BGP Best Current Practices

ISP Workshops

Last updated 25th May 2018

Acknowledgements

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

Philip Smith

Configuring BGP

Where do we start?

Cisco IOS Good Practices

- ISPs should start off with the following BGP commands as a basic template:

```
router bgp 64511
 bgp deterministic-med
 distance bgp 200 200 200
 no synchronization
 no auto-summary
```

Replace with public ASN

Make ebgp and ibgp distance the same

- If supporting more than just IPv4 unicast neighbours

```
 no bgp default ipv4-unicast
```

  - Turns off IOS assumption that all neighbours will exchange IPv4 prefixes

EBGP Configuration Best Practices

- Industry standard is described in RFC8212
  - https://tools.ietf.org/html/rfc8212
  - External BGP (EBGP) Route Propagation Behaviour without Policies

- NB: BGP in Cisco IOS 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 Configuration Best Practices

- 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
```

What is BGP for??

What is an IGP not for?

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 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**

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 ISP 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
  - May 2018: 393000 /24s in IPv4 table of 702000 prefixes
- The same is happening for /48s with IPv6
  - May 2018: 23000 /48s in IPv6 table of 49000 prefixes

Configuring Aggregation – Cisco IOS

- ISP has 100.66.0.0/19 address block
- To put into BGP as an aggregate:

```
router bgp 64511
 address-family ipv4
  network 100.66.0.0 mask 255.255.224.0
ip route 100.66.0.0 255.255.224.0 null0
```

- The static route is a “pull up” route
  - More specific prefixes within this address block ensure connectivity to ISP’s customers
  - “Longest match” lookup

Aggregation

- Address block should be announced to the Internet as an aggregate
- Subprefixes of address block should NOT be announced to Internet unless for traffic engineering
  - See BGP Multihoming presentations
- Aggregate should be generated internally
  - Not on the network borders!

Announcing Aggregate – Cisco IOS

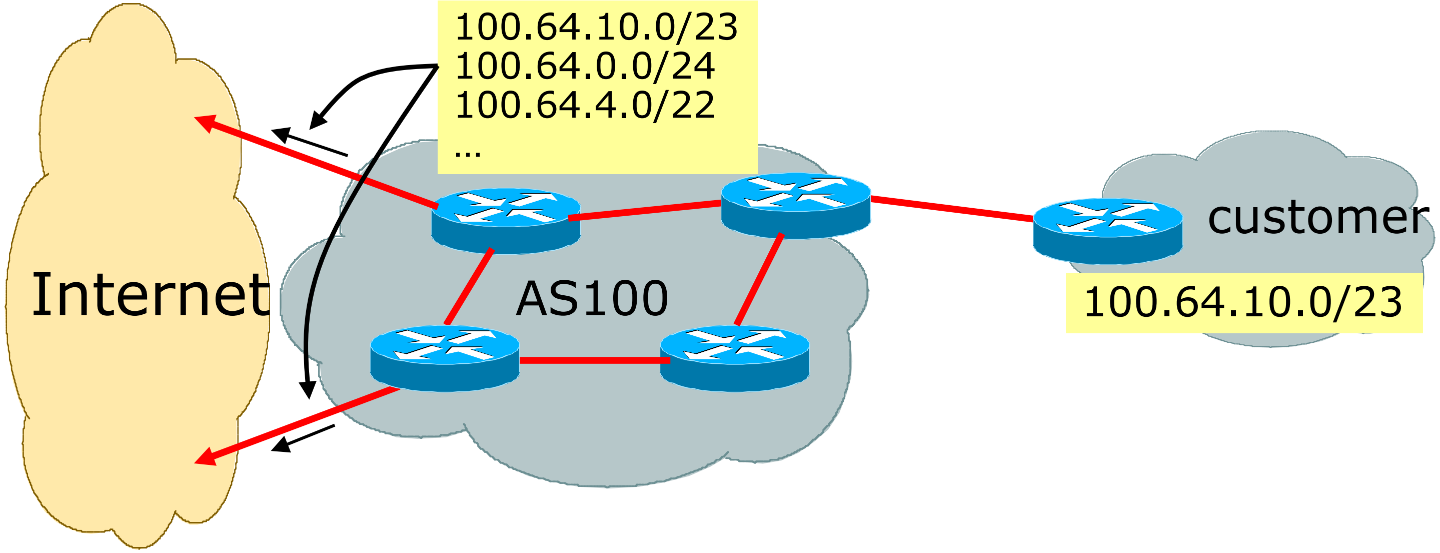
- Configuration Example

```
router bgp 64511
 address-family ipv4
  network 100.66.0.0 mask 255.255.224.0
  neighbor 100.67.10.1 remote-as 101
  neighbor 100.67.10.1 prefix-list out-filter out
  neighbor 100.67.10.1 prefix-list default in
  neighbor 100.67.10.1 activate
!
ip route 100.66.0.0 255.255.224.0 null0
!
ip prefix-list out-filter permit 100.66.0.0/19
ip prefix-list out-filter deny 0.0.0.0/0 le 32
!
ip prefix-list default permit 0.0.0.0/0
```

Announcing an Aggregate

- ISPs 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 in the Internet. But now?
  - IPv4 run-out is having an impact

Aggregation – Example

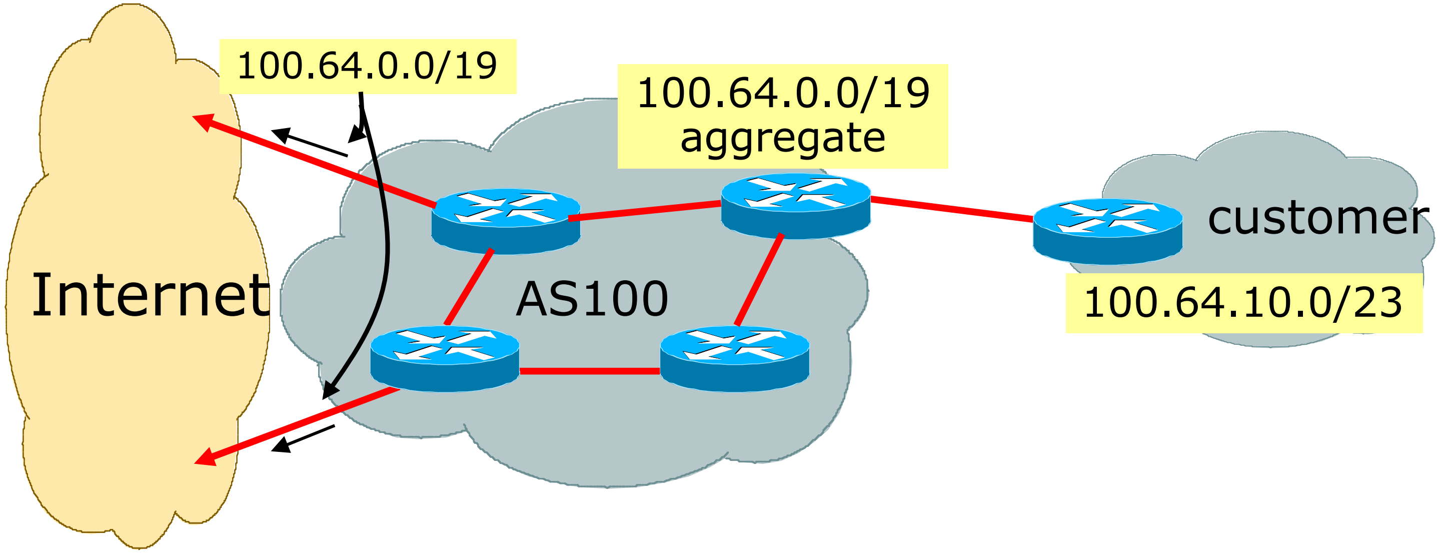


- Customer has /23 network assigned from AS100's /19 address block
- AS100 announces customers' individual networks to the Internet

Aggregation – Bad Example

- Customer link goes down
  - Their /23 network becomes unreachable
  - /23 is withdrawn from AS100's iBGP
- Their ISP 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 ISP
  - 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 ISP's 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 AS100's /19 address block
- AS100 announced /19 aggregate to the Internet

Aggregation – Good Example

- Customer link goes down
  - Their /23 network becomes unreachable
  - /23 is withdrawn from AS100's iBGP
- /19 aggregate is still being announced
  - No BGP hold down problems
  - No BGP propagation delays
  - No damping by other ISPs

- Customer link returns
- Their /23 network is visible again
  - The /23 is re-injected into AS100'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 ISPs 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 (May 2018)

- Current IPv4 Internet Routing Table Statistics

BGP Routing Table Entries	701924
Prefixes after maximum aggregation	269661
Unique prefixes in Internet	337019
/24s announced	392667
ASNs in use	60819

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

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

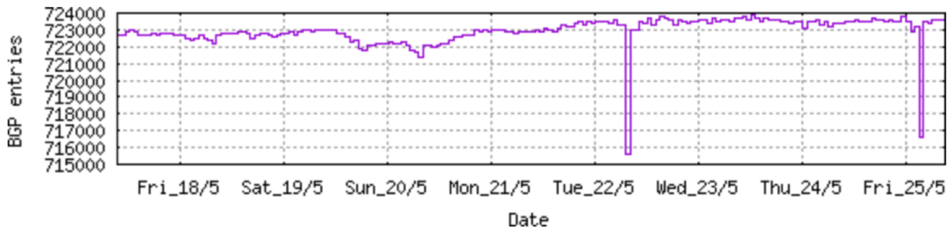
- Also computes the size of the routing table assuming ISPs performed optimal aggregation
- Website allows searches and computations of aggregation to be made on a per AS basis
  - Flexible and powerful tool to aid ISPs
  - Intended to show how greater efficiency in terms of BGP table size can be obtained 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

A list of advertisements of address blocks and Autonomous System numbers where there is no matching allocation data.

Status Summary

Table History

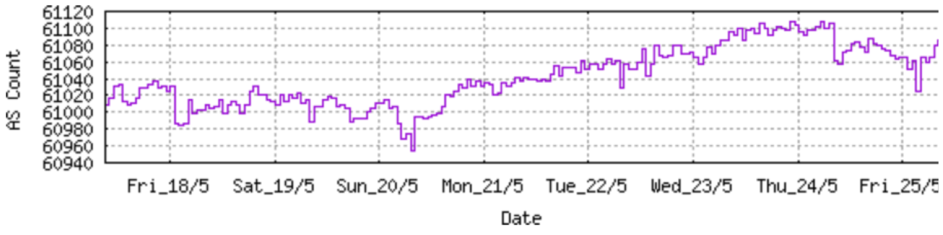
Date	Prefixes	CIDR Aggregated
18-05-18	722664	389624
19-05-18	722770	389323
20-05-18	722170	389080
21-05-18	722938	389747
22-05-18	723399	390485
23-05-18	723475	390045
24-05-18	723456	390481
25-05-18	723797	390473



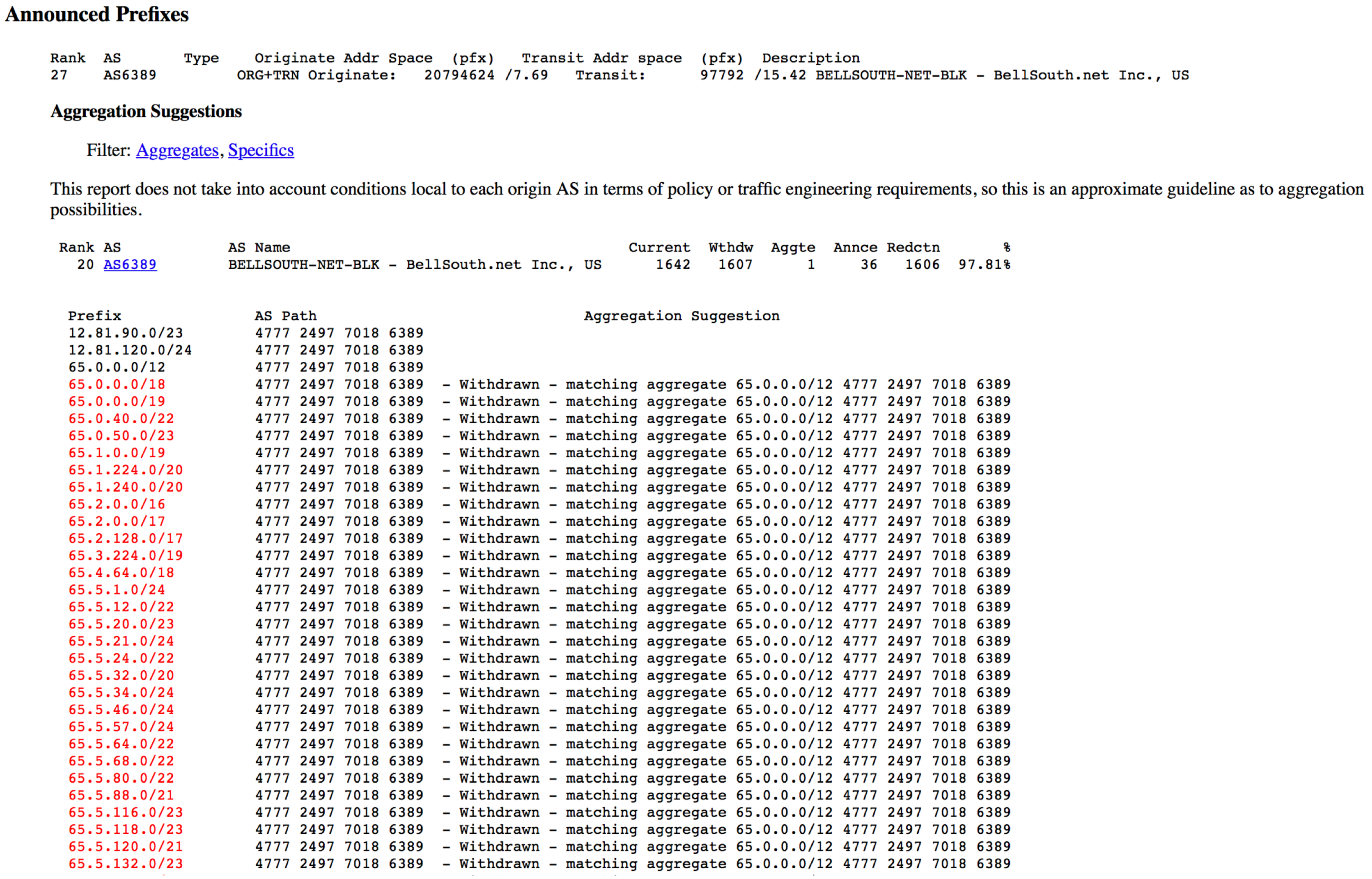
Plot: BGP Table Size

AS Summary

61089	Number of ASes in routing system
22713	Number of ASes announcing only one prefix
5420	Largest number of prefixes announced by an AS AS4538: ERX-CERNET-BKB China Education and Research Network Center, CN
119056384	Largest address span announced by an AS (/32s) AS4134: CHINANET-BACKBONE No.31,Jin-rong Street, CN



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
27    AS6389         ORG+TRN Originate:    20794624 /7.69   Transit:      97792 /15.42 BELLSOUTH-NET-BLK - BellSouth.net Inc., US
```

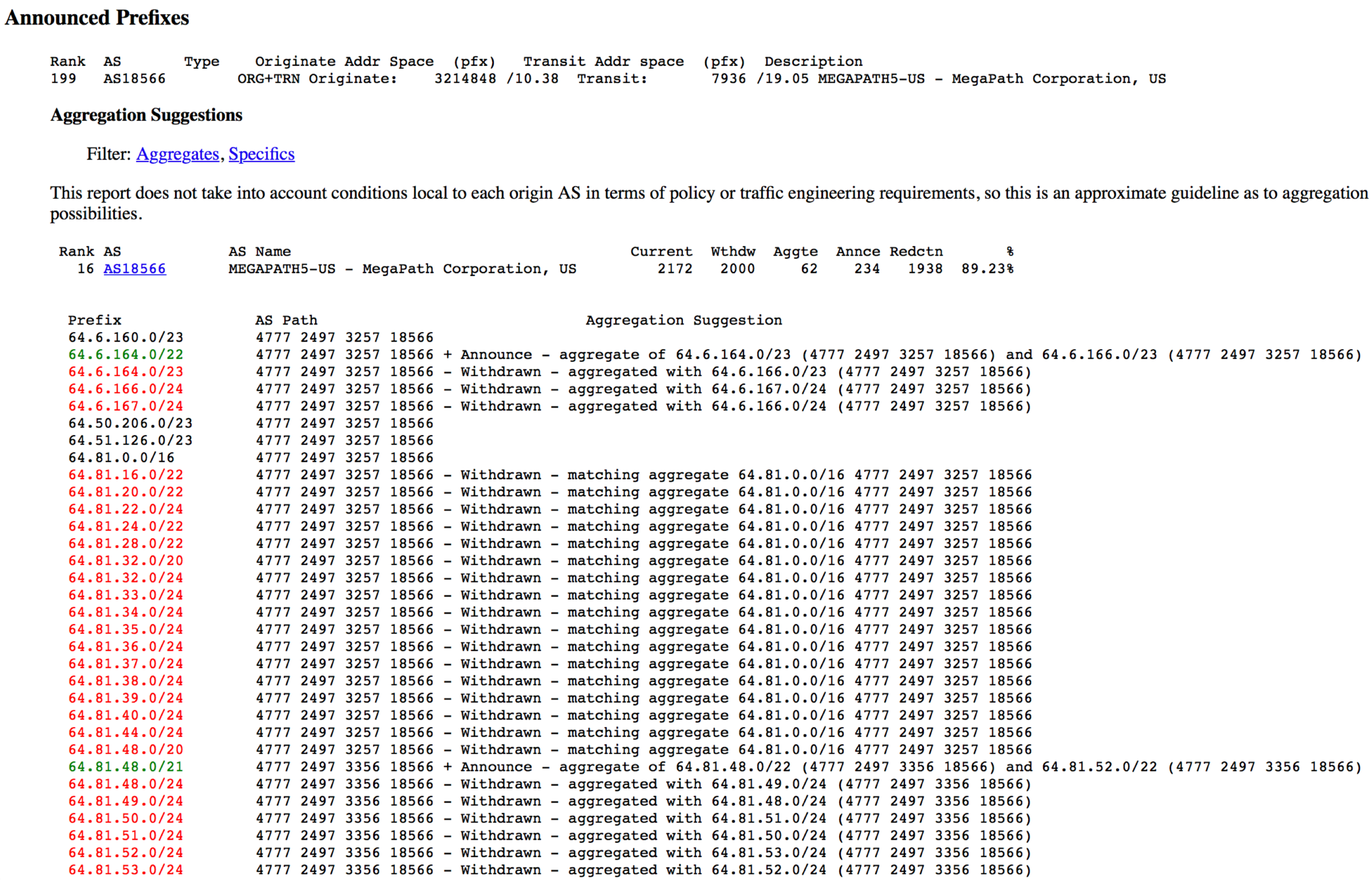
Aggregation Suggestions

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      %
   20 AS6389           BELLSOUTH-NET-BLK - BellSouth.net Inc., US    1642   1607      1     36   1606  97.81%


  Prefix               AS Path                              Aggregation Suggestion
  12.81.90.0/23        4777 2497 7018 6389
  12.81.120.0/24       4777 2497 7018 6389
  65.0.0.0/12          4777 2497 7018 6389
  65.0.0.0/18          4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.0.0.0/19          4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.0.40.0/22         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.0.50.0/23         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.1.0.0/19          4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.1.224.0/20        4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.1.240.0/20        4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.2.0.0/16          4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.2.0.0/17          4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.2.128.0/17        4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.3.224.0/19        4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.4.64.0/18         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.1.0/24          4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.12.0/22         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.20.0/23         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.21.0/24         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.24.0/22         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.32.0/20         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.34.0/24         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.46.0/24         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.57.0/24         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.64.0/22         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.68.0/22         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.80.0/22         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.88.0/21         4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.116.0/23        4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.118.0/23        4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.120.0/21        4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
  65.5.132.0/23        4777 2497 7018 6389   - Withdrawn - matching aggregate 65.0.0.0/12 4777 2497 7018 6389
```



Announced Prefixes

```
Rank  AS        Type    Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
199   AS18566        ORG+TRN Originate:    3214848 /10.38  Transit:       7936 /19.05 MEGAPATH5-US - MegaPath Corporation, US
```

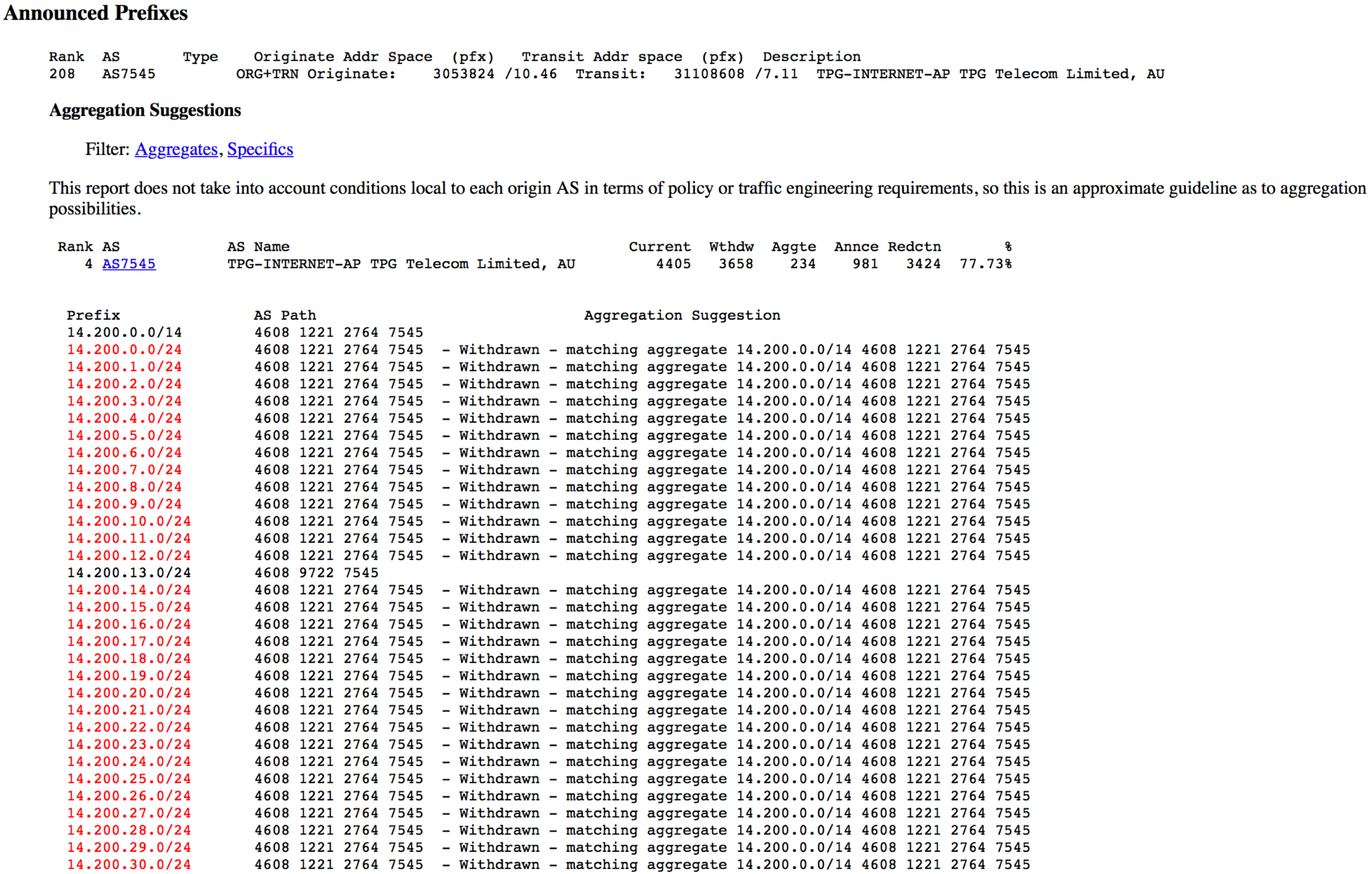
Aggregation Suggestions

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      %
   16 AS18566       MEGAPATH5-US - MegaPath Corporation, US    2172   2000     62    234   1938  89.23%
```

```
  Prefix              AS Path                              Aggregation Suggestion
  64.6.160.0/23       4777 2497 3257 18566
  64.6.164.0/22       4777 2497 3257 18566 + Announce - aggregate of 64.6.164.0/23 (4777 2497 3257 18566) and 64.6.166.0/23 (4777 2497 3257 18566)
  64.6.164.0/23       4777 2497 3257 18566 - Withdrawn - aggregated with 64.6.166.0/23 (4777 2497 3257 18566)
  64.6.166.0/24       4777 2497 3257 18566 - Withdrawn - aggregated with 64.6.167.0/24 (4777 2497 3257 18566)
  64.6.167.0/24       4777 2497 3257 18566 - Withdrawn - aggregated with 64.6.166.0/24 (4777 2497 3257 18566)
  64.50.206.0/23      4777 2497 3257 18566
  64.51.126.0/23      4777 2497 3257 18566
  64.81.0.0/16        4777 2497 3257 18566
  64.81.16.0/22       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.20.0/22       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.22.0/24       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.24.0/22       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.28.0/22       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.32.0/20       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.32.0/24       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.33.0/24       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.34.0/24       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.35.0/24       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.36.0/24       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.37.0/24       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.38.0/24       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.39.0/24       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.40.0/24       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.44.0/24       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.48.0/20       4777 2497 3257 18566 - Withdrawn - matching aggregate 64.81.0.0/16 4777 2497 3257 18566
  64.81.48.0/21       4777 2497 3356 18566 + Announce - aggregate of 64.81.48.0/22 (4777 2497 3356 18566) and 64.81.52.0/22 (4777 2497 3356 18566)
  64.81.48.0/24       4777 2497 3356 18566 - Withdrawn - aggregated with 64.81.49.0/24 (4777 2497 3356 18566)
  64.81.49.0/24       4777 2497 3356 18566 - Withdrawn - aggregated with 64.81.48.0/24 (4777 2497 3356 18566)
  64.81.50.0/24       4777 2497 3356 18566 - Withdrawn - aggregated with 64.81.51.0/24 (4777 2497 3356 18566)
  64.81.51.0/24       4777 2497 3356 18566 - Withdrawn - aggregated with 64.81.50.0/24 (4777 2497 3356 18566)
  64.81.52.0/24       4777 2497 3356 18566 - Withdrawn - aggregated with 64.81.53.0/24 (4777 2497 3356 18566)
  64.81.53.0/24       4777 2497 3356 18566 - Withdrawn - aggregated with 64.81.52.0/24 (4777 2497 3356 18566)
```



Announced Prefixes

```
Rank  AS        Type     Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
208   AS7545          ORG+TRN Originate:    3053824 /10.46  Transit:    31108608 /7.11  TPG-INTERNET-AP TPG Telecom Limited, AU
```

Aggregation Suggestions

Filter: Aggregates, Specifics

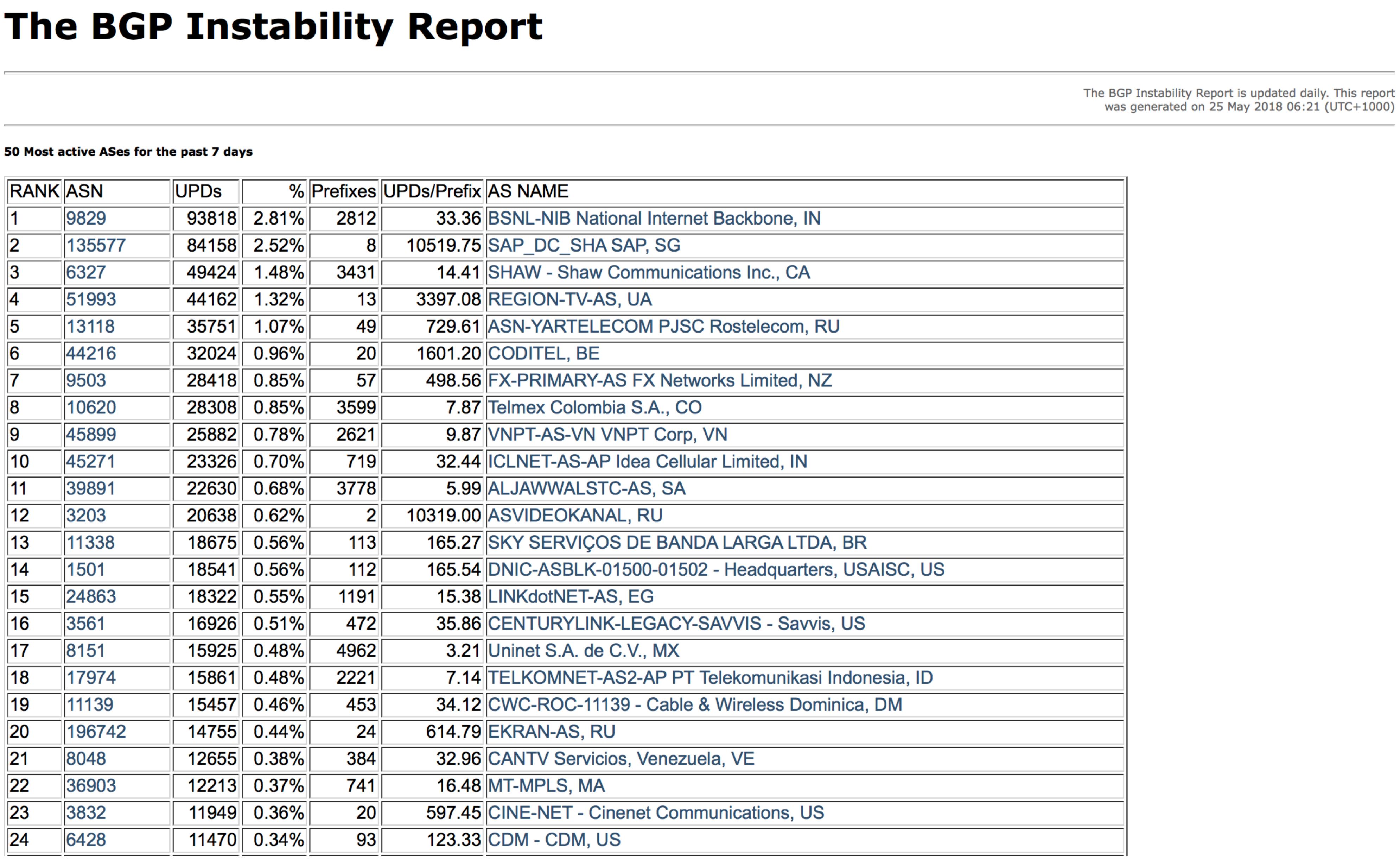
This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

```
 Rank AS            AS Name                                 Current  Wthdw  Aggte  Annce Redctn       %
    4 AS7545        TPG-INTERNET-AP TPG Telecom Limited, AU     4405   3658    234    981   3424  77.73%


  Prefix             AS Path                             Aggregation Suggestion
  14.200.0.0/14      4608 1221 2764 7545
  14.200.0.0/24      4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.1.0/24      4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.2.0/24      4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.3.0/24      4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.4.0/24      4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.5.0/24      4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.6.0/24      4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.7.0/24      4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.8.0/24      4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.9.0/24      4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.10.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.11.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.12.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.13.0/24     4608 9722 7545
  14.200.14.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.15.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.16.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.17.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.18.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.19.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.20.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.21.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.22.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.23.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.24.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.25.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.26.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.27.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.28.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.29.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
  14.200.30.0/24     4608 1221 2764 7545  - Withdrawn - matching aggregate 14.200.0.0/14 4608 1221 2764 7545
```

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

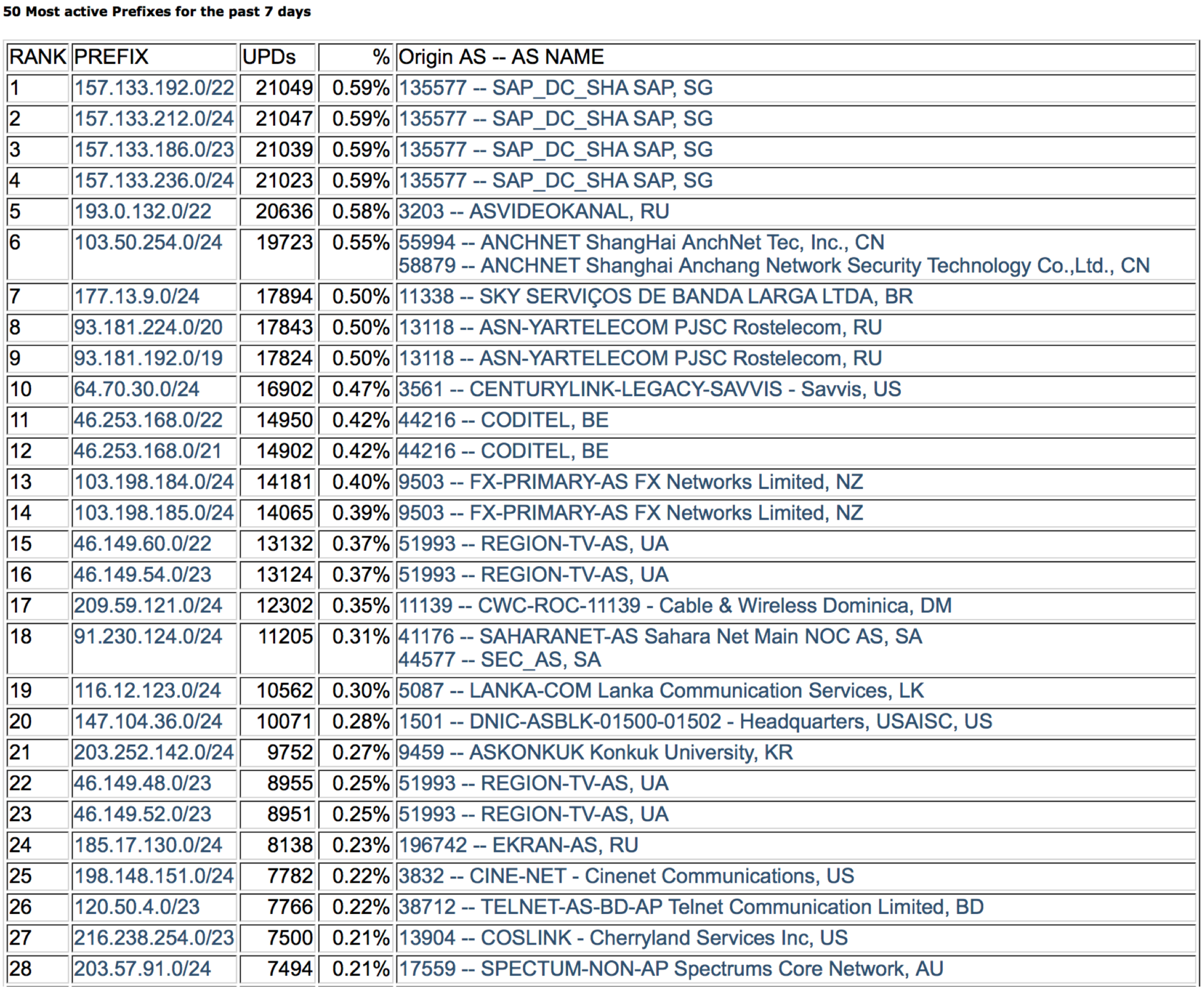


The BGP Instability Report

The BGP Instability Report is updated daily. This report
was generated on 25 May 2018 06:21 (UTC+1000)

50 Most active ASes for the past 7 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	9829	93818	2.81%	2812	33.36	BSNL-NIB National Internet Backbone, IN
2	135577	84158	2.52%	8	10519.75	SAP_DC_SHA SAP, SG
3	6327	49424	1.48%	3431	14.41	SHAW - Shaw Communications Inc., CA
4	51993	44162	1.32%	13	3397.08	REGION-TV-AS, UA
5	13118	35751	1.07%	49	729.61	ASN-YARTELECOM PJSC Rostelecom, RU
6	44216	32024	0.96%	20	1601.20	CODITEL, BE
7	9503	28418	0.85%	57	498.56	FX-PRIMARY-AS FX Networks Limited, NZ
8	10620	28308	0.85%	3599	7.87	Telmex Colombia S.A., CO
9	45899	25882	0.78%	2621	9.87	VNPT-AS-VN VNPT Corp, VN
10	45271	23326	0.70%	719	32.44	ICLNET-AS-AP Idea Cellular Limited, IN
11	39891	22630	0.68%	3778	5.99	ALJAWWALSTC-AS, SA
12	3203	20638	0.62%	2	10319.00	ASVIDEOKANAL, RU
13	11338	18675	0.56%	113	165.27	SKY SERVIÇOS DE BANDA LARGA LTDA, BR
14	1501	18541	0.56%	112	165.54	DNIC-ASBLK-01500-01502 - Headquarters, USAISC, US
15	24863	18322	0.55%	1191	15.38	LINKdotNET-AS, EG
16	3561	16926	0.51%	472	35.86	CENTURYLINK-LEGACY-SAVVIS - Savvis, US
17	8151	15925	0.48%	4962	3.21	Uninet S.A. de C.V., MX
18	17974	15861	0.48%	2221	7.14	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia, ID
19	11139	15457	0.46%	453	34.12	CWC-ROC-11139 - Cable & Wireless Dominica, DM
20	196742	14755	0.44%	24	614.79	EKRAN-AS, RU
21	8048	12655	0.38%	384	32.96	CANTV Servicios, Venezuela, VE
22	36903	12213	0.37%	741	16.48	MT-MPLS, MA
23	3832	11949	0.36%	20	597.45	CINE-NET - Cinenet Communications, US
24	6428	11470	0.34%	93	123.33	CDM - CDM, US



50 Most active Prefixes for the past 7 days

RANK	PREFIX	UPDs	%	Origin AS -- AS NAME
1	157.133.192.0/22	21049	0.59%	135577 -- SAP_DC_SHA SAP, SG
2	157.133.212.0/24	21047	0.59%	135577 -- SAP_DC_SHA SAP, SG
3	157.133.186.0/23	21039	0.59%	135577 -- SAP_DC_SHA SAP, SG
4	157.133.236.0/24	21023	0.59%	135577 -- SAP_DC_SHA SAP, SG
5	193.0.132.0/22	20636	0.58%	3203 -- ASVIDEOKANAL, RU
6	103.50.254.0/24	19723	0.55%	55994 -- ANCHNET ShangHai AnchNet Tec, Inc., CN 58879 -- ANCHNET Shanghai Anchang Network Security Technology Co.,Ltd., CN
7	177.13.9.0/24	17894	0.50%	11338 -- SKY SERVIÇOS DE BANDA LARGA LTDA, BR
8	93.181.224.0/20	17843	0.50%	13118 -- ASN-YARTELECOM PJSC Rostelecom, RU
9	93.181.192.0/19	17824	0.50%	13118 -- ASN-YARTELECOM PJSC Rostelecom, RU
10	64.70.30.0/24	16902	0.47%	3561 -- CENTURYLINK-LEGACY-SAVVIS - Savvis, US
11	46.253.168.0/22	14950	0.42%	44216 -- CODITEL, BE
12	46.253.168.0/21	14902	0.42%	44216 -- CODITEL, BE
13	103.198.184.0/24	14181	0.40%	9503 -- FX-PRIMARY-AS FX Networks Limited, NZ
14	103.198.185.0/24	14065	0.39%	9503 -- FX-PRIMARY-AS FX Networks Limited, NZ
15	46.149.60.0/22	13132	0.37%	51993 -- REGION-TV-AS, UA
16	46.149.54.0/23	13124	0.37%	51993 -- REGION-TV-AS, UA
17	209.59.121.0/24	12302	0.35%	11139 -- CWC-ROC-11139 - Cable & Wireless Dominica, DM
18	91.230.124.0/24	11205	0.31%	41176 -- SAHARANET-AS Sahara Net Main NOC AS, SA 44577 -- SEC_AS, SA
19	116.12.123.0/24	10562	0.30%	5087 -- LANKA-COM Lanka Communication Services, LK
20	147.104.36.0/24	10071	0.28%	1501 -- DNIC-ASBLK-01500-01502 - Headquarters, USAISC, US
21	203.252.142.0/24	9752	0.27%	9459 -- ASKONKUK Konkuk University, KR
22	46.149.48.0/23	8955	0.25%	51993 -- REGION-TV-AS, UA
23	46.149.52.0/23	8951	0.25%	51993 -- REGION-TV-AS, UA
24	185.17.130.0/24	8138	0.23%	196742 -- EKRAN-AS, RU
25	198.148.151.0/24	7782	0.22%	3832 -- CINE-NET - Cinenet Communications, US
26	120.50.4.0/23	7766	0.22%	38712 -- TELNET-AS-BD-AP Telnet Communication Limited, BD
27	216.238.254.0/23	7500	0.21%	13904 -- COSLINK - Cherryland Services Inc, US
28	203.57.91.0/24	7494	0.21%	17559 -- SPECTUM-NON-AP Spectrums Core Network, AU

Receiving Prefixes

Receiving Prefixes

- There are three scenarios for receiving prefixes from other ASNs
  - 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

- ISPs should only accept prefixes which have been assigned or allocated to their downstream customer
- If ISP has assigned address space to its customer, then the customer IS entitled to announce it back to his ISP
- If the ISP 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 all 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
```

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 193.128.0.0/16

inetnum:        193.128.0.0 - 193.133.255.255
netname:        UK-PIPEX-193-128-133
country:        GB
org:            ORG-UA24-RIPE
admin-c:        WERT1-RIPE
tech-c:         UPHM1-RIPE
status:         ALLOCATED UNSPECIFIED
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:        2002-06-25T15:05:40Z
last-modified:  2016-10-31T12:20:01Z
source:         RIPE
```

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

Receiving Prefixes from customer: Cisco IOS

- For Example:
  - Downstream has 100.69.0.0/20 block
  - Should only announce this to upstreams
  - Upstreams should only accept this from them
- Configuration on upstream

```
router bgp 100
 address-family ipv4
  neighbor 100.67.10.1 remote-as 101
  neighbor 100.67.10.1 prefix-list customer in
  neighbor 100.67.10.1 prefix-list default out
  neighbor 100.67.10.1 activate
!
ip prefix-list customer permit 100.69.0.0/20
!
ip prefix-list default permit 0.0.0.0/0
```

Receiving Prefixes: From Peers

- A peer is an ISP 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 the IRRToolSet

https://github.com/irrtoolset/irrtoolset

Receiving Prefixes from peer: Cisco IOS

- For Example:
  - Peer has 220.50.0.0/16, 61.237.64.0/18 and 81.250.128.0/17 address blocks
- Configuration on local router

```
router bgp 100
 address-family ipv4
  neighbor 100.67.10.1 remote-as 101
  neighbor 100.67.10.1 prefix-list my-peer in
  neighbor 100.67.10.1 prefix-list my-prefix out
  neighbor 100.67.10.1 activate
!
ip prefix-list my-peer permit 220.50.0.0/16
ip prefix-list my-peer permit 61.237.64.0/18
ip prefix-list my-peer permit 81.250.128.0/17
ip prefix-list my-peer deny 0.0.0.0/0 le 32
!
ip prefix-list my-prefix permit 100.67.16.0/20
```

Receiving Prefixes:
From Upstream/Transit Provider

- Upstream/Transit Provider is an ISP who you pay to give you transit to the WHOLE Internet
- Receiving prefixes from them is not desirable unless really necessary
  - Traffic Engineering – see BGP Multihoming presentations
- 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

- Downstream Router Configuration

```
router bgp 100
 address-family ipv4
  network 100.66.0.0 mask 255.255.224.0
  neighbor 100.65.7.1 remote-as 101
  neighbor 100.65.7.1 prefix-list infilter in
  neighbor 100.65.7.1 prefix-list outfilter out
  neighbor 100.65.7.1 activate
!
ip prefix-list infilter permit 0.0.0.0/0
!
ip prefix-list outfilter permit 100.66.0.0/19
```

Receiving Prefixes:
From Upstream/Transit Provider

- Upstream Router Configuration

```
router bgp 101
 address-family ipv4
  neighbor 100.65.7.2 remote-as 100
  neighbor 100.65.7.2 default-originate
  neighbor 100.65.7.2 prefix-list cust-in in
  neighbor 100.65.7.2 prefix-list cust-out out
  neighbor 100.65.7.2 activate
!
ip prefix-list cust-in permit 100.66.0.0/19
!
ip prefix-list cust-out permit 0.0.0.0/0
```

Receiving Prefixes: From Upstream/Transit Provider

- If 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"
  - www.team-cymru.org/Services/Bogons/http.html
- For IPv4 also consult:
  - www.rfc-editor.org/rfc/rfc6441.txt (BCP171)
- For IPv6 also consult:
  - www.space.net/~gert/RIPE/ipv6-filters.html
- Bogon Route Server:
  - www.team-cymru.org/Services/Bogons/routeserver.html
  - Supplies a BGP feed (IPv4 and/or IPv6) of address blocks which should not appear in the BGP table

Receiving IPv4 Prefixes

```
router bgp 100
 network 101.10.0.0 mask 255.255.224.0
 neighbor 100.65.7.1 remote-as 101
 neighbor 100.65.7.1 prefix-list in-filter in
!
ip prefix-list in-filter deny 0.0.0.0/0             ! Default
ip prefix-list in-filter deny 0.0.0.0/8 le 32       ! RFC1122 local host
ip prefix-list in-filter deny 10.0.0.0/8 le 32      ! RFC1918
ip prefix-list in-filter deny 100.64.0.0/10 le 32   ! RFC6598 shared address
ip prefix-list in-filter deny 101.10.0.0/19 le 32   ! Local prefix
ip prefix-list in-filter deny 127.0.0.0/8 le 32     ! Loopback
ip prefix-list in-filter deny 169.254.0.0/16 le 32  ! Auto-config
ip prefix-list in-filter deny 172.16.0.0/12 le 32   ! RFC1918
ip prefix-list in-filter deny 192.0.0.0/24 le 32    ! RFC6598 IETF protocol
ip prefix-list in-filter deny 192.0.2.0/24 le 32    ! TEST1
ip prefix-list in-filter deny 192.168.0.0/16 le 32  ! RFC1918
ip prefix-list in-filter deny 198.18.0.0/15 le 32   ! Benchmarking
ip prefix-list in-filter deny 198.51.100.0/24 le 32 ! TEST2
ip prefix-list in-filter deny 203.0.113.0/24 le 32  ! TEST3
ip prefix-list in-filter deny 224.0.0.0/3 le 32     ! Multicast & Experimental
ip prefix-list in-filter deny 0.0.0.0/0 ge 25       ! Prefixes >/24
ip prefix-list in-filter permit 0.0.0.0/0 le 32
```

Receiving IPv6 Prefixes

```
router bgp 100
 network 2020:3030::/32
 neighbor 2020:3030::1 remote-as 101
 neighbor 2020:3030::1 prefix-list v6in-filter in
!
ipv6 prefix-list v6in-filter permit 64:ff9b::/96          ! RFC6052 v4v6trans
ipv6 prefix-list v6in-filter deny 2001::/23 le 128        ! RFC2928 IETF prot
ipv6 prefix-list v6in-filter deny 2001:2::/48 le 128      ! Benchmarking
ipv6 prefix-list v6in-filter deny 2001:10::/28 le 128     ! ORCHID
ipv6 prefix-list v6in-filter deny 2001:db8::/32 le 128    ! Documentation
ipv6 prefix-list v6in-filter deny 2002::/16 le 128        ! Deny all 6to4
ipv6 prefix-list v6in-filter deny 2020:3030::/32 le 128   ! Local Prefix
ipv6 prefix-list v6in-filter deny 3ffe::/16 le 128        ! Formerly 6bone
ipv6 prefix-list v6in-filter permit 2000::/3 le 48        ! Global Unicast
ipv6 prefix-list v6in-filter deny ::/0 le 128
```

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 ISPs to be good Internet citizens

Prefixes into iBGP

Injecting prefixes into iBGP

- Use iBGP to carry customer prefixes
  - Don’t use IGP
- Point static route to customer interface
- Use BGP network statement
- As long as static route exists (interface active), prefix will be in BGP

Router Configuration: network statement

- Example:

```
interface loopback 0
 ip address 100.64.3.1 255.255.255.255
!
interface Serial 5/0
 ip unnumbered loopback 0
 ip verify unicast reverse-path
!
ip route 100.71.10.0 255.255.252.0 Serial 5/0
!
router bgp 100
 address-family ipv4
  network 100.71.10.0 mask 255.255.252.0
!
```

Injecting prefixes into iBGP

- Interface flap will result in prefix withdraw and reannounce
  - use "**`ip route . . . permanent`**"
- Many ISPs redistribute static routes into BGP rather than using the network statement
  - Only do this if you understand why

Router Configuration: redistribute static

- Example:

```
ip route 100.71.10.0 255.255.252.0 Serial 5/0
!
router bgp 100
 address-family ipv4
  redistribute static route-map static-to-bgp
<snip>
!
route-map static-to-bgp permit 10
 match ip address prefix-list ISP-block
 set origin igp
 set community 100:1000
<snip>
!
ip prefix-list ISP-block permit 100.71.10.0/22 le 30
```

Injecting prefixes into iBGP

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

Summary

- Best Practices Covered:
  - When to use BGP
  - When to use ISIS/OSPF
  - Aggregation
  - Receiving Prefixes
  - Prefixes into BGP

Route Origin Authorisation

Steps to securing the Routing System

What is RPKI?

- Resource Public Key Infrastructure (RPKI)
- A robust security framework for verifying the association between resource holder and their Internet resources
- Created to address the issues in RFC 4593 “Generic Threats to Routing Protocols”
- Helps to secure Internet routing by validating routes
  - Proof that prefix announcements are coming from the legitimate holder of the resource

RFC 6480 – An Infrastructure to Support Secure Internet Routing (Feb 2012)

Benefits of RPKI - Routing

- Prevents route hijacking
  - A prefix originated by an AS without authorisation
  - Reason: malicious intent
- Prevents mis-origination
  - A prefix that is mistakenly originated by an AS which does not own it
  - Also route leakage
  - Reason: configuration mistake / fat finger

BGP Security (BGPsec)

- Extension to BGP that provides improved security for BGP routing
- Being worked on by the SIDR Working Group at IETF
- Implemented via a new optional non-transitive BGP attribute that contains a digital signature
- Two components:
  - BGP Prefix Origin Validation (using RPKI)
  - BGP Path Validation

Route Origin Authorisation (ROA)

- A digital object that contains a list of address prefixes and one AS number
- It is an authority created by a prefix holder to authorise an AS Number to originate one or more specific route advertisements
- Publish a ROA using your RIR member portal

Router Origin Validation

- Router must support RPKI
- Checks an RP cache / validator
- Validation returns 3 states:
  - Valid = when authorization is found for prefix X
  - Invalid = when authorization is found for prefix X but not from ASN Y
  - Unknown = when no authorization data is found
- Vendor support:
  - Cisco IOS – available in release 15.2
  - Cisco IOS/XR – available in release 4.3.2
  - Juniper – available in release 12.2
  - Nokia – available in release R12.0R4
  - Huawei – newly available – release TBA

Configure Router to Use Cache

- Point router to the local RPKI cache
  - Server listens on port 43779
  - Cisco IOS example:

```
router bgp 64512
 bgp rpki server tcp 10.0.0.3 port 43779 refresh 60
```

BGP Table (IPv4)

```
RPKI validation codes: V valid, I invalid, N Not found

Network          Metric LocPrf Path
N*>  1.0.4.0/24       0        37100 6939 4637 1221 38803 56203 i
N*>  1.0.5.0/24       0        37100 6939 4637 1221 38803 56203 i
...
V*>  1.9.0.0/16       0        37100 4788 i
N*>  1.10.8.0/24      0        37100 10026 18046 17408 58730 i
N*>  1.10.64.0/24     0        37100 6453 3491 133741 i
...
V*>  1.37.0.0/16      0        37100 4766 4775 i
N*>  1.38.0.0/23      0        37100 6453 1273 55410 38266 i
N*>  1.38.0.0/17      0        37100 6453 1273 55410 38266 {38266} i
...
I*   5.8.240.0/23     0        37100 44217 3178 i
I*   5.8.241.0/24     0        37100 44217 3178 i
I*   5.8.242.0/23     0        37100 44217 3178 i
I*   5.8.244.0/23     0        37100 44217 3178 i
...
```

Courtesy of SEACOM: http://as37100.net

BGP Table (IPv6)

```
RPKI validation codes: V valid, I invalid, N Not found

Network              Metric LocPrf Path
N*>  2001::/32            0        37100 6939 i
N*   2001:4:112::/48      0        37100 112 i
...
V*>  2001:240::/32        0         37100 2497 i
N*>  2001:250::/48        0         37100 6939 23911 45
N*>  2001:250::/32        0         37100 6939 23911 23910 i
...
V*>  2001:348::/32        0         37100 2497 7679 i
N*>  2001:350::/32        0         37100 2497 7671 i
N*>  2001:358::/32        0         37100 2497 4680 i
...
I*   2001:1218:101::/48   0         37100 6453 8151 278 i
I*   2001:1218:104::/48   0         37100 6453 8151 278 i
N*   2001:1221::/48       0         37100 2914 8151 28496 i
N*>  2001:1228::/32       0         37100 174 18592 i
...
```

Courtesy of SEACOM: http://as37100.net

RPKI BGP State: Valid

```
BGP routing table entry for 2001:240::/32, version 109576927
Paths: (2 available, best #2, table default)
  Not advertised to any peer
  Refresh Epoch 1
  37100 2497
    2C0F:FEB0:11:2::1 (FE80::2A8A:1C00:1560:5BC0) from
                                  2C0F:FEB0:11:2::1 (105.16.0.131)
      Origin IGP, metric 0, localpref 100, valid, external, best
      Community: 37100:2 37100:22000 37100:22004 37100:22060
      path 0828B828 RPKI State valid
      rx pathid: 0, tx pathid: 0x0
```

Courtesy of SEACOM: http://as37100.net

RPKI BGP State: Invalid

```
BGP routing table entry for 2001:1218:101::/48, version 149538323
Paths: (2 available, no best path)
  Not advertised to any peer
  Refresh Epoch 1
  37100 6453 8151 278
    2C0F:FEB0:B:3::1 (FE80::86B5:9C00:15F5:7C00) from
                                  2C0F:FEB0:B:3::1 (105.16.0.162)
      Origin IGP, metric 0, localpref 100, valid, external
      Community: 37100:1 37100:12
      path 0DA7D4FC RPKI State invalid
      rx pathid: 0, tx pathid: 0
```

Courtesy of SEACOM: http://as37100.net

RPKI BGP State: Not Found

```
BGP routing table entry for 2001:200::/32, version 124240929
Paths: (2 available, best #2, table default)
  Not advertised to any peer
  Refresh Epoch 1
  37100 2914 2500
    2C0F:FEB0:11:2::1 (FE80::2A8A:1C00:1560:5BC0) from
                                2C0F:FEB0:11:2::1 (105.16.0.131)
      Origin IGP, metric 0, localpref 100, valid, external, best
      Community: 37100:1 37100:13
      path 19D90E68 RPKI State not found
      rx pathid: 0, tx pathid: 0x0
```

Courtesy of SEACOM: http://as37100.net

Using RPKI

- Network operators can make decisions based on RPKI state:
  - Invalid – discard the prefix
  - Not found – let it through (maybe low local preference)
  - Valid – let it through (high local preference)
- Some operators even considering making “not found” a discard event
  - But then Internet IPv4 BGP table would shrink to about 20k prefixes and the IPv6 BGP table would shrink to about 3k prefixes!

RPKI Summary

- All AS operators must consider deploying
- An important step to securing the routing system
  - Origin validation
- Doesn’t secure the path, but that’s the next hurdle to cross
- With origin validation, the opportunities for malicious or accidental mis-origination disappear

Configuration Tips

Of passwords, tricks and templates

iBGP and IGPs Reminder!

- Make sure loopback is configured on router
  - iBGP between loopbacks, NOT real 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 ISPs 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 ISPs 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 ASes in a typical AS_PATH in the Internet today
  - The Internet is around 5 ASes deep on average
  - Largest AS_PATH is usually 16-20 ASNs

```
neighbor x.x.x.x maxas-limit 20
```

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

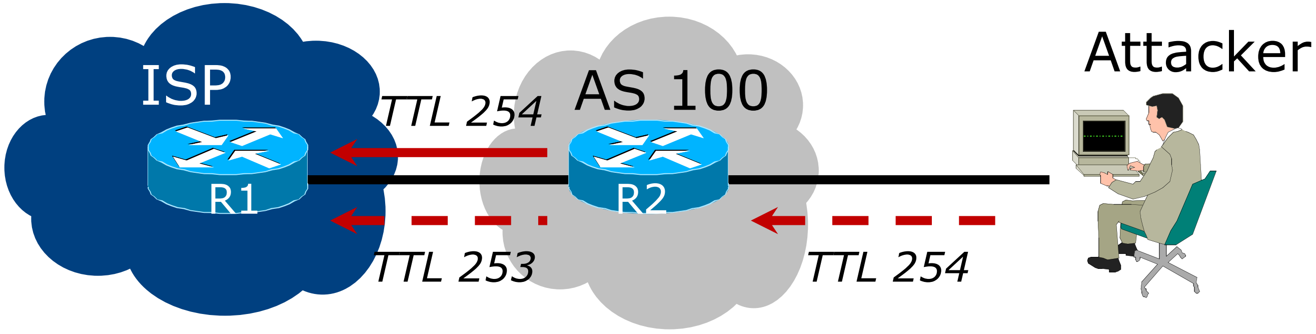
- Allow configuration of the maximum number of prefixes a BGP router will receive from a peer
- Two level control:
  - Warning threshold: log warning message
  - Maximum: tear down the BGP peering, manual intervention required to restart

```
neighbor <x.x.x.x> maximum-prefix <max> [restart N] [<threshold>] [warning-only]
```

- *restart* is an optional keyword which will restart the BGP session N minutes after being torn down
- *threshold* is an optional parameter between 1 to 100
  - Specify the percentage of <max> that will cause a warning message to be generated. Default is 75%.
- *warning-only* is an optional keyword which allows log messages to be generated but peering session will not be torn down

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

BGP TTL ‘hack’

- Configuration example:

```
neighbor 100.121.0.2 ttl-security hops 1
```

- BGP neighbour status:

```
Router# sh ip bgp neigh 100.121.0.2
...
Mininum incoming TTL 254, Outgoing TTL 255
Local host: 100.121.0.1, Local port: 41103
Foreign host: 100.121.0.2, Foreign port: 179
```

- The neighbour must set the same configuration
  - If they don’t, the BGP session will not come up

Templates

- Good practice to configure templates for everything
  - Vendor defaults tend not to be optimal or even very useful for ISPs
  - ISPs create their own defaults by using configuration templates
- eBGP and iBGP examples follow
  - Also see Team Cymru's BGP templates
    - http://www.team-cymru.org/documents.html

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 – it will consume considerable amounts of extra memory for BGP
- Remove private ASes 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/manrs
  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
  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
- 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 to 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

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 Best Current Practices

ISP Workshops