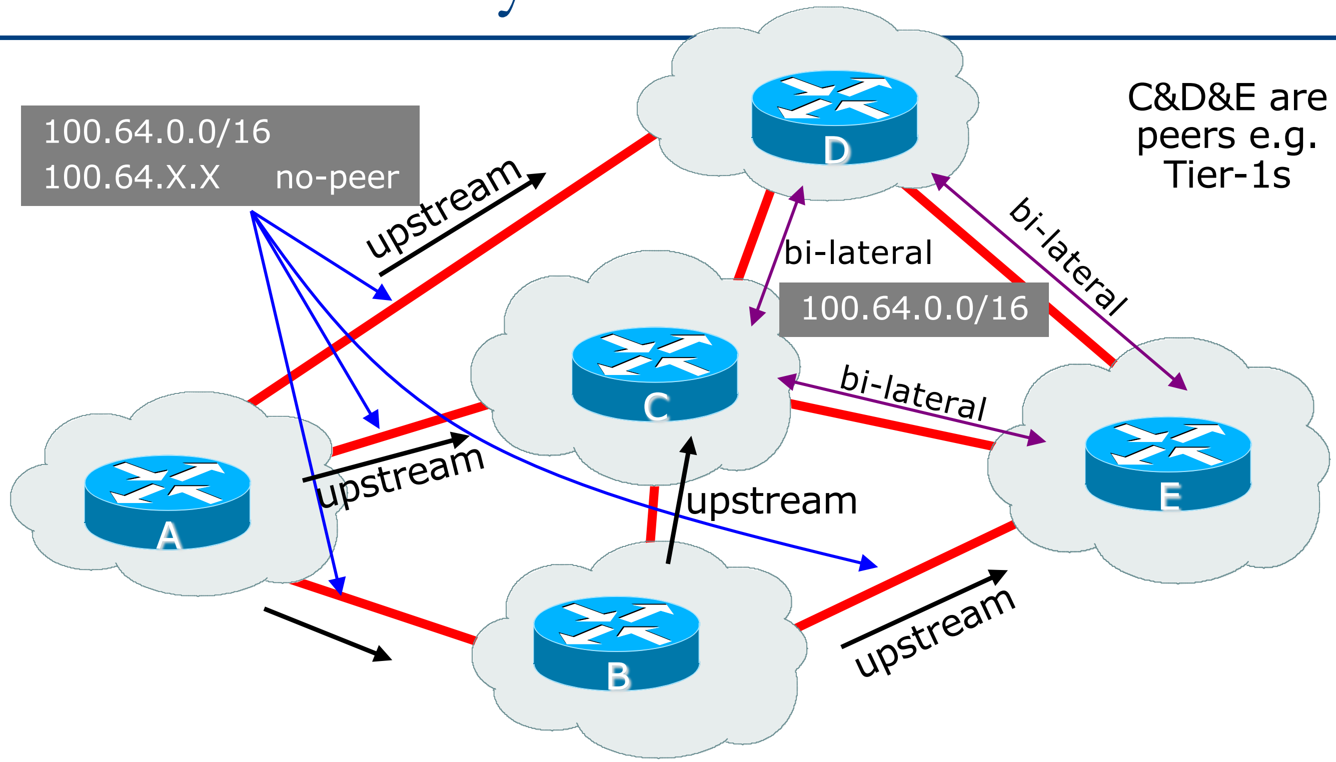
- AS64501 announces aggregate and subprefixes
  - Intention is to improve loadsharing by leaking subprefixes to upstream AS64502 only
- Subprefixes marked with no-export community
- Router G in AS64502 does not announce prefixes with no-export community set

No-Advertise Community



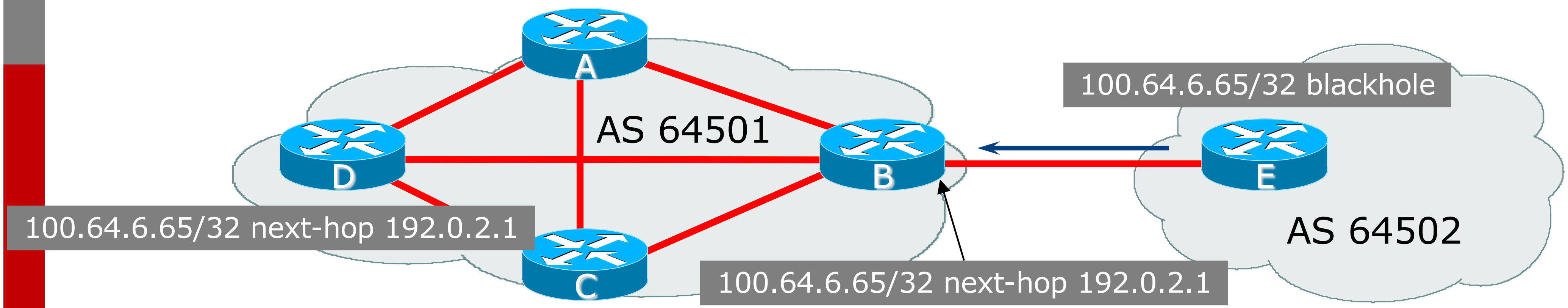
- Used to not advertise a prefix in IBGP
  - B hears 0.0.0.0/0 from EBGP peer E
  - Tags 0.0.0.0/0 as *no-advertise*
  - B will (automatically) not announce prefix to A, C or D
  - Easier/more scalable than using a prefix filter

No-Peer Community



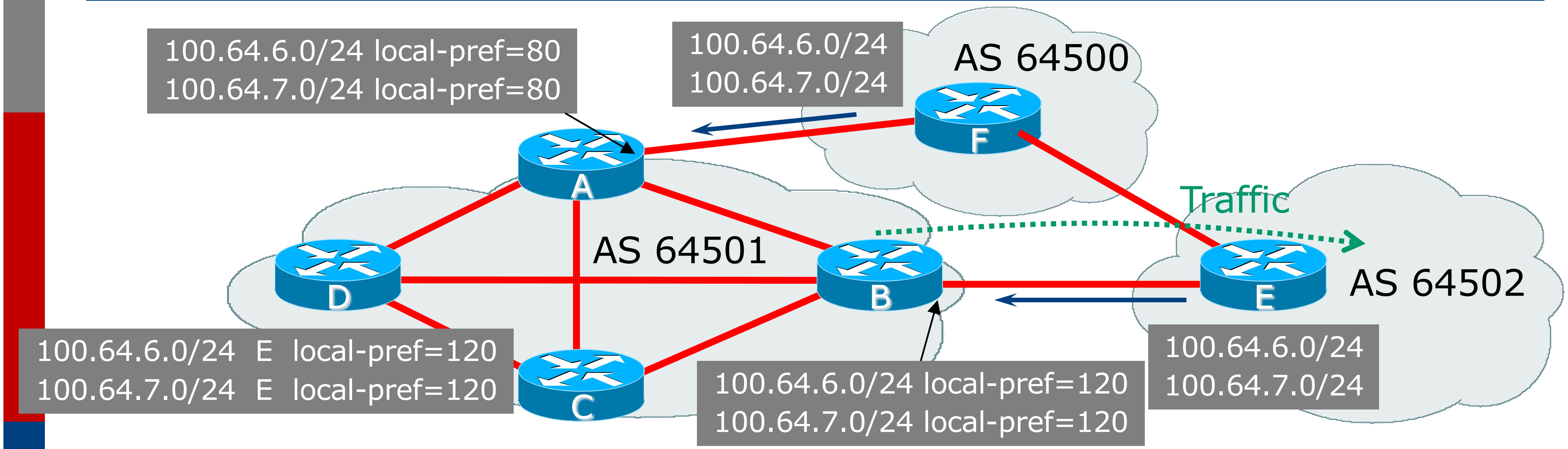
- Sub-prefixes marked with no-peer community are not sent to bi-lateral peers
  - They are only sent to upstream providers

Blackhole Community



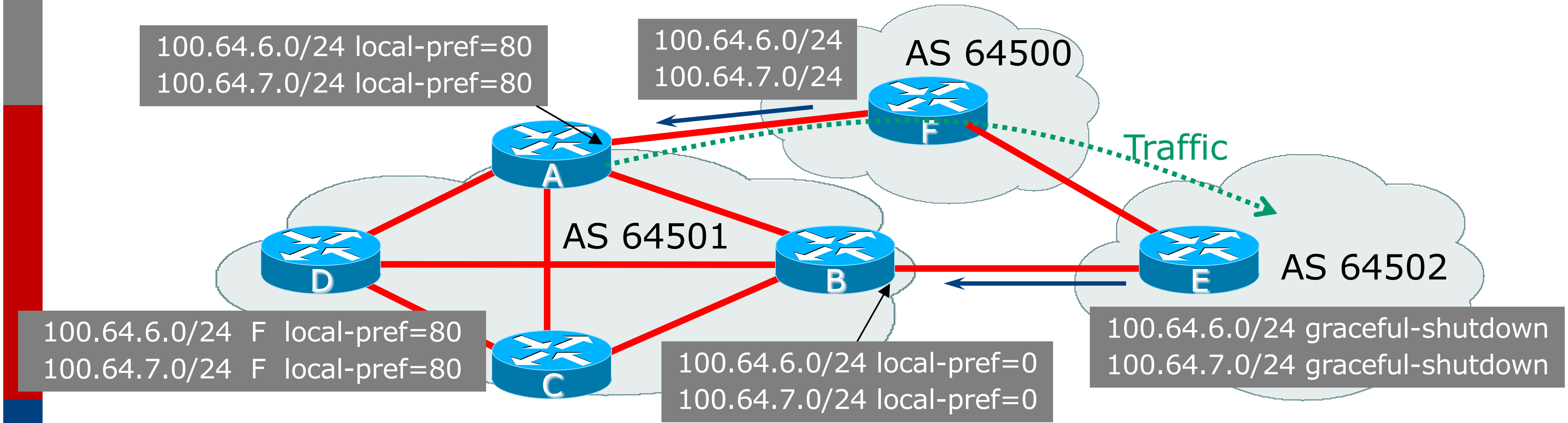
- Used to signal to a BGP neighbour to null route traffic
  - Router E sets *blackhole* community
  - Router B detects *blackhole* community on incoming EBGP announcements and sets *next-hop* to 192.0.2.1
  - 192.0.2.1 is routed to Null interface on all routers within the Autonomous System
  - All traffic to 100.64.6.65 is Null routed

Graceful-Shutdown Community (before)



- Used to inform an EBGP peer that the peering will be going down soon
  - Steady state is primary path between AS64502 and 64501 is via routers E & B
  - AS64502 wants to shutdown direct link, which means traffic will use path via AS64500
  - Graceful-Shutdown ensures that this can be achieved without traffic loss by informing AS64501 that the link is going away

Graceful-Shutdown Community (after)



- Used to inform an EBGP peer that the peering will be going down soon
  - Router E sets *graceful-shutdown* community
  - Router B detects *graceful-shutdown* community on incoming EBGP announcements and sets *local-preference* to 0
  - Best path to 100.64.6.0/24 and 100.64.7.0/24 is now via Router F
  - Allows graceful transition of external best path from Router E to Router F

Service Provider use of Communities

Some examples of how Network Operators make life easier for themselves

BGP Communities

- Communities are generally set at the edge of the service provider network
  - Customer edge: customer prefixes belong to different communities depending on the services they have purchased
  - Internet edge: transit provider prefixes belong to different communities, depending on the loadsharing or traffic engineering requirements of the local Service Provider, or what the demands from its BGP customers might be
- Two simple examples follow to explain the concept

Community Example: Customer Edge

- Service Providers tag prefixes learned from their BGP and static customers with communities
  - To identify services the customer may have purchased
  - To identify prefixes which are part of the Provider’s own address space
  - To identify customer independent address space
  - To control prefix distribution in IBGP
  - To control prefix announcements to customers and upstreams
  - (amongst several other reasons)

Community Example: Customer Edge

- No need to alter filters at the network border when adding a new customer
- New customer simply is added to the appropriate community
  - Border filters already in place take care of announcements
  - $\Rightarrow$ Ease of operation!

Community Example: Customer Edge

Community Value	Description
X:1000	Aggregates
X:1001	Subprefixes of AS X aggregates
X:1005	Static (non-BGP) customer Provider Independent addresses
X:2000	BGP customers who get Transit
X:2100	BGP customers announced to Private Peers
X:2200	BGP customers announced at IXPs
X:2500	BGP customers announced to other BGP customers
X:3100	Routes received from Private Peers
X:3200	Routes received from IXP peers

Community Example: Internet Edge

- This demonstrates how communities might be used at the peering edge of a service provider network
- Service Provider has four types of BGP peers:
  - Customer
  - IXP peer
  - Private peer
  - Transit provider
- The prefixes received from each can be classified using communities
- Customers can opt to receive any or all of the above

Community Example: Internet Edge

- Referring to Customer Edge assignment table:
  - BGP customer who buys local connectivity gets X:2500
  - BGP customer who buys local and IXP connectivity receives community X:2500 and X:3200
  - BGP customer who buys full peer connectivity receives community X:2500, X:3100, and X:3200
- Customer who wants "the Internet" gets everything
  - Gets default route originated by aggregation router
  - Or pays money to get the full BGP table! 😀

Community Example: Internet Edge

- No need to create customised filters when adding customers
  - Border router already sets communities
  - Installation engineers pick the appropriate community set when establishing the customer BGP session
  - $\Rightarrow$ Ease of operation!
- Communities also available for customers to do traffic engineering with Network Operator's peers and upstreams
  - Common examples in the following table

Communities for EBGP

Community Value	Action	Description
X:80	`set local-preference 80`	Backup path
X:120	`set local-preference 120`	Primary path (over-ride BGP path selection default)
X:1	`set as-path prepend X`	Single prepend when announced to X's upstreams
X:2	`set as-path prepend X X`	Double prepend when announced to X's upstreams
X:3	`set as-path prepend X X X`	Triple prepend when announced to X's upstreams
X:666	`set ip next-hop 192.0.2.1`	Blackhole route – very useful for DoS attack mitigation (RFC7999)

Community Example – Summary

- Two examples of customer edge and internet edge can be combined to form a simple community solution for network operator prefix policy control
- More experienced operators tend to have more sophisticated options available
  - Advice is to start with the easy examples given, and then proceed onwards as experience is gained

Network Operator BGP Communities

- There are no recommended Network Operator BGP communities apart from
  - RFC1998
  - The well-known communities
    - www.iana.org/assignments/bgp-well-known-communities
- Efforts have been made to document from time to time
  - Collection of Network Operator communities at www.onesc.net/communities
  - NANOG Tutorial:
    - www.nanog.org/meetings/nanog40/presentations/BGPcommunities.pdf
- Network Operator policy is usually published
  - On the Operator’s website
  - Referenced in the AS Object in the IRR

Example: NTT

Community	Local-Pref	Description
(default)	120	customer
65520:nnnn	50	this community will only set the local preference within the connected country, not beyond
65530:nnnn	50	this community will only set the local preference within the connected region, not beyond
2914:435	50	only beyond the connected country
2914:436	50	only beyond the connected region
2914:450	96	customer fallback
2914:460	98	peer backup
2914:470	100	peer
2914:480	110	customer backup
2914:490	120	customer default
2914:666		**blackhole**

Customers wanting to alter their route announcements to other customers

NTT BGP customers may choose to prepend to all other NTT BGP customers with the following communities:

Community	Description
2914:411	prepends o/b to customer 1x
2914:412	prepends o/b to customer 2x
2914:413	prepends o/b to customer 3x

More info at https://www.gin.ntt.net/support-center/policies-procedures/routing/

Example: Verizon Europe

```
aut-num:        AS702
descr:          Verizon Business EMEA - Commercial IP service provider in Europe
<snip>
remarks:        --------------------------------------------------------------
                Verizon Business filters out inbound prefixes longer than /24.
                We also filter any networks within AS702:RS-INBOUND-FILTER.
                --------------------------------------------------------------
                VzBi uses the following communities with its customers:
                702:80     Set Local Pref 80 within AS702
                702:120    Set Local Pref 120 within AS702
                702:20     Announce only to VzBi AS'es and VzBi customers
                702:30     Keep within Europe, don't announce to other VzBi AS's
                702:1      Prepend AS702 once at edges of VzBi to Peers
                702:2      Prepend AS702 twice at edges of VzBi to Peers
                702:3      Prepend AS702 thrice at edges of VzBi to Peers
                --------------------------------------------------------------
                Advanced communities for customers
                702:7020   Do not announce to AS702 peers with a scope of
                National but advertise to Global Peers, European
                Peers and VzBi customers.
                702:7001   Prepend AS702 once at edges of VzBi to AS702
                peers with a scope of National.
                702:7002   Prepend AS702 twice at edges of VzBi to AS702
                peers with a scope of  National.
<snip>
                --------------------------------------------------------------
                Additional details of the VzBi communities are located at:
                http://www.verizonbusiness.com/uk/customer/bgp/
```

And many more!

Example: Arelion

```
aut-num:        AS1299
descr:          Arelion, f/k/a Telia Carrier
<snip>
remarks:        BGP COMMUNITY SUPPORT FOR AS1299 TRANSIT CUSTOMERS:
remarks:
remarks:        Community Action (default local pref 200)
remarks:        -----------------------------------------
remarks:        1299:50 Set local pref 50 within AS1299 (lowest possible)
remarks:        1299:150 Set local pref 150 within AS1299 (equal to peer, backup)
remarks:        1299:1y050 Set local pref 50 in region y
remarks:        1299:1y150 Set local pref 150 in region y
remarks:        Where y is:
remarks:        0= outside own continent
remarks:        2= Europe
remarks:        5= North America
remarks:        7= Asia Pacific
<snip>
remarks:        European peers
remarks:        Community Action
remarks:        --------- ------
remarks:        1299:200x All peers Europe incl:
remarks:
remarks:        1299:252x NTT/2914
remarks:        1299:253x Zayo/6461
remarks:        1299:254x Orange/5511
remarks:        1299:256x Lumen/3356
remarks:        1299:257x Verizon/702
<snip>
remarks:        Where x is number of prepends (x=0,1,2,3) or do NOT announce (x=9)
```

And many many more!

Example: BT Ignite

```
aut-num:       AS5400
descr:         BT
<snip>
remarks:         Communities scheme:
remarks:         The following BGP communities can be set by BT
remarks:         BGP customers to affect announcements to major peers.
remarks:
remarks:         5400:NXXX
remarks:         N=1          not announce
remarks:         N=2          prepend an extra "5400 5400" on announcement
remarks:         Valid values for XXX:
remarks:         000          All peers and transits
remarks:         500          All transits
remarks:         503          Colt AS3356
remarks:         509          Arelion AS1299
remarks:         002          Sprint AS1239
remarks:         004          Vodafone Global Network AS1273
remarks:         005          Verizon EMEA AS702
remarks:         014          DTAG AS3320
remarks:         016          Orange AS5511
remarks:         018          Tata Communications Ltd AS6453
remarks:         023          GTT Communications AS3257
remarks:         045          Telecom Italia Sparkle AS6762
remarks:         073          GTT Communications AS286
remarks:         169          Cogent AS174
remarks:         177          Telxius Cable AS12956
remarks:         177          Telefonica Germany GmbH AS6805
remarks:         190          Comcast AS7922
remarks:         191          Highwinds Network Group AS12989
<snip>
```

And many more!

Example: Level3

```
aut-num:      AS3356
descr:        Level 3 Communications
<snip>
remarks:        --------------------------------------------------------
remarks:        customer traffic engineering communities - Suppression
remarks:        --------------------------------------------------------
remarks:        64960:XXX - announce to AS XXX if 65000:0
remarks:        65000:0   - announce to customers but not to peers
remarks:        65000:XXX - do not announce at peerings to AS XXX
remarks:        --------------------------------------------------------
remarks:        customer traffic engineering communities - Prepending
remarks:        --------------------------------------------------------
remarks:        65001:0   - prepend once  to all peers
remarks:        65001:XXX - prepend once  at peerings to AS XXX
remarks:        65002:0   - prepend twice to all peers
remarks:        65002:XXX - prepend twice at peerings to AS XXX
<snip>
remarks:        --------------------------------------------------------
remarks:        customer traffic engineering communities - LocalPref
remarks:        --------------------------------------------------------
remarks:        3356:70   - set local preference to 70
remarks:        3356:80   - set local preference to 80
remarks:        3356:90   - set local preference to 90
remarks:        --------------------------------------------------------
remarks:        customer traffic engineering communities - Blackhole
remarks:        --------------------------------------------------------
remarks:        3356:9999 - blackhole (discard) traffic
<snip>
```

And many more!

Creating your own community policy

- Consider creating communities to give policy control to customers
  - Reduces technical support burden
  - Reduces the amount of router reconfiguration, and the chance of mistakes
  - Use previous Network Operator and configuration examples as a guideline

BGP Communities

- There are no “standard communities” for Network Operators
- Best practices today consider that Network Operators should use BGP communities extensively for:
  - Scaling IBGP
  - Multihoming support of traffic engineering
- Look in the Network Operator AS Object in the IRR or on their website for documented community support

BGP Techniques for Network Operators

- BGP Basics
- Scaling BGP
- Using Communities
- Deploying BGP in a Service Provider Network

Deploying BGP in a Service Provider Network

Okay, so we've learned all about BGP now; how do we use it on our network??

Deploying BGP

- The role of IGPs and iBGP
- EBGP default behaviour
- Aggregation
- Receiving Prefixes
- Configuration Tips

The role of IBGP and IGP

Ships in the night?

Or

Good foundations?

BGP versus OSPF/ISIS

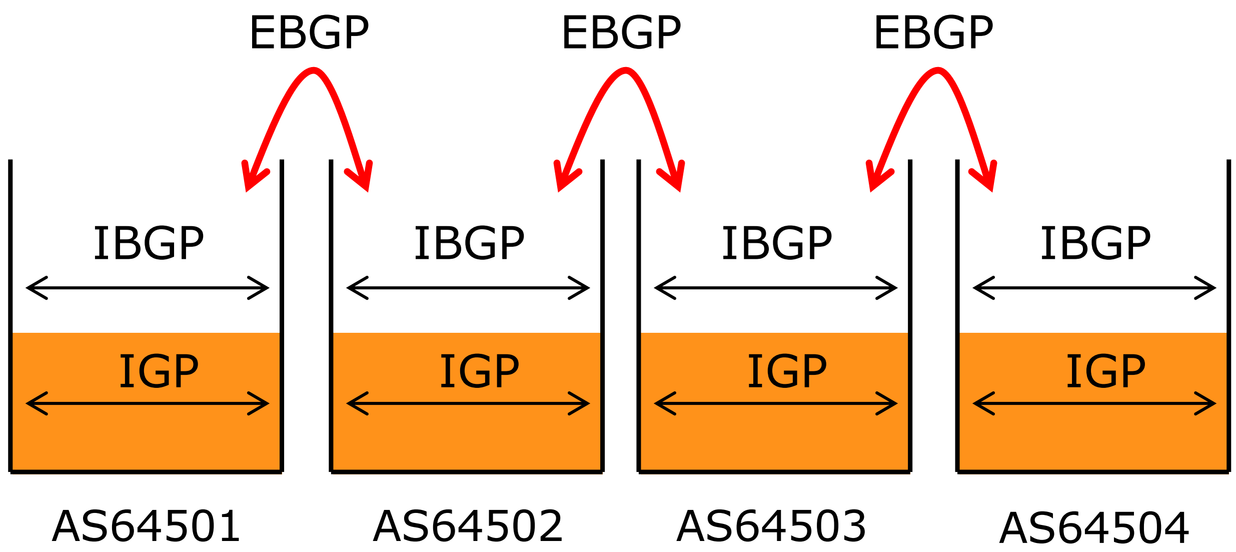
- Internal Routing Protocols (IGPs)
  - Examples are IS-IS and OSPF
  - Used for carrying **infrastructure** addresses
  - NOT used for carrying Internet prefixes or customer prefixes
  - Design goal is to **minimise** number of prefixes in IGP to aid **scalability** and **rapid convergence**

BGP versus OSPF/IS-IS

- BGP is used
  - Internally (IBGP)
  - Externally (EBGP)
- IBGP is used to carry:
  - Some/all Internet prefixes across backbone
  - Customer prefixes
- EBGP is used to:
  - Exchange prefixes with other ASes
  - Implement routing policy

BGP/IGP model used in Service Provider networks

- Model representation



BGP versus OSPF/IS-IS

- DO NOT:
  - Distribute BGP prefixes into an IGP
  - Distribute IGP routes into BGP
  - Use an IGP to carry customer prefixes
- **YOUR NETWORK WILL NOT SCALE**

Injecting prefixes into IBGP

- Use IBGP to carry customer prefixes
  - Don’t ever use IGP
- Point static route to customer interface
- Enter network into BGP process
  - Ensure that implementation options are used so that the prefix always remains in IBGP, regardless of state of interface
  - i.e. avoid IBGP flaps caused by interface flaps

EBGP Default Behaviour

Changing legacy defaults

EBGP Default Behaviour

- Industry standard is described in RFC8212
  - https://tools.ietf.org/html/rfc8212
  - External BGP (EBGP) Route Propagation Behaviour without Policies
- NB: BGP in many implementations is permissive by default
  - This is contrary to industry standard and RFC8212
- Configuring BGP peering without using filters means:
  - All best paths on the local router are passed to the neighbour
  - All routes announced by the neighbour are received by the local router
  - Can have disastrous consequences (see RFC8212)

EBGP Default Behaviour

- Best practice is to ensure that each EBGP neighbour has inbound and outbound filter applied:

```
router bgp 64511
 address-family ipv4
  neighbor 100.64.0.1 remote-as 64510
  neighbor 100.64.0.1 prefix-list as64510-in in
  neighbor 100.64.0.1 prefix-list as64510-out out
  neighbor 100.64.0.1 activate
```

EBGP Default Behaviour

- FRR turns on RFC8212 support by default:
  - https://frrouting.org/

```
frr.pfs.lab(config)# router bgp 64512 view LAB
frr.pfs.lab(config-router)# bgp ?
<snip>
ebgp-requires-policy          Require in and out policy for eBGP peers (RFC8212)
<snip>
```

- No prefixes will be sent or received to external peers in the absence of inbound and outbound policy

Aggregation

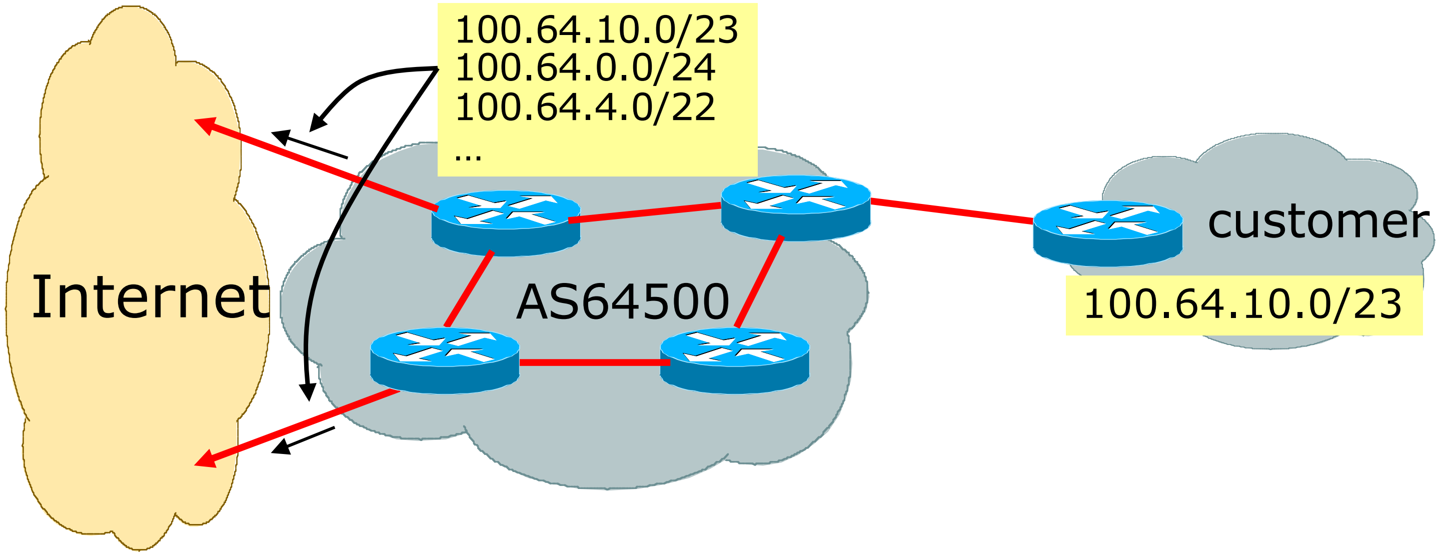
Aggregation

- Aggregation means announcing the address block received from the RIR to the other ASes connected to your network
- Subprefixes of this aggregate may be:
  - Used internally in the provider network
  - Announced to other ASes to aid with multihoming
- Too many operators are still thinking about class Cs, resulting in a proliferation of /24s in the Internet routing table
  - October 2024: 588702 /24s in IPv4 table of 963199 prefixes
- The same is happening for /48s with IPv6
  - October 2024: 97306 /48s in IPv6 table of 203243 prefixes

Announcing an Aggregate

- Network Operators who don't and won't aggregate are held in poor regard by community
- Registries publish their minimum allocation size
  - For IPv4:
    - /24
  - For IPv6:
    - /48 for assignment, /32 for allocation
- Until 2010, there was no real reason to see anything longer than a /22 IPv4 prefix on the Internet. But now?
  - IPv4 run-out is having an impact
  - It is expected that eventually the global IPv4 table will be mostly /24s

Aggregation – Example

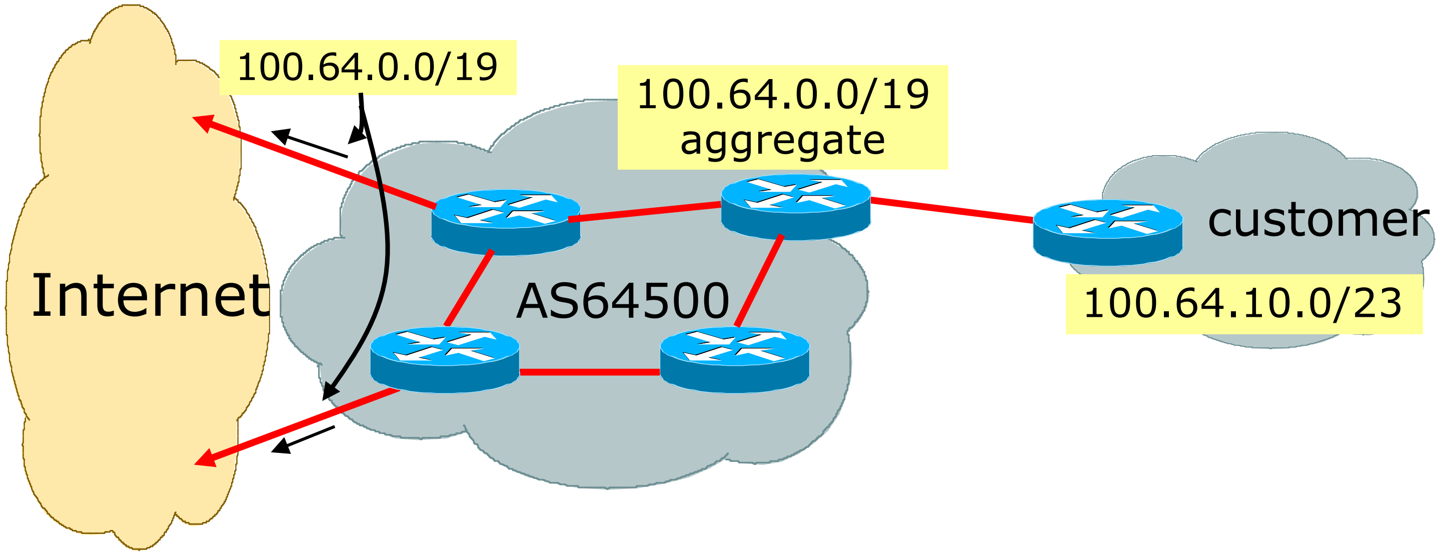


- Customer has /23 network assigned from AS64500's /19 address block
- AS64500 announces customers' individual networks to the Internet

Aggregation – Bad Example

- Customer link goes down
  - Their /23 network becomes unreachable
  - /23 is withdrawn from AS64500's IBGP
- Their Service Provider doesn't aggregate its /19 network block
  - /23 network withdrawal announced to peers
  - Starts rippling through the Internet
  - Added load on all Internet backbone routers as network is removed from routing table
- Customer link returns
  - Their /23 network is now visible to their Service Provider
  - Their /23 network is re-advertised to peers
  - Starts rippling through Internet
  - Load on Internet backbone routers as network is reinserted into routing table
  - Some Network Operators suppress the flaps
  - Internet may take 10-20 min or longer to be visible
  - Where is the Quality of Service???

Aggregation – Example



- Customer has /23 network assigned from AS64500's /19 address block
- AS64500 announced /19 aggregate to the Internet

Aggregation – Good Example

- Customer link goes down
  - Their /23 network becomes unreachable
  - /23 is withdrawn from AS64500's IBGP
- /19 aggregate is still being announced
  - No BGP hold down problems
  - No BGP propagation delays
  - No damping by other network operators
- Customer link returns
- Their /23 network is visible again
  - The /23 is re-injected into AS64500's IBGP
- The whole Internet becomes visible immediately
- Customer has Quality of Service perception

Aggregation – Summary

- Good example is what everyone should do!
  - Adds to Internet stability
  - Reduces size of routing table
  - Reduces routing churn
  - Improves Internet QoS for **everyone**
- Bad example is what too many still do!
  - Why? Lack of knowledge?
  - Laziness?

Separation of IBGP and EBGP

- Many Network Operators do not understand the importance of separating IBGP and EBGP
  - IBGP is where all customer prefixes are carried
  - EBGP is used for announcing aggregate to Internet and for Traffic Engineering
- Do NOT do traffic engineering with customer originated IBGP prefixes
  - Leads to instability similar to that mentioned in the earlier bad example
  - Even though aggregate is announced, a flapping subprefix will lead to instability for the customer concerned
- Generate traffic engineering prefixes on the Border Router

The Internet Today (October 2024)

- Current IPv4 Internet Routing Table Statistics

BGP Routing Table Entries	963199
Prefixes after maximum aggregation	366428
Unique prefixes in Internet	467599
/24s announced	588702
ASNs in use	76301

  - (maximum aggregation is calculated by Origin AS)
  - (unique prefixes > max aggregation means that operators are announcing prefixes from their blocks without a covering aggregate)

The Internet Today (October 2024)

- Current IPv6 Internet Routing Table Statistics

BGP Routing Table Entries	203243
/48s announced	97306
ASNs in use	33488

Efforts to improve aggregation

- The CIDR Report
  - Initiated and operated for many years by Tony Bates
  - Now combined with Geoff Huston’s routing analysis
    - www.cidr-report.org
    - (covers both IPv4 and IPv6 BGP tables)
  - Results e-mailed on a weekly basis to most operations lists around the world
  - Lists the top 30 service providers who could do better at aggregating
- RIPE Routing WG aggregation recommendations
  - IPv4: RIPE-399 — www.ripe.net/ripe/docs/ripe-399.html
  - IPv6: RIPE-532 — www.ripe.net/ripe/docs/ripe-532.html

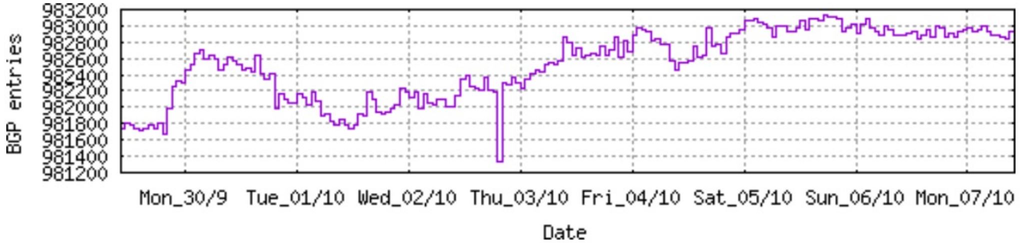
Efforts to Improve Aggregation
The CIDR Report

- Also computes the size of the routing table assuming network operators performed optimal aggregation
- Website allows searches and computations of aggregation to be made on a per AS basis
  - Flexible and powerful tool to aid Network Operators
  - Intended to show how greater efficiency in terms of BGP table size can be obtained without loss of routing and policy information
  - Shows what forms of origin AS aggregation could be performed and the potential benefit of such actions to the total table size
  - Very effectively challenges the traffic engineering excuse

Status Summary

Table History

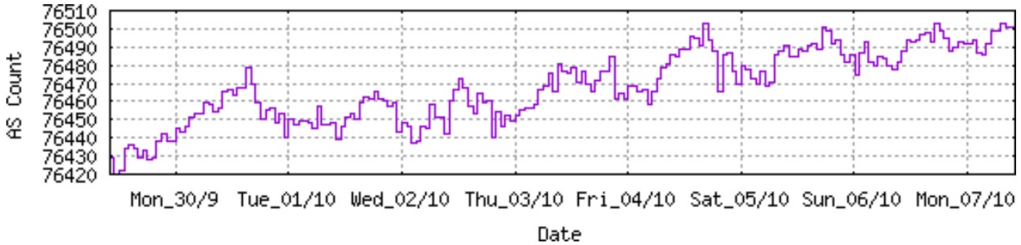
Date	Prefixes	CIDR Aggregated
30-09-24	982307	550818
01-10-24	982058	550748
02-10-24	982183	550106
03-10-24	982295	550684
04-10-24	982692	550561
05-10-24	982961	550344
06-10-24	983019	550218
07-10-24	982954	552866



Plot: BGP Table Size

AS Summary

76507	Number of ASes in routing system
26755	Number of ASes announcing only one prefix
11582	Largest number of prefixes announced by an AS AS8151: UNINET, MX
229178368	Largest address span announced by an AS (/32s) AS749: DNIC-AS-00749, US



Plot: AS count
Plot: Average announcements per origin AS
Report: ASes ordered by originating address span
Report: ASes ordered by transit address span
Report: Autonomous System number-to-name mapping (from Registry WHOIS data)

Announced Prefixes

Rank	AS	Type	Originate Addr Space	(pfx)	Transit Addr space	(pfx)	Description
89	AS6389	ORG+TRN	Originate:	8144640 /9.04	Transit:	49664 /16.40	BELLSOUTH-NET-BLK, US

Long term deaggregator – BellSouth in the US

Aggregation Suggestions

Filter: Aggregates, Specifics

This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

Rank	AS	AS Name	Current	Wthdw	Aggte	Annce	Redctn	%
175	AS6389	BELLSOUTH-NET-BLK, US	652	344	33	341	311	47.70%

Prefix	AS Path	Aggregation Suggestion
12.81.120.0/24	4608 7575 6461 7018 6389	
12.130.209.0/24	4608 7575 6461 7018 6389 6389 6389 6389	
65.5.0.0/16	4608 7575 2914 7018 6389	
65.5.64.0/22	4608 7575 2914 7018 6389	- Withdrawn - matching aggregate 65.5.0.0/16 4608 7575 2914 7018 6389
65.5.118.0/23	4608 7575 2914 7018 6389	- Withdrawn - matching aggregate 65.5.0.0/16 4608 7575 2914 7018 6389
65.5.160.0/21	4608 7575 6461 7018 6389	+ Announce - aggregate of 65.5.160.0/22 (4608 7575 6461 7018 6389) and 65.5.164.0/22 (4608 7575 6461 7018 6389)
65.5.160.0/22	4608 7575 6461 7018 6389	- Withdrawn - aggregated with 65.5.164.0/22 (4608 7575 6461 7018 6389)
65.5.164.0/22	4608 7575 6461 7018 6389	- Withdrawn - aggregated with 65.5.160.0/22 (4608 7575 6461 7018 6389)
65.5.172.0/22	4608 7575 2914 7018 6389	- Withdrawn - matching aggregate 65.5.0.0/16 4608 7575 2914 7018 6389
65.5.200.0/21	4608 7575 2914 7018 6389	- Withdrawn - matching aggregate 65.5.0.0/16 4608 7575 2914 7018 6389
65.5.228.0/22	4608 7575 6461 7018 6389	
65.5.232.0/22	4608 7575 6461 7018 6389	
65.5.236.0/22	4608 7575 2914 7018 6389	- Withdrawn - matching aggregate 65.5.0.0/16 4608 7575 2914 7018 6389
65.5.240.0/22	4608 7575 6461 7018 6389	
65.5.244.0/22	4608 7575 2914 7018 6389	- Withdrawn - matching aggregate 65.5.0.0/16 4608 7575 2914 7018 6389
65.5.248.0/21	4608 7575 6461 7018 6389	+ Announce - aggregate of 65.5.248.0/22 (4608 7575 6461 7018 6389) and 65.5.252.0/22 (4608 7575 6461 7018 6389)
65.5.248.0/22	4608 7575 6461 7018 6389	- Withdrawn - aggregated with 65.5.252.0/22 (4608 7575 6461 7018 6389)
65.5.252.0/22	4608 7575 6461 7018 6389	- Withdrawn - aggregated with 65.5.248.0/22 (4608 7575 6461 7018 6389)
65.6.0.0/15	4608 7575 2914 7018 6389	
65.6.192.0/22	4777 2516 1299 7018 6389	
65.6.196.0/22	4608 7575 6461 7018 6389	
65.7.64.0/18	4608 7575 6461 7018 6389	
65.12.0.0/14	4608 7575 2914 7018 6389	
65.12.32.0/20	4608 7575 6461 7018 6389	
65.13.84.0/22	4608 7575 6461 7018 6389	
65.13.92.0/22	4608 7575 6461 7018 6389	
65.13.120.0/22	4608 7575 2914 7018 6389	- Withdrawn - matching aggregate 65.12.0.0/14 4608 7575 2914 7018 6389
65.13.124.0/22	4777 2516 1299 7018 6389	
65.13.136.0/22	4608 7575 2914 7018 6389	- Withdrawn - matching aggregate 65.12.0.0/14 4608 7575 2914 7018 6389
65.13.176.0/21	4608 7575 2914 7018 6389	- Withdrawn - matching aggregate 65.12.0.0/14 4608 7575 2914 7018 6389
65.13.184.0/21	4608 7575 6461 7018 6389	
65.13.192.0/22	4608 7575 6461 7018 6389	

Long term deaggregator – FPT in Vietnam

Announced Prefixes

```
Rank  AS       Type    Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
121   AS18403        ORG+TRN Originate:    6144000 /9.45   Transit:     435456 /13.27 FPT-AS-AP FPT Telecom Company, VN
```

Aggregation Suggestions

Filter: Aggregates, Specifics

This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

```
 Rank AS          AS Name                                    Current  Wthdw  Aggte  Annce Redctn       %
    9 AS18403     FPT-AS-AP FPT Telecom Company, VN             4449   4033     86    502   3947  88.72%
```

```
 Prefix                AS Path                             Aggregation Suggestion
 1.52.0.0/14           4608 4635 18403 18403 18403
 1.52.0.0/18           4608 4635 18403 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/14 4608 4635 18403 18403 18403
 1.52.0.0/14           4608 4826 18403 18403 + Announce - aggregate of 1.52.0.0/15 (4608 4826 18403 18403) and 1.54.0.0/15 (4608 4826 18403 18403)
 1.52.0.0/20           4608 4826 18403 18403 - Withdrawn - aggregated with 1.52.16.0/20 (4608 4826 18403 18403)
 1.52.0.0/23           4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/20 4608 4826 18403 18403
 1.52.2.0/23           4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/20 4608 4826 18403 18403
 1.52.4.0/24           4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/20 4608 4826 18403 18403
 1.52.5.0/24           4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/20 4608 4826 18403 18403
 1.52.6.0/24           4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/20 4608 4826 18403 18403
 1.52.7.0/24           4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/20 4608 4826 18403 18403
 1.52.8.0/23           4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/20 4608 4826 18403 18403
 1.52.10.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/20 4608 4826 18403 18403
 1.52.11.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/20 4608 4826 18403 18403
 1.52.12.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/20 4608 4826 18403 18403
 1.52.13.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/20 4608 4826 18403 18403
 1.52.14.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/20 4608 4826 18403 18403
 1.52.15.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.0.0/20 4608 4826 18403 18403
 1.52.16.0/20          4608 4826 18403 18403 - Withdrawn - aggregated with 1.52.0.0/20 (4608 4826 18403 18403)
 1.52.16.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
 1.52.17.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
 1.52.18.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
 1.52.19.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
 1.52.20.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
 1.52.21.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
 1.52.22.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
 1.52.23.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
 1.52.24.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
 1.52.25.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
 1.52.26.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
 1.52.27.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
 1.52.28.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
 1.52.29.0/24          4608 4826 18403 18403 - Withdrawn - matching aggregate 1.52.16.0/20 4608 4826 18403 18403
```

Announced Prefixes

```
Rank  AS       Type    Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
131   AS7545         ORG+TRN Originate:    5497856 /9.61   Transit:    3503360 /10.26 TPG-INTERNET-AP TPG Telecom Limited, AU
```

Aggregation Suggestions

Filter: Aggregates, Specifics

Long term deaggregator – TPG in Australia

This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

```
Rank AS            AS Name                                     Current  Wthdw  Aggte  Annce Redctn      %
   4 AS7545        TPG-INTERNET-AP TPG Telecom Limited, AU        5917   5403    161    675   5242  88.59%
```

```
Prefix                AS Path                         Aggregation Suggestion
14.2.0.0/17           4608 7575 7545      + Announce - aggregate of 14.2.0.0/18 (4608 7575 7545) and 14.2.64.0/18 (4608 7575 7545)
14.2.0.0/19           4608 7575 7545      - Withdrawn - aggregated with 14.2.32.0/19 (4608 7575 7545)
14.2.32.0/19          4608 7575 7545      - Withdrawn - aggregated with 14.2.0.0/19 (4608 7575 7545)
14.2.32.0/21          4608 7575 7545      - Withdrawn - matching aggregate 14.2.32.0/19 4608 7575 7545
14.2.40.0/21          4608 7575 7545      - Withdrawn - matching aggregate 14.2.32.0/19 4608 7575 7545
14.2.48.0/21          4608 7575 7545      - Withdrawn - matching aggregate 14.2.32.0/19 4608 7575 7545
14.2.56.0/21          4608 7575 7545      - Withdrawn - matching aggregate 14.2.32.0/19 4608 7575 7545
14.2.64.0/19          4608 7575 7545      - Withdrawn - aggregated with 14.2.96.0/19 (4608 7575 7545)
14.2.96.0/19          4608 7575 7545      - Withdrawn - aggregated with 14.2.64.0/19 (4608 7575 7545)
14.2.128.0/18         4608 7575 7545
14.2.192.0/20         4608 7575 7545
14.200.0.0/14         4608 7575 7545
14.200.0.0/24         4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.1.0/24         4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.2.0/24         4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.3.0/24         4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.4.0/24         4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.5.0/24         4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.6.0/24         4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.7.0/24         4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.8.0/24         4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.9.0/24         4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.10.0/24        4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.11.0/24        4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.12.0/24        4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.13.0/24        4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.14.0/24        4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.15.0/24        4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.16.0/24        4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.17.0/24        4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.18.0/24        4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
14.200.19.0/24        4608 7575 7545      - Withdrawn - matching aggregate 14.200.0.0/14 4608 7575 7545
```

Announced Prefixes

```
Rank  AS        Type    Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
50    AS12479         ORG+TRN Originate:   14329856 /8.23   Transit:     429312 /13.29 UNI2-AS, ES
```

Aggregation Suggestions

Filter: Aggregates, Specifics

Long term deaggregator – Orange in Spain

This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

```
 Rank AS          AS Name                                    Current  Wthdw  Aggte  Annce Redctn       %
    5 AS12479     UNI2-AS, ES                                   7640   5429    709   2920   4720  61.78%
```

```
  Prefix              AS Path                            Aggregation Suggestion
  1.178.224.0/19      4608 1221 4637 5511 12479
  1.178.224.0/21      4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 1.178.224.0/19 4608 1221 4637 5511 12479
  1.178.232.0/21      4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 1.178.224.0/19 4608 1221 4637 5511 12479
  1.178.240.0/20      4777 2516 1299 5511 12479 + Announce - aggregate of 1.178.240.0/21 (4777 2516 1299 5511 12479) and 1.178.248.0/21 (4777 2516 1299 5511 12479)
  1.178.240.0/21      4777 2516 1299 5511 12479 - Withdrawn - aggregated with 1.178.248.0/21 (4777 2516 1299 5511 12479)
  1.178.248.0/21      4777 2516 1299 5511 12479 - Withdrawn - aggregated with 1.178.240.0/21 (4777 2516 1299 5511 12479)
  37.11.0.0/16        4608 1221 4637 5511 12479
  37.11.0.0/22        4777 2516 1299 5511 12479
  37.11.8.0/22        4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 37.11.0.0/16 4608 1221 4637 5511 12479
  37.11.16.0/22       4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 37.11.0.0/16 4608 1221 4637 5511 12479
  37.11.20.0/22       4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 37.11.0.0/16 4608 1221 4637 5511 12479
  37.11.24.0/21       4777 2516 1299 5511 12479 + Announce - aggregate of 37.11.24.0/22 (4777 2516 1299 5511 12479) and 37.11.28.0/22 (4777 2516 1299 5511 12479)
  37.11.24.0/22       4777 2516 1299 5511 12479 - Withdrawn - aggregated with 37.11.28.0/22 (4777 2516 1299 5511 12479)
  37.11.28.0/22       4777 2516 1299 5511 12479 - Withdrawn - aggregated with 37.11.24.0/22 (4777 2516 1299 5511 12479)
  37.11.32.0/22       4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 37.11.0.0/16 4608 1221 4637 5511 12479
  37.11.36.0/22       4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 37.11.0.0/16 4608 1221 4637 5511 12479
  37.11.40.0/22       4777 2516 1299 5511 12479
  37.11.44.0/24       4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 37.11.0.0/16 4608 1221 4637 5511 12479
  37.11.45.0/24       4777 2516 1299 5511 12479
  37.11.46.0/23       4777 2516 1299 5511 12479
  37.11.48.0/22       4777 2516 1299 5511 12479
  37.11.52.0/22       4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 37.11.0.0/16 4608 1221 4637 5511 12479
  37.11.56.0/23       4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 37.11.0.0/16 4608 1221 4637 5511 12479
  37.11.58.0/23       4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 37.11.0.0/16 4608 1221 4637 5511 12479
  37.11.60.0/22       4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 37.11.0.0/16 4608 1221 4637 5511 12479
  37.11.64.0/22       4777 2516 1299 5511 12479
  37.11.68.0/22       4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 37.11.0.0/16 4608 1221 4637 5511 12479
  37.11.72.0/22       4777 2516 1299 5511 12479
  37.11.76.0/22       4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 37.11.0.0/16 4608 1221 4637 5511 12479
  37.11.80.0/22       4608 1221 4637 5511 12479 - Withdrawn - matching aggregate 37.11.0.0/16 4608 1221 4637 5511 12479
  37.11.88.0/21       4777 2516 1299 5511 12479 + Announce - aggregate of 37.11.88.0/22 (4777 2516 1299 5511 12479) and 37.11.92.0/22 (4777 2516 1299 5511 12479)
  37.11.88.0/22       4777 2516 1299 5511 12479 - Withdrawn - aggregated with 37.11.92.0/22 (4777 2516 1299 5511 12479)
```

Importance of Aggregation

- Size of routing table
  - Router Memory is not so much of a problem as it was in the 1990s
  - Routers routinely carry over 2 million prefixes
- Convergence of the Routing System
  - This is a problem
  - Bigger table takes longer for CPU to process
  - BGP updates take longer to deal with
  - BGP Instability Report tracks routing system update activity
  - bgpupdates.potaroo.net/instability/bgpupd.html

The BGP Instability Report

The BGP Instability Report is updated daily. This report was generated on 07 October 2024 06:23 (UTC+1000)

50 Most active ASes for the past 14 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	16509	408962	3.76%	10998	37.19	AMAZON-02, US
2	8151	229838	2.12%	11862	19.38	UNINET, MX
3	19429	201817	1.86%	693	291.22	ETB - Colombia, CO
4	9829	157254	1.45%	2004	78.47	BSNL-NIB National Internet Backbone, IN
5	14754	115042	1.06%	455	252.84	TELECOMUNICACIONES DE GUATEMALA, SOCIEDAD ANONIMA, GT
6	45899	108995	1.00%	3267	33.36	VNPT-AS-VN VNPT Corp, VN
7	12849	95536	0.88%	586	163.03	HOTNET-IL HOTmobile, IL
8	264681	94435	0.87%	52	1816.06	Sociedad de Telecomunicaciones Netsouth SPA, CL
9	36903	92733	0.85%	1249	74.25	MT-MPLS, MA
10	5639	86473	0.80%	170	508.66	Telecommunication Services of Trinidad and Tobago, TT
11	7552	73764	0.68%	3991	18.48	VIETEL-AS-AP Viettel Group, VN
12	39891	70552	0.65%	4602	15.33	ALJAWWALSTC-AS, SA
13	4155	69488	0.64%	2155	32.25	USDA-1, US
14	58224	69381	0.64%	1439	48.21	TCI, IR
15	6057	68990	0.64%	578	119.36	Administracion Nacional de Telecomunicaciones, UY
16	42337	66435	0.61%	710	93.57	RESPINA-AS, IR
17	10620	65157	0.60%	3546	18.37	Telmex Colombia S.A., CO
18	647	62366	0.57%	415	150.28	DNIC-ASBLK-00616-00665, US
19	45271	57905	0.53%	878	65.95	ICLNET-AS-AP Idea Cellular Limited, IN
20	36914	55012	0.51%	545	100.94	KENET-AS, KE
21	149038	51590	0.47%	2	25795.00	UCGCL-AS-AP UNIQUE COMM GROUP COMPANY LIMITED, MM
22	12220	51235	0.47%	8	6404.38	I-EVOLVE-TECHNOLOGY-SERVICES, US
23	367	51094	0.47%	2874	17.78	DNIC-ASBLK-00306-00371, US
24	37187	48152	0.44%	55	875.49	SKYBAND, MW

50 Most active Prefixes for the past 14 days

RANK	PREFIX	UPDs	%	Origin AS -- AS NAME
1	66.133.36.0/24	34009	0.30%	12684 -- SES-LUX-AS, LU
2	170.244.214.0/24	25966	0.23%	266508 -- PONTO WIFI LTDA ME, BR
3	103.177.86.0/24	25924	0.23%	149038 -- UCGCL-AS-AP UNIQUE COMM GROUP COMPANY LIMITED, MM
4	103.177.87.0/24	25666	0.23%	149038 -- UCGCL-AS-AP UNIQUE COMM GROUP COMPANY LIMITED, MM
5	143.255.59.0/24	22630	0.20%	33182 -- DIMENOC, US
6	103.248.132.0/22	19934	0.18%	132829 -- NCPL-AS-AP Nevigate Communications S Pte Ltd, SG
7	45.172.92.0/22	19784	0.18%	265566 -- TELESISTEMAS PENINSULARES SA DE CV, MX
8	197.216.59.0/24	18803	0.17%	11259 -- ANGOLATELECOM, AO
9	107.154.97.0/24	18720	0.17%	19551 -- INCAPSULA, US
10	146.71.102.0/24	17917	0.16%	53850 -- GORILLASERVERS, US
11	103.223.2.0/24	15446	0.14%	135445 -- IDNIC-AIRPAY-AS-ID PT. Airpay International Indonesia, ID
12	124.195.190.0/24	15310	0.14%	38684 -- CMBDAEJEON-AS-KR CMB Daejeon Broadcasting Co,.Ltd, KR
13	186.227.7.0/24	14310	0.13%	262765 -- Net Facil Sistemas Eletronicos Ltda ME, BR
14	202.181.232.0/23	13724	0.12%	7540 -- HKCIX-AS-AP HongKong Commercial Internet Exchange, HK
15	207.167.116.0/22	13480	0.12%	7954 -- IMMENSE-NETWORKS, US
16	45.129.17.0/24	13234	0.12%	208417 -- EONSCOPE, US
17	185.32.70.0/24	12790	0.11%	51269 -- HEXATOM, FR
18	112.198.160.0/22	12633	0.11%	4775 -- GLOBE-TELECOM-AS Globe Telecoms, PH
19	138.84.126.0/23	12581	0.11%	4775 -- GLOBE-TELECOM-AS Globe Telecoms, PH
20	138.99.97.0/24	12285	0.11%	28657 -- MD Brasil - Tecnologia da Informacao Ltda, BR
21	185.18.201.0/24	11132	0.10%	47855 -- PRIME Moscow branch, RU
22	130.137.230.0/24	11066	0.10%	16509 -- AMAZON-02, US
23	130.137.63.0/24	11056	0.10%	16509 -- AMAZON-02, US
24	72.237.213.0/24	10922	0.10%	12220 -- I-EVOLVE-TECHNOLOGY-SERVICES, US
25	72.43.207.0/24	10922	0.10%	12220 -- I-EVOLVE-TECHNOLOGY-SERVICES, US
26	72.237.212.0/24	10903	0.10%	12220 -- I-EVOLVE-TECHNOLOGY-SERVICES, US
27	177.72.32.0/21	10891	0.10%	262540 -- CBNET TELECOM EIRELI, BR
28	189.90.24.0/22	10607	0.09%	265141 -- RBT Internet, BR
29	209.22.66.0/24	10388	0.09%	2046 -- DNIC-AS-02046, US

The BGP IPv6 Instability Report

This report is updated daily. The current report was generated on 7 October 2024 01:14 (UTC+1000)

---

50 Most active ASes for the past 14 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	210842	287064	7.39%	196	1464.61	RKZED-AS, ID
2	11172	175698	4.52%	7186	24.45	Alestra, S. de R.L. de C.V., MX
3	20473	165251	4.25%	1687	97.96	AS-VULTR, US
4	16509	138514	3.57%	5293	26.17	AMAZON-02, US
5	52965	107824	2.78%	119	906.08	1TELECOM SERVICOS DE TECNOLOGIA EM INTERNET LTDA, BR
6	40138	98245	2.53%	42	2339.17	MDNET, US
7	8151	84918	2.19%	435	195.21	UNINET, MX
8	53122	65670	1.69%	3	21890.00	super midia tv a cabo ltda, BR
9	14080	61011	1.57%	232	262.98	Telmex Colombia S.A., CO
10	272112	59484	1.53%	96	619.62	TELECABLE DOMINICANO, S.A., DO
11	4767	57768	1.49%	1	57768.00	AIT-CS-ASN Computer Science, TH
12	263390	52151	1.34%	23	2267.43	FNT Telecomunicacoes e Acesso a Redes de Internet, BR
13	400339	50763	1.31%	6	8460.50	TRINITY-CYBER-01, US
14	7545	50714	1.31%	2950	17.19	TPG-INTERNET-AP TPG Telecom Limited, AU
15	53667	39792	1.02%	1172	33.95	PONYNET, US
16	202256	38991	1.00%	576	67.69	LAWLIETNET, CN
17	42298	37421	0.96%	662	56.53	GCC-MPLS-PEERING GCC MPLS peering, QA
18	7296	35165	0.91%	4	8791.25	AS7296, US
19	21664	32782	0.84%	31	1057.48	AMZN-BYOASN, US
20	263608	31376	0.81%	5	6275.20	WSNET TELECOM LTDA ME, BR
21	149697	31340	0.81%	24	1305.83	GIS-AS-ID PT Global Internet Solusindo, ID
22	36969	28395	0.73%	5	5679.00	MTL-AS, MW
23	37693	24318	0.63%	343	70.90	TUNISIANA, TN
24	60539	23503	0.60%	2050	11.46	HUICAST_TELECOM, HK
25	206569	22572	0.58%	3	7524.00	PAULHENRI-ZIMMERLIN, FR

50 Most active Prefixes for the past 14 days

RANK	PREFIX	UPDs	%	Origin AS -- AS NAME
1	2403:e240::/32	57768	1.40%	4767 -- AIT-CS-ASN Computer Science, TH
2	2606:ab40:100::/48	50748	1.23%	400339 -- TRINITY-CYBER-01, US
3	2606:6e00:8000::/35	35099	0.85%	7296 -- AS7296, US
4	2804:fdc::/32	31372	0.76%	263608 -- WSNET TELECOM LTDA ME, BR
5	2804:1e10:fffe::/48	24059	0.58%	53122 -- super midia tv a cabo ltda, BR
6	2804:1e10:ffff::/48	21774	0.53%	53122 -- super midia tv a cabo ltda, BR
7	2804:1e10::/32	19838	0.48%	53122 -- super midia tv a cabo ltda, BR
8	2a13:79c0:100::/40	19618	0.47%	200235 -- CTO-EXTERNE cto-externe, FR
9	2407:5440::/48	19242	0.47%	141145 -- GIGANET-AS-ID PT Giga Digital Nusantara, ID
10	2804:8b6c:1000::/37	18837	0.46%	273731 -- MGDATA TECNOLOGIA LTDA, BR
11	2a0c:b641:302::/47	17562	0.42%	204210 -- ZEUSPACKAGINGLTD, IE
12	2a0e:97c0:78f::/48	16443	0.40%	210397 -- WOLKEN-AS, DE
13	2602:fbbc::/46	15366	0.37%	400498 -- BPLLC-AS-01, US
14	2804:393c:7700::/40	14825	0.36%	266020 -- ICLICK TELECOM, BR
15	2804:6f8:c3e8::/48	14666	0.35%	52848 -- IAGENTE SISTEMAS PARA COMUNICACAO, BR
16	2001:43f8:d60::/48	13662	0.33%	328162 -- ICOLO, KE
17	2a0e:8f02:f06e::/48	13478	0.33%	214959 -- DEUTNET, FR
18	2a10:ccc0:cccc::/46	12156	0.29%	151194 -- STELIGHT-AS-AP Zhu Yucheng, CN
19	2402:9880:500::/40	11978	0.29%	58744 -- BODYTRACE-HK Mirror Tower, HK
20	2c0f:f988::/32	11367	0.28%	37353 -- SEACOM-AS, ZA
21	2404:2280:15b::/48	10940	0.26%	24429 -- TAOBAO Zhejiang Taobao Network Co.,Ltd, CN
22	2806:202::/32	9185	0.22%	28458 -- IENTC S DE RL DE CV, MX
23	2402:e580:73f7::/48	8910	0.22%	40138 -- MDNET, US
24	2a0a:6044:bb02::/48	8635	0.21%	215956 -- MYIP-AS MyIP.be ASN, BE
25	2605:9cc0:c0f::/48	8413	0.20%	16509 -- AMAZON-02, US 21664 -- AMZN-BYOASN, US
26	2605:9cc0:c08::/48	8351	0.20%	16509 -- AMAZON-02, US 21664 -- AMZN-BYOASN, US
27	2605:9cc0:c0a::/48	8176	0.20%	16509 -- AMAZON-02, US 21664 -- AMZN-BYOASN, US

Aggregation: Summary

- Aggregation on the Internet could be MUCH better
  - 50% saving on Internet routing table size is quite feasible
  - Tools are available
  - Commands on the routers are not hard
  - CIDR-Report webpage

Receiving Prefixes

Receiving Prefixes

- There are three scenarios for receiving prefixes from other ASes
  - Customer talking BGP
  - Peer talking BGP
  - Upstream/Transit talking BGP
- Each has different filtering requirements and need to be considered separately

Receiving Prefixes: From Customers

- Network Operators must only accept prefixes which have been assigned or allocated to their downstream customer
- If Network Operator has assigned address space to its customer, then the customer IS entitled to announce it back to their Network Operator
- If the Network Operator has NOT assigned address space to its customer, then:
  - Check in the five RIR databases to see if this address space really has been assigned to the customer
  - The tool: whois –h jwhois.apnic.net x.x.x.0/24
    - (jwhois is "joint whois" and queries the 5 RIR databases)

Receiving Prefixes: From Customers

- Example use of whois to check if customer is entitled to announce address space:

```
$ whois -h jwhois.apnic.net 202.12.29.0

inetnum:        202.12.29.0 - 202.12.29.255
netname:        APNIC-SERVICES-AU
descr:          Asia Pacific Network Information Centre
descr:          Regional Internet Registry for the Asia-Pacific Region
descr:          6 Cordelia Street
descr:          South Brisbane
geoloc:         27.4731138 153.0141194
country:        AU
admin-c:        AIC1-AP
tech-c:         AIC1-AP
mnt-by:         APNIC-HM
mnt-irt:        IRT-APNIC-IS-AP
status:         ASSIGNED PORTABLE
changed:        hm-changed@apnic.net 20170327
changed:        hm-changed@apnic.net 20170331
source:         APNIC
```

inetnum – means it is an address delegation to an entity

Portable – means its an assignment to the customer, the customer can announce it to you

Receiving Prefixes: From Customers

- Example use of whois to check if customer is entitled to announce address space:

```
$ whois -h jwhois.apnic.net 194.15.141.0

inetnum:        194.15.141.0 - 194.15.141.255
netname:        INETTECH
country:        SE
org:            ORG-ITAS2-RIPE
admin-c:        KEL5-RIPE
tech-c:         KEL5-RIPE
status:         ASSIGNED PI
mnt-by:         RIPE-NCC-END-MNT
mnt-by:         KURTIS-PP-MNT
mnt-routes:     KURTIS-PP-MNT
mnt-domains:    KURTIS-PP-MNT
created:        2003-12-04T09:33:09Z
last-modified:  2016-04-14T08:21:55Z
source:         RIPE
sponsoring-org: ORG-NIE1-RIPE
```

inetnum – means it is an address delegation to an entity

Assigned PI – means its an assignment to the customer, the customer can announce it to you

Receiving Prefixes: From Customers

- Example use of whois to check if customer is entitled to announce address space:

```
$ whois -h jwhois.apnic.net 193.128.0.0/22

inetnum:        193.128.0.0 - 193.128.6.255
netname:        UK-PIPEX-19931014
country:        GB
org:            ORG-UA24-RIPE
admin-c:        WERT1-RIPE
tech-c:         UPHM1-RIPE
status:         ALLOCATED PA
remarks:        Please send abuse notification to abuse@uk.uu.net
mnt-by:         RIPE-NCC-HM-MNT
mnt-by:         AS1849-MNT
mnt-routes:     AS1849-MNT
mnt-routes:     WCOM-EMEA-RICE-MNT
mnt-irt:        IRT-MCI-GB
created:        2018-07-30T09:42:04Z
last-modified:  2018-07-30T09:42:04Z
source:         RIPE # Filtered
```

inetnum – means it is an address delegation to an entity

ALLOCATED – means that this is Provider Aggregatable address space and can only be announced by the provider holding the allocation (in this case Verizon UK)

Receiving Prefixes: From Peers

- A peer is a Network Operator with whom you agree to exchange prefixes you originate into the Internet routing table
  - Prefixes you accept from a peer are only those they have indicated they will announce
  - Prefixes you announce to your peer are only those you have indicated you will announce

Receiving Prefixes: From Peers

- Agreeing what each will announce to the other:
  - Exchange of e-mail documentation as part of the peering agreement, and then ongoing updates

    OR
  - Use of the Internet Routing Registry and configuration tools such as:
    - IRRToolSet: https://github.com/irrtoolset/irrtoolset
    - bgpq4: https://github.com/bgp/bgpq4
      (uses NTT's IRR database by default)

Receiving Prefixes: From Upstream/Transit Provider

- Upstream/Transit Provider is a Network Operator who you pay to give you transit to the WHOLE Internet
- Receiving prefixes from them is not desirable unless really necessary
  - Traffic Engineering
- Ask upstream/transit provider to either:
  - originate a default-route

    OR
  - announce one prefix you can use as default

Receiving Prefixes: From Upstream/Transit Provider

- If it is necessary to receive prefixes from any provider, care is required.
  - Don't accept default (unless you need it)
  - Don't accept your own prefixes
- Special use prefixes for IPv4 and IPv6:
  - http://www.rfc-editor.org/rfc/rfc6890.txt
- For IPv4:
  - Don't accept prefixes longer than /24 (?)
    - /24 was the historical class C
- For IPv6:
  - Don't accept prefixes longer than /48 (?)
    - /48 is the design minimum delegated to a site

Receiving Prefixes: From Upstream/Transit Provider

- Check Team Cymru's list of "bogons"
  - http://www.team-cymru.com/bogon-reference-http
- For IPv4 also consult:
  - https://www.rfc-editor.org/rfc/rfc6441.txt (BCP171)
- Bogon Route Server:
  - https://www.team-cymru.com/bogon-reference-bgp
  - Supplies a BGP feed (IPv4 and/or IPv6) of address blocks which should not appear in the BGP table

Receiving IPv4 Prefixes

```
deny 0.0.0.0/0                      ! Default
deny 0.0.0.0/8 to /32               ! RFC1122 local host
deny 10.0.0.0/8 to /32              ! RFC1918
deny 100.64.0.0/10 to /32           ! RFC6598 shared address
deny 101.10.0.0/19 to /32           ! Local prefix
deny 127.0.0.0/8 to /32             ! Loopback
deny 169.254.0.0/16 to /32          ! Auto-config
deny 172.16.0.0/12 to /32           ! RFC1918
deny 192.0.0.0/24 to /32            ! RFC6598 IETF protocol
deny 192.0.2.0/24 to /32            ! TEST1
deny 192.168.0.0/16 to /32          ! RFC1918
deny 198.18.0.0/15 to /32           ! Benchmarking
deny 198.51.100.0/24 to /32         ! TEST2
deny 203.0.113.0/24 to /32          ! TEST3
deny 224.0.0.0/3 to /32             ! Multicast & Experimental
deny 0.0.0.0/0 from /25 to /32      ! Prefixes >/24
deny subnets of your own address space
permit everything else
```

Receiving IPv6 Prefixes

```
permit 64:ff9b::/96              ! RFC6052 v4v6trans
deny 2001::/23 to /128           ! RFC2928 IETF prot
deny 2001:2::/48 to /128         ! Benchmarking
deny 2001:10::/28 to /128        ! ORCHID
deny 2001:db8::/32 to /128       ! Documentation
deny 2002::/16 to /128           ! Deny all 6to4
deny 2020:3030::/32 to /128      ! Local Prefix
deny 3ffe::/16 to /128           ! Old 6bone
deny subnets of your own address space
permit 2000::/3 to 48            ! Global Unicast
deny ::/0 to /128                ! Deny everything else
```

Note: These filters block Teredo (serious security risk) and 6to4 (deprecated by RFC7526)

Receiving Prefixes

- Paying attention to prefixes received from customers, peers and transit providers assists with:
  - The integrity of the local network
  - The integrity of the Internet
- Responsibility of all Network Operators to be good Internet citizens

Receiving BGP attributes

Receiving BGP attributes

- BGP attributes are sent as part of the BGP updates for each prefix
- Common attributes operators need to be aware of, for routing best practice, are:
  - MED
  - AS numbers (only public ASNs are routable)
  - BGP Communities

Receiving Prefixes: MEDs?

- MEDs are used by EBGP neighbours to indicate preferred entry point into their network over two or more links with their neighbour
  - Allows the operator to determine entry path into their network
    - Might have unintended consequences within their peer's network
  - Many operators will override MEDs attached to BGP announcements by setting their own local-preference values

Receiving Prefixes: Bogon ASNs?

- What about prefixes originated by bogon AS numbers?
  - Public ranges are 1-64495 (excluding 23456) and 131072-458751
    - IANA is distributing AS blocks to the RIRs from the latter range
  - All other ASNs are either for documentation, or for private use, or are unassigned
    - And any prefixes originating from those need to be dropped
    - Configuration error? Malicious intent?
- What would the AS_PATH filter look like?
  - Challenging with regular expression (as per IOS)
  - Easier with AS ranges (as per Bird or JunOS)

Filtering bogon ASNs – BIRD

- Here is a function showing how to filter bogon ASNs, as described previously:

```
function as_path_contains_bogons()
int set invalid_asns;
{
    invalid_asns = [
        0,                         # Reserved
        23456,                     # Transition AS
        64496..64511,              # Documentation ASNs
        64512..65534,              # Private ASNs
        65535,                     # Reserved
        65536..65551,              # Documentation ASNs
        65552..131071,             # Reserved
        458752..4199999999,        # IANA Reserved
        4200000000..4294967294,    # Private ASNs
        4294967295                 # Reserved
    ];
    return bgp_path ~ invalid_asns;
}
```

Filtering bogon ASNs – FRR

- Here is an AS-PATH regexp showing how to filter bogon ASNs:

```
bgp as-path access-list Bogon_ASNs deny _0_
bgp as-path access-list Bogon_ASNs deny _23456_
bgp as-path access-list Bogon_ASNs deny _6449[6-9]_
bgp as-path access-list Bogon_ASNs deny _64[5-9][0-9][0-9]_
bgp as-path access-list Bogon_ASNs deny _6[5-9][0-9][0-9][0-9]_
bgp as-path access-list Bogon_ASNs deny _[7-9][0-9][0-9][0-9][0-9]_
bgp as-path access-list Bogon_ASNs deny _1[0-2][0-9][0-9][0-9][0-9]_
bgp as-path access-list Bogon_ASNs deny _130[0-9][0-9][0-9]_
bgp as-path access-list Bogon_ASNs deny _1310[0-6][0-9]_
bgp as-path access-list Bogon_ASNs deny _13107[0-1]_
bgp as-path access-list Bogon_ASNs deny _45875[2-9]_
bgp as-path access-list Bogon_ASNs deny _4587[6-9][0-9]_
bgp as-path access-list Bogon_ASNs deny _458[8-9][0-9][0-9]_
bgp as-path access-list Bogon_ASNs deny _459[0-9][0-9][0-9]_
bgp as-path access-list Bogon_ASNs deny _4[6-9][0-9][0-9][0-9][0-9]_
bgp as-path access-list Bogon_ASNs deny _[5-9][0-9][0-9][0-9][0-9][0-9]_
bgp as-path access-list Bogon_ASNs deny _[0-9][0-9][0-9][0-9][0-9][0-9][0-9]_
bgp as-path access-list Bogon_ASNs deny _[0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9]_
bgp as-path access-list Bogon_ASNs deny _[0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9]_
bgp as-path access-list Bogon_ASNs deny _[0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9]_
bgp as-path access-list Bogon_ASNs permit .*
```

Receiving Prefixes: BGP Communities?

- BGP communities are attached to BGP announcements to indicate:
  - Internal policy within an AS
  - External policy supported by a peer, for:
    - Onward routing policy/traffic engineering
    - Filtering (eg Remotely Triggered Blackhole Filtering)
    - Traffic engineering between the two networks
- Different BGP implementations have different default BGP community behaviours – consult:
  - Vendor documentation
  - https://www.rfc-editor.org/rfc/rfc8642.txt for discussion of some of the issues operators need to be aware of

Receiving Prefixes: BGP Communities

- Don't accept community values that are not expected
  - Match expected values
  - Overwrite received community values with your own default value

```
ip community-list standard lp-250 permit 65534:250
!
route-map ebgp-import permit 5
 description Set high preference
 match community lp-250
 set local-preference 250
 set community 65534:100
!
route-map ebgp-import permit 10
 description Set our default community
 set community 65534:100
!
```

We only expect this community from our EBGP neighbour

Cisco IOS: this overwrites all incoming community values

Receiving Prefixes: BGP Communities

- Don't send community values that are not needed by peer
  - This avoids propagating your internal communities to other networks

```
route-map ebgp-export permit 5
 description Tell upstream to set local-pref 250
 set community 65534:250
!
```

Cisco IOS: this overwrites all other community values

- Propagate all communities within the AS (by IBGP)
  - This may need changes to your equipment's default!

Receiving BGP attributes

- Care is needed when receiving prefixes, to be aware of some of the optional BGP attributes that may be attached
  - BGP communities are only intended for policy decisions within an AS or between two peering ASes
  - MEDs may have unexpected consequences for traffic flows on the peer's network
  - Bogon ASNs, like bogon address space, must never be used or announced to the global Internet

Interconnection Best Practices

PeeringDB and the Internet Routing Registry

Interconnection Best Practices

- Types of Peering
- Using the PeeringDB and IXPDB
- Using the Internet Routing Registry

Types of Peering (1)

- Private Peering
  - Where two network operators agree to interconnect their networks, and exchange their respective routes, for the purpose of ensuring their customers can reach each other directly over the peering link
- Settlement Free Peering
  - No traffic charges
  - **The most common form of peering**
- Paid Peering
  - Where two operators agree to exchange traffic charges for a peering relationship

Types of Peering (2)

- Bi-lateral Peering
  - Very similar to Private Peering, but usually takes place at a public peering point (IXP)
- Multilateral Peering
  - Takes place at Internet Exchange Points, where operators all peer with each other via a Route Server
- Mandatory Multilateral Peering
  - Where operators are forced to peer with each other as condition of IXP membership
  - **Strongly discouraged: Has no record of success**

Types of Peering (3)

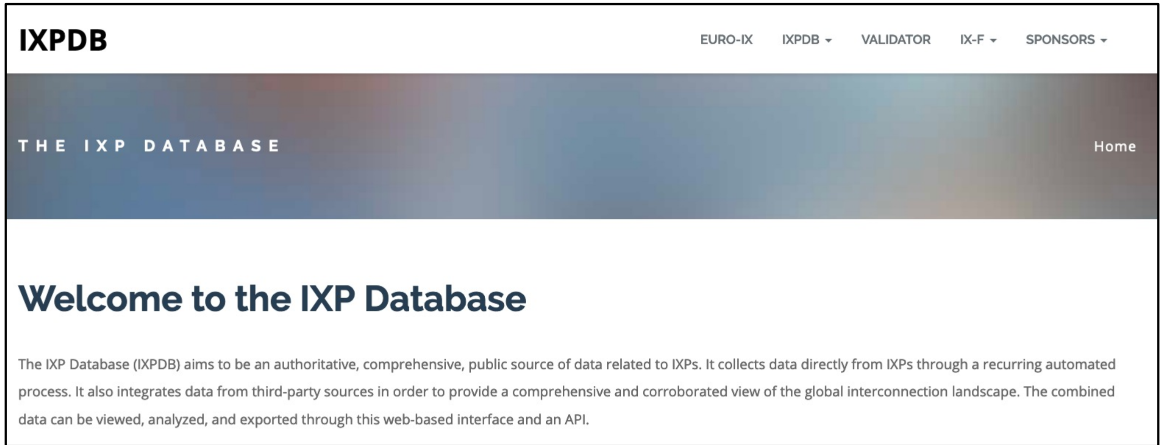
- Open Peering
  - Where a network operator publicly states that they will peer with all parties who approach them for peering
  - Commonly found at IXPs where the network operator participates via the Route Server (RS)
- Selective Peering
  - Where a network operator's peering policy depends on the nature of the operator who requests peering with them
  - At IXPs, the operator will not peer with RS but will only peer bilaterally
- Restrictive Peering
  - Where a network operator decides who its peering partners are, and is generally not approachable to considering peering opportunities

Types of Peering (4)

- The Peering Database documents network operator peering policies
  - https://www.peeringdb.com
- All operators with an AS are recommended to register in the PeeringDB
  - All operators who are considering peering or are peering must be in the PeeringDB to enhance their peering opportunities
- Participation in peering fora is encouraged too
  - Global Peering Forum (GPF) – (for North American peering)
  - Regional Peering Fora (Europe, Middle East, Africa, Asia, Caribbean, Latin America)
  - Many countries now have their own Peering Fora

Types of Peering (5)

- The IXPDB documents IXPs and their participants around the world
  - https://ixpdb.euro-ix.net/en/
- All Internet Exchange Point operators should register their IXP in the database
  - IXPs using IXP Manager will have this happen as part of the IXP Manager set up
  - Provides the LAN IP addresses of each member to facilitate automation



PeeringDB

Search here for a network, IX, or facility.

Advanced Search | v2 Search (Beta)

pfsinoz

English (English)

HKIX Gold Sponsor

Peers	Connections	Open Peers	Total Speed	% with IPv6
296	381	186	14.7T	82

Organization	Hong Kong Internet eXchange Limited
Also Known As	
Long Name	Hong Kong Internet Exchange
City	Hong Kong
Country	HK
Continental Region	Asia Pacific
Media Type	Ethernet
Service Level	Not Disclosed
Terms	Not Disclosed
Last Updated	2020-01-22T04:24:06Z
Notes ?	

Contact Information

Company Website	https://www.hkix.net/
Traffic Stats Website	https://www.hkix.net/hkix/stat/aggt/hkix-aggregate.html
Technical Email	noc@hkix.net
Technical Phone ?	+85239439900
Policy Email	info@hkix.net
Policy Phone ?	+85239438800
Sales Email	
Sales Phone ?	
Health Check	

Peers at this Exchange Point

Filter

Peer Name / IPv4	ASN / IPv6	Speed	Policy ?
2012 Limited 123.255.90.135	4658 2001:7fa:0:1::ca28:a087	10G	Selective
2012 Limited 123.255.90.122	4658 2001:7fa:0:1::ca28:a07a	1G	Selective
ACE CDN 123.255.91.67	139341 2001:7fa:0:1::ca28:a143	100G	Open
ACE CDN 123.255.91.79	139341 2001:7fa:0:1::ca28:a14f	100G	Open
ACME Universal 123.255.91.24	56190	1G	Open
ADVANCED HOSTERS 123.255.91.178	39572 2001:7fa:0:1::ca28:a1b2	100G	Selective
Advanced Information Co. 123.255.91.191	38047 2001:7fa:0:1::ca28:a1bf	10G	Open
Advanced Wireless Network Co. Ltd.(IIG) 123.255.92.80	45430 2001:7fa:0:1::ca28:a250	100G	Selective
AgotoZ HK 123.255.90.175	141167 2001:7fa:0:1::ca28:a0af	10G	Open
Akamai Prolexic DDoS Mitigation 123.255.91.26	32787 2001:7fa:0:1::ca28:a11a		Selective
Akamai Technologies 123.255.91.95	20940 2001:7fa:0:1::ca28:a15f	300G	Open
Akamai Technologies	20940	400G	Open

PeeringDB

Search here for a network, IX, or facility.

Advanced Search   v2 Search (Beta)

pfsinoz

English (English)

Amazon.com Diamond Sponsor

Organization	Amazon.com, Inc.
Also Known As	Amazon Web Services
Long Name	
Company Website	http://www.amazon.com
ASN	16509
IRR as-set/route-set	AS16509:AS-AMAZON
Route Server URL	
Looking Glass URL	
Network Type	Enterprise
IPv4 Prefixes	12000
IPv6 Prefixes	5000
Traffic Levels	Not Disclosed
Traffic Ratios	Balanced
Geographic Scope	Global
Protocols Supported	☑ Unicast IPv4 ○ Multicast ☑ IPv6 ☑ Never via route servers
Last Updated	2023-09-01T08:36:56Z
Public Peering Info Updated	2023-08-25T01:57:49
Peering Facility Info Updated	2023-08-29T15:43:36
Contact Info Updated	2020-12-01T12:29:55Z
Notes	**AWS Peering - https://peering.aws/** **Peering requests:** When submitting a peering request, please address the specific regional contact listed below only for the location of your request. *(Example: peering requests for London should use peering-*

Public Peering Exchange Points

Filter

Exchange / IPv4	ASN / IPv6	Speed	RS Peer
1-IX EU 185.1.254.91	16509 2001:7f8:115:1::91	100G	○
AKL-IX (Auckland NZ) 43.243.21.112	16509 2001:7fa:11:6:0:407d:0:1	100G	○
AKL-IX (Auckland NZ) 43.243.21.113	16509 2001:7fa:11:6:0:407d:0:2	100G	○
AMS-IX 80.249.210.100	16509 2001:7f8:1::a501:6509:1	600G	○
AMS-IX 80.249.210.217	16509 2001:7f8:1::a501:6509:2	600G	○
AMS-IX Chicago 206.108.115.36	16509 2001:504:38:1:0:a501:6509:1	100G	○
AMS-IX Hong Kong 103.247.139.74	16509 2001:df0:296::a501:6509:2	10G	○
AMS-IX Hong Kong 103.247.139.10	16509 2001:df0:296::a501:6509:1	10G	○
AMS-IX Mumbai 223.31.200.29	16509 2001:e48:44:100b:0:a501:6509:2	10G	○
AMS-IX Mumbai 223.31.200.30	16509 2001:e48:44:100b:0:a501:6509:1	10G	○
AMS-IX Singapore 112.137.24.238	16509 2a00:8422:ae5::a501:6509:1	10G	○
AMS-IX Singapore	16509	10G	○

Interconnection Facilities

Filter

PeeringDB

Search here for a network, IX, or facility.

Advanced Search

v2 Search (Beta)

pfsinoz

English (English)

Arelion (Twelve99)

Organization	Arelion
Also Known As	f/k/a Telia Carrier
Long Name	
Company Website	https://www.arelion.com/
ASN	1299
IRR as-set/route-set ?	RIPE::AS1299:AS-TWELVE99
Route Server URL	
Looking Glass URL	https://lg.twelve99.net/
Network Type	NSP
IPv4 Prefixes ?	600000
IPv6 Prefixes ?	130000
Traffic Levels	100+Tbps
Traffic Ratios	Balanced
Geographic Scope	Global
Protocols Supported	Unicast IPv4, Multicast, IPv6, Never via route servers
Last Updated	2023-08-25T12:04:42Z
Public Peering Info Updated	
Peering Facility Info Updated	2023-08-23T15:46:01
Contact Info Updated	2023-06-20T13:36:16
Notes ?	AS1299 is matching RPKI validation state and reject invalid prefixes from peers and customers. Our looking-glass marks validation state for all prefixes. Please review your registered ROAs to reduce number of invalid prefixes. **All trouble ticket requests or support related emails should be sent to support@arelion.com.**

Public Peering Exchange Points

Filter

Exchange / IPv4	ASN / IPv6	Speed	RS Peer

No filter matches.
You may filter by **Exchange**, **ASN** or **Speed.**

Interconnection Facilities

Filter

Facility / ASN	Country / City
123.NET - DC1 - 24700 Northwestern Hwy. 1299	United States of America Southfield
1530 SWIFT - NOCIX 1299	United States of America North Kansas City
1623 Farnam 1299	United States of America Omaha
365 Data Centers Buffalo (BU1) 1299	United States of America Buffalo
365 Data Centers Detroit (DT1) 1299	United States of America Southfield
365 Data Centers Nashville (NA1) 1299	United States of America Nashville
365 Data Centers Tampa (TA1) 1299	United States of America Tampa
3U Rechenzentrum Berlin 1299	Germany Berlin
910Telecom Denver 1299	United States of America Denver
Aitelecom Mérida 1299	Mexico Mérida

Internet Routing Registry

- Many major transit providers and several content providers pay attention to what is contained in the Internet Routing Registry
  - There are many IRRs operating, the most commonly used being those hosted by the Regional Internet Registries, RADB, and some transit providers
- Best practice for any AS holder is to document their routing policy in the IRR
  - A route-object is the absolute minimum requirement

Internet Routing Registry

- IRR objects can be created via the database web-interfaces or submitted via email
- Policy language used to be known as RPSL
- Problems:
  - IRR contains a lot of outdated information
  - Network operators not following best practices
- Some network operators now using RPKI and ROAs to securely indicate the origin AS of their routes
  - Takes priority over IRR entries
  - RPKI and ROAs covered in other presentations

Which Internet Routing Registry database to use?

- Members of a Regional Internet Registry are strongly encouraged to use their RIR's Internet Routing Registry instance
  - Usually managed via the RIR's member portal giving easy access for creation and update of objects
  - Provided as part of the RIR's services to its members
- Operators who do not belong to any RIR generally use:
  - Their upstream transit provider's Routing Registry (if provided)
  - The RADB (https://www.radb.net)
    - Placing objects in the RADB requires an annual subscription fee
    - RADB now uses IRRDv4 – objects with RPKI **Invalid** cannot be created; existing RPKI **Invalid** objects will NOT be visible in a query, nor can they be modified

Route Object: Purpose

- Documents which Autonomous System number is originating the route listed
- Required by many major transit providers
  - They build their customer and peer filter based on the route-objects listed in the IRR
  - Referring to at least the 5 RIR routing registries and the RADB
  - Some operators run their own Routing Registry
    - May require their customers to place a Route Object there (if not using the 5 RIR or RADB versions of the IRR)

Route Object: Examples

```
route:          202.144.128.0/20
descr:          DRUKNET-BLOCK-A1
country:        BT
notify:         ioc@bt.bt
mnt-by:         MAINT-BT-DRUKNET
origin:         AS18024
last-modified:  2018-09-18T09:37:40Z
source:         APNIC
```

This declares that AS18024 is the origin of 202.144.128.0/20

```
route6:         2405:D000::/32
descr:          DRUKNET-IPV6-BLOCK
origin:         AS17660
notify:         netops@bt.bt
mnt-by:         MAINT-BT-DRUKNET
last-modified:  2010-07-21T03:46:02Z
source:         APNIC
```

This declares that AS17660 is the origin of 2405:D000::/32

AS Object: Purpose

- Documents peering policy with other Autonomous Systems
  - Lists network information
  - Lists contact information
  - Lists routes announced to neighbouring autonomous systems
  - Lists routes accepted from neighbouring autonomous systems
- Some operators pay close attention to what is contained in the AS Object
  - Some configure their border router BGP policy based on what is listed in the AS Object

AS Object: Example

```
aut-num:        AS17660
as-name:        DRUKNET-AS
descr:          DrukNet ISP, Bhutan Telecom, Thimphu
country:        BT
org:            ORG-BTL2-AP
import:         from AS6461   action pref=100;     accept ANY
export:         to AS6461     announce AS-DRUKNET-TRANSIT
import:         from AS2914   action pref=150;     accept ANY
export:         to AS2914     announce AS-DRUKNET-TRANSIT
<snip>
import:         from AS135666 action pref=250;   accept AS135666
export:         to AS135666   announce {0.0.0.0/0} AS-DRUKNET-TRANSIT
admin-c:        DNO1-AP
tech-c:         DNO1-AP
notify:         netops@bt.bt
mnt-irt:        IRT-BTTELECOM-BT
mnt-by:         APNIC-HM
mnt-lower:      MAINT-BT-DRUKNET
mnt-routes:     MAINT-BT-DRUKNET
last-modified:  2019-06-09T22:40:10Z
source:         APNIC
```

Examples of inbound and outbound policies – RPSL

AS-Set: Purpose

- The AS-Set is used by network operators to group AS numbers they provide transit for in an easier to manage form
  - Convenient for more complicated policy declarations
  - Used mostly by network operators who build their EBGP filters from their IRR entries
  - Commonly used at Internet Exchange Points to handle large numbers of peers

AS-Set: Example

```
as-set:         AS-DRUKNET-TRANSIT
descr:          DrukNet transit networks
members:        AS17660
members:        AS132232
members:        AS134715
members:        AS135666
members:        AS137925
members:        AS59219
members:        AS18024
members:        AS18025
members:        AS137994
members:        AS140695
members:        AS151498
members:        AS151955
members:        AS152317
members:        AS138558
admin-c:        DNO1-AP
tech-c:         DNO1-AP
notify:         netops@bt.bt
mnt-by:         MAINT-BT-DRUKNET
last-modified:  2024-09-16T04:35:58Z
source:         APNIC
```

Lists all the autonomous systems within the AS-DRUKNET-TRANSIT group

Hierarchical AS-Set

- The usage of hierarchical AS-Set (RFC2622) is strongly recommended now (and required for APNIC IRR) – this helps resolve name collisions

```
as-set:         AS-GEMNET
descr:          GEMNET LLC
country:        MN
members:        AS9934, AS9484, AS10219, AS9789,
                AS38038, AS24496, AS24559, AS4850,
<snip>
tech-c:         GA263-AP
admin-c:        GA263-AP
mnt-by:         MAINT-GEMNET-MN
mnt-lower:      MAINT-GEMNET-MN
last-modified:  2023-09-26T01:25:15Z
source:         APNIC
```

VS

```
as-set:         AS-GEMNET
descr:          GEMNET s.r.o. ASes
members:        AS59479
members:        AS202733
tech-c:         DUMY-RIPE
admin-c:        DUMY-RIPE
mnt-by:         GEMNETCZ-MNT
created:        2013-08-19T09:49:13Z
last-modified:  2024-08-27T14:09:27Z
source:         RIPE
```

- Solution: AS-Set name changes to AS45204:AS-GEMNET
- Consult https://sanog.org/resources/sanog41/SANOG41_Conference-Recent-IRR-changes_Maz.pdf for more information and migration steps

Summary

- PeeringDB
  - An industry Best Practice so that:
    - Network operators can promote the interconnects they participate in and attract more peering partners
- IXPDB
  - An industry Best Practice so that:
    - Internet Exchange Points can show their participants and help make the interconnect more attractive for potential participants
- IRR
  - An industry Best Practice:
    - So that network operators can document which autonomous system is originating their prefixes
    - Used by network operators to filter prefixes received from their customers and peers

Configuration Tips

Of passwords, tricks and templates

IBGP and IGPs Reminder!

- Make sure loopback is configured on router
  - IBGP between loopbacks, NOT physical interfaces
- Make sure IGP carries loopback IPv4 /32 and IPv6 /128 address
- Consider the DMZ nets:
  - Use unnumbered interfaces?
  - Use next-hop-self on IBGP neighbours
  - Or carry the DMZ IPv4 /30s and IPv6 /127s in the IBGP
  - Basically, keep the DMZ nets out of the IGP!

IBGP: Next-hop-self

- BGP speaker announces external network to IBGP peers using router's local address (loopback) as next-hop
- Used by many service providers on edge routers
  - Preferable to carrying DMZ point-to-point link addresses in the IGP
  - Reduces size of IGP to just core infrastructure
  - Alternative to using unnumbered interfaces
  - Helps scale network
  - Many network operators consider this "best practice"

Limiting AS Path Length

- Some BGP implementations have problems with long AS_PATHS
  - Memory corruption
  - Memory fragmentation
- Even using AS_PATH prepends, it is not normal to see more than 20 ASNs in a typical AS_PATH in the Internet Routing Table today
  - The Internet is around 5 ASes deep on average
  - Largest AS_PATH is usually 16-20 ASNs

Limiting AS Path Length

- Some announcements have ridiculous lengths of AS-paths
  - This example is an error in one IPv6 implementation

```
*> 3FFE:1600::/24       22 11537 145 12199 10318 10566 13193 1930 2200 3425 293 5609 5430
13285 6939 14277 1849 33 15589 25336 6830 8002 2042 7610 i
```

  - This example shows 100 prepends (for no obvious reason)

```
*>i193.105.15.0         2516 3257 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
50404 i
```

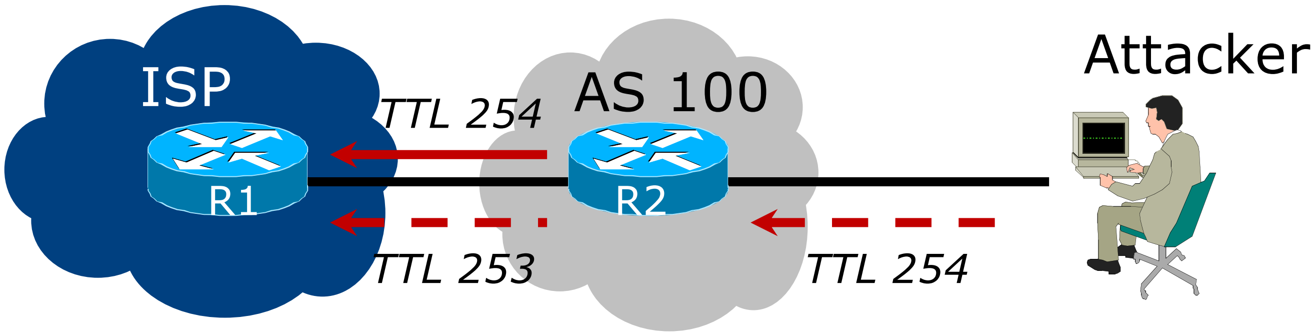
- If your implementation supports it, consider limiting the maximum AS-path length you will accept

BGP Maximum Prefix Tracking

- Allow configuration of the maximum number of prefixes a BGP router will receive from a peer
  - Supported by good BGP implementations
- Usually have two level control for prefix count:
  - Reaches warning threshold: log a warning message
    - Threshold is configurable
  - Reaches maximum:
    - Only send warnings
    - Tear down BGP, manual intervention required to restart
    - Tear down BGP and automatically restart after a delay (configurable)

BGP TTL "hack"

- Implement RFC5082 on BGP peerings
  - (Generalised TTL Security Mechanism)
  - Neighbour sets TTL to 255
  - Local router expects TTL of incoming BGP packets to be 254
  - No one apart from directly attached devices can send BGP packets which arrive with TTL of 254, so any possible attack by a remote miscreant is dropped due to TTL mismatch



BGP TTL "hack"

- TTL Hack:
  - Both neighbours must agree to use the feature
  - TTL check is much easier to perform than MD5
  - (Called BTSH – BGP TTL Security Hack)
- Provides "security" for BGP sessions
  - In addition to packet filters of course
  - MD5 should still be used for messages which slip through the TTL hack
  - See https://www.nanog.org/meetings/nanog27/presentations/meyer.pdf for more details

Templates

- Good practice to configure templates for everything
  - Vendor defaults tend not to be optimal or even very useful for network operators
  - Network operators create their own defaults by using configuration templates
- EBGP and IBGP examples follow
  - Also see Team Cymru's BGP templates
    - http://www.team-cymru.com/community-services

IBGP Template Example

- IBGP between loopbacks!
- Next-hop-self
  - Keep DMZ and external point-to-point out of IGP
- Always send communities in IBGP
  - Otherwise BGP policy accidents will happen
  - (Default on some vendor implementations, optional on others)
- Hardwire BGP to version 4
  - Yes, this is being paranoid!
  - Prevents accidental configuration of BGP version 3 which is still supported in some implementations

IBGP Template
Example continued

- Use passwords on IBGP session
  - Not being paranoid, VERY necessary
  - It's a secret shared between you and your peer
  - If arriving packets don't have the correct MD5 hash, they are ignored
  - Helps defeat miscreants who wish to attack BGP sessions
- Powerful preventative tool, especially when combined with filters and the TTL "hack"

EBGP Template Example

- BGP damping
  - Do **NOT** use it unless you understand the impact
  - Do **NOT** use the vendor defaults without thinking
- Cisco’s Soft Reconfiguration
  - Do **NOT** use unless troubleshooting or doing Route Origin Validation – it will consume considerable amounts of extra memory for BGP
- Remove private ASNs from announcements
  - Common omission today
- Use extensive filters, with “backup”
  - Use AS-path filters to backup prefix filters
  - Keep policy language for implementing policy, rather than basic filtering

EBGP Template
Example continued

- Use password agreed between you and peer on EBGP session
- Use maximum-prefix tracking
  - Router will warn you if there are sudden increases in BGP table size, bringing down EBGP if desired
- Limit maximum as-path length inbound
- Log changes of neighbour state
  - …and monitor those logs!
- Make BGP admin distance higher than that of any IGP
  - Otherwise, prefixes heard from outside your network could override your IGP!!

Mutually Agreed Norms for Routing Security

Industry Best Practices to ensure Security of the Routing System

MANRS

Routing Security

- Implement the recommendations in https://www.manrs.org
  1. Prevent propagation of incorrect routing information
     - Filter BGP peers, in & out!
  2. Prevent traffic with spoofed source addresses
     - BCP38 – Unicast Reverse Path Forwarding
  3. Facilitate communication between network operators
     - NOC to NOC Communication
     - Up-to-date details in Route and AS Objects, and PeeringDB
  4. Facilitate validation of routing information
     - Route Origin Authorisation using RPKI

MANRS

MANRS 1)

- Filtering prefixes inbound and outbound
  - RFC8212 requires all EBGP implementations to reject prefixes received and announced in the absence of any policy

- Advice: ***Never*** set up an EBGP session without inbound and outbound prefix filters
  - If full table required, block at least the bogons (see earlier)

MANRS 2)

- Implementing BCP 38
  - Unicast Reverse Path Forwarding
  - (Deny outbound traffic from customers which has spoofed source addresses)

- Advice: implement uRPF on ***all*** single-homed customer facing interfaces
  - Cheaper (CPU & RAM) than implementing packet filters

MANRS 3)

- Facilitate NOC to NOC communication
  - Know the **direct** NOC contacts for your customer Network Operators, your peer Network Operators, and your upstream Network Operators
  - This is not calling their "customer support line"
  - Make sure NOC contact info is part of any service contract
  - Up to date info in Route and AS Objects
  - Up to date AS info in PeeringDB

- Advice: NOC contact info for all connected Autonomous Networks is known to your NOC

MANRS 4)

- Facilitate validation of Routing Information
  - RPKI and Route Origin Authorisation (ROA)
  - All routes originated need to be signed to indicate that your AS is authorised to originate these routes
    - Helps secure the global routing system
- Advice: Sign ROAs for all originated routes using RPKI
  - And make sure all customer originated routes are also signed
  - Validate received routes from all peers
    - High priority for validated routes
    - Discard invalid routes
    - Low priority for unsigned routes

MANRS summary

- If your organisation supports and implements all 4 techniques in your network
  - Then join MANRS
  - https://www.manrs.org/join/
    - MANRS for Operators
    - MANRS for IXPs
    - MANRS for CDN & Cloud Providers

Summary

- Use configuration templates
- Standardise the configuration
- Be aware of standard "tricks" to avoid compromise of the BGP session
- Anything to make your life easier, network less prone to errors, network more likely to scale
- Implement the four fundamentals of MANRS
- It's all about scaling – if your network won't scale, then it won't be successful

BGP Techniques for Network Operators

The End!