BGP Best Current Practices

ISP Workshops

Configuring BGP

Where do we start?

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

Cisco IOS Good Practices

- BGP in Cisco IOS is permissive by default
- 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
- Good practice is to ensure that each eBGP neighbour has inbound and outbound filter applied:

```
router bgp 64511
 neighbor 1.2.3.4 remote-as 64510
 neighbor 1.2.3.4 prefix-list as64510-in in
 neighbor 1.2.3.4 prefix-list as64510-out out
```

What is BGP for??

What is an IGP not for?

BGP versus OSPF/ISIS

- Internal Routing Protocols (IGPs)
  - Examples are ISIS 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/ISIS

- 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/ISIS

- 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
  - December 2015: 314000 /24s in IPv4 table of 573000 prefixes
- The same is happening for /48s with IPv6
  - December 2015: 11400 /48s in IPv6 table of 25100 prefixes

Configuring Aggregation – Cisco IOS

- ISP has 101.10.0.0/19 address block
- To put into BGP as an aggregate:

```
router bgp 64511
 network 101.10.0.0 mask 255.255.224.0
ip route 101.10.0.0 255.255.224.0 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

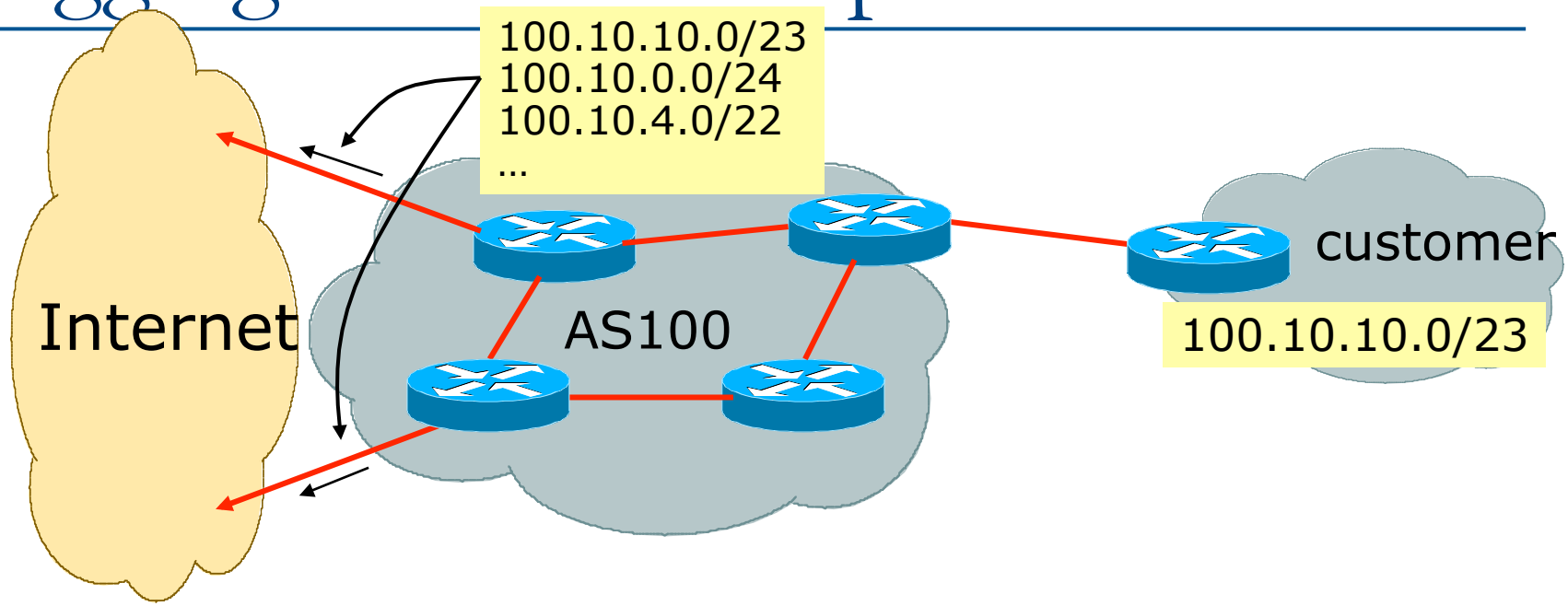
- Configuration Example

```
router bgp 64511
 network 101.10.0.0 mask 255.255.224.0
 neighbor 102.102.10.1 remote-as 101
 neighbor 102.102.10.1 prefix-list out-filter out
!
ip route 101.10.0.0 255.255.224.0 null0
!
ip prefix-list out-filter permit 101.10.0.0/19
ip prefix-list out-filter deny 0.0.0.0/0 le 32
!
```

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:
    - Now ranging from a /20 to a /24 depending on RIR
    - Different sizes for different address blocks
    - (APNIC changed its minimum allocation to /24 in October 2010)
  - For IPv6:
    - /48 for assignment, /32 for allocation
- Until recently there was no real reason to see anything longer than a /22 IPv4 prefix in the Internet
  - Maybe IPv4 run-out is starting to have an impact?

Aggregation – Example

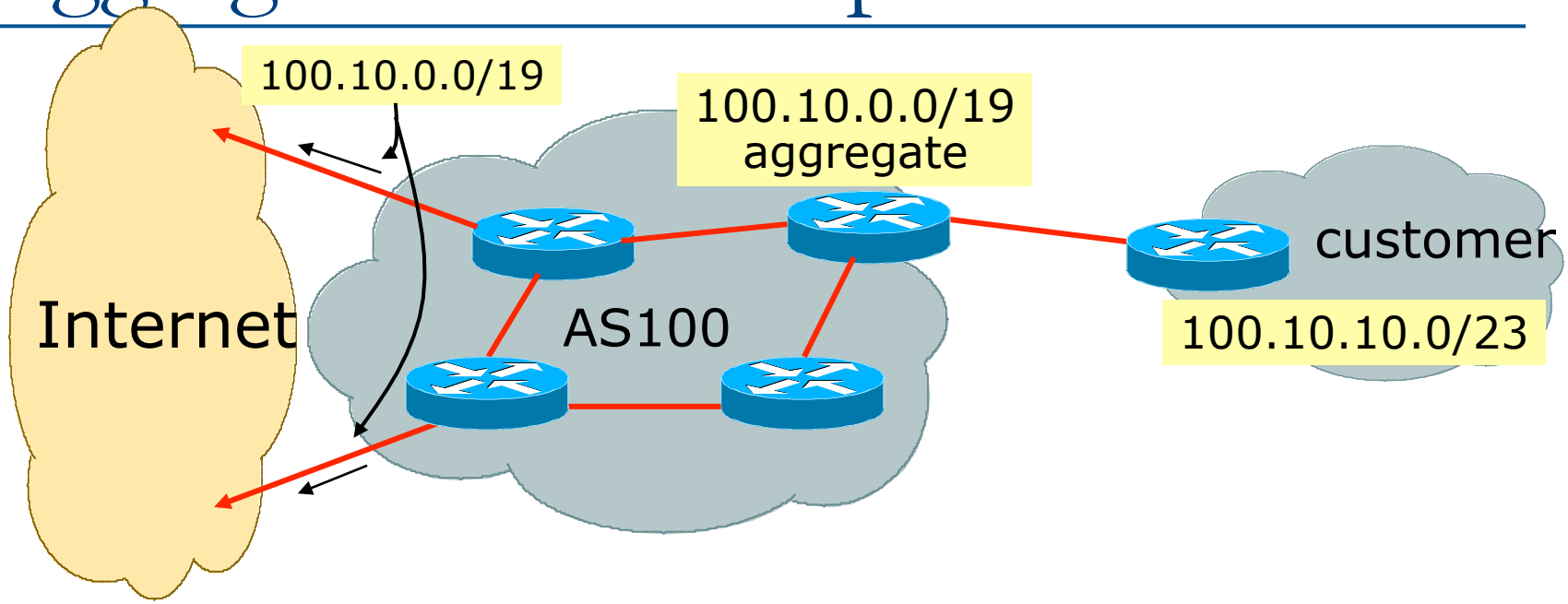


- Customer has /23 network assigned from AS100's /19 address block
- AS100 announces customers' individual networks to the Internet

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 (December 2015)

- Current Internet Routing Table Statistics
  - BGP Routing Table Entries 573136
  - Prefixes after maximum aggregation 212475
  - Unique prefixes in Internet 278860
  - Prefixes smaller than registry alloc 188598
  - /24s announced 313799
  - ASes in use 52219

  - (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

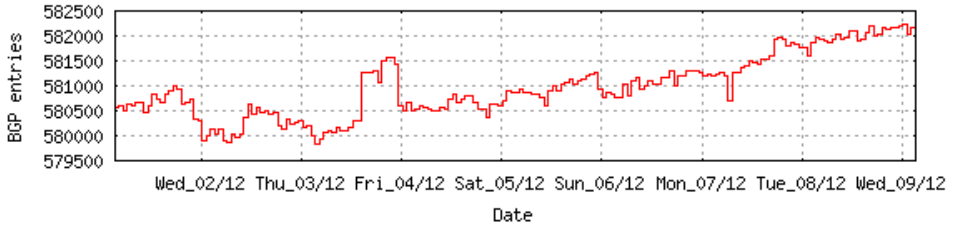
- Also computes the size of the routing table assuming ISPs performed optimal aggregation
- Website allows searches and computations of aggregation to be made on a per AS basis
  - Flexible and powerful tool to aid ISPs
  - Intended to show how greater efficiency in terms of BGP table size can be obtained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

Status Summary

Table History

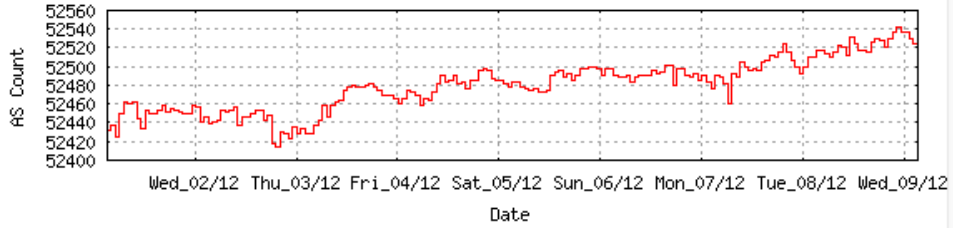
Date	Prefixes	CIDR Aggregated
02-12-15	580311	313738
03-12-15	580298	313913
04-12-15	580588	314014
05-12-15	580613	314275
06-12-15	580927	314460
07-12-15	581257	314665
08-12-15	581766	315027
09-12-15	582187	315075

Plot: BGP Table Size



AS Summary

52526	Number of ASes in routing system
20721	Number of ASes announcing only one prefix
5612	Largest number of prefixes announced by an AS AS4538: ERX-CERNET-BKB China Education and Research Network Center,CN
120893696	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)

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	%
5	AS6389	BELLSOUTH-NET-BLK - BellSouth.net Inc.,US	2508	2467	4	45	2463	98.21%

Prefix	AS Path	Aggregation Suggestion
12.81.90.0/23	6939 7018 6389	
12.81.120.0/24	6939 7018 6389	
12.83.5.0/24	6939 7018 6389	
12.83.7.0/24	6939 7018 6389	
65.0.0.0/12	6939 7018 6389	
65.0.0.0/18	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.0.0.0/19	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.0.40.0/22	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.0.50.0/23	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.0.64.0/18	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.0.128.0/18	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.0.192.0/19	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.0.224.0/19	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.1.0.0/19	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.1.32.0/19	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.1.64.0/19	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.1.224.0/20	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.1.240.0/20	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.2.0.0/16	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.2.0.0/17	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.2.128.0/17	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.3.224.0/19	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.4.64.0/18	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.4.192.0/18	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.1.0/24	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.12.0/22	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.16.0/22	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.20.0/23	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.21.0/24	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.22.0/23	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.24.0/22	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.28.0/22	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.32.0/20	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.34.0/24	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.46.0/24	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.57.0/24	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.64.0/22	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.68.0/22	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.74.0/23	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389
65.5.76.0/22	6939 7018 6389	- Withdrawn - matching aggregate 65.0.0.0/12 6939 7018 6389

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	2213	1409	209	1013	1200	54.23%

Prefix	AS Path	Aggregation Suggestion
64.6.160.0/23	6939 18566	
64.6.164.0/23	6939 18566	
64.6.166.0/24	6939 2828 18566	
64.6.167.0/24	6939 18566	
64.50.206.0/23	6939 18566	
64.51.126.0/23	6939 18566	
64.81.16.0/22	6939 1299 3356 18566	
64.81.20.0/22	6939 18566	
64.81.22.0/24	6939 18566	- Withdrawn - matching aggregate 64.81.20.0/22 6939 18566
64.81.24.0/21	6939 1299 3356 18566	+ Announce - aggregate of 64.81.24.0/22 (6939 1299 3356 18566) and 64.81.28.0/22 (6939 1299 3356
64.81.24.0/22	6939 1299 3356 18566	- Withdrawn - aggregated with 64.81.28.0/22 (6939 1299 3356 18566)
64.81.28.0/22	6939 1299 3356 18566	- Withdrawn - aggregated with 64.81.24.0/22 (6939 1299 3356 18566)
64.81.32.0/20	6939 1299 18566	
64.81.32.0/24	6939 1299 18566	- Withdrawn - matching aggregate 64.81.32.0/20 6939 1299 18566
64.81.33.0/24	6939 1299 18566	- Withdrawn - matching aggregate 64.81.32.0/20 6939 1299 18566
64.81.34.0/24	6939 1299 18566	- Withdrawn - matching aggregate 64.81.32.0/20 6939 1299 18566
64.81.35.0/24	6939 1299 18566	- Withdrawn - matching aggregate 64.81.32.0/20 6939 1299 18566
64.81.36.0/24	6939 1299 18566	- Withdrawn - matching aggregate 64.81.32.0/20 6939 1299 18566
64.81.37.0/24	6939 1299 18566	- Withdrawn - matching aggregate 64.81.32.0/20 6939 1299 18566
64.81.38.0/24	6939 1299 18566	- Withdrawn - matching aggregate 64.81.32.0/20 6939 1299 18566
64.81.39.0/24	6939 1299 18566	- Withdrawn - matching aggregate 64.81.32.0/20 6939 1299 18566
64.81.40.0/24	6939 1299 18566	- Withdrawn - matching aggregate 64.81.32.0/20 6939 1299 18566
64.81.44.0/24	6939 1299 18566	- Withdrawn - matching aggregate 64.81.32.0/20 6939 1299 18566
64.81.48.0/20	6939 1299 3356 18566	
64.81.48.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.49.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.50.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.51.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.52.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.53.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.54.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.55.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.56.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.57.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.58.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.59.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.60.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.61.0/24	6939 1299 3356 18566	- Withdrawn - matching aggregate 64.81.48.0/20 6939 1299 3356 18566
64.81.64.0/20	6939 1299 3356 18566	

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 1 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 08 December 2015 06:39 (UTC+1000)

50 Most active ASes for the past 7 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	2635	473609	7.53%	84	5638.20	AUTOMATTIC - Automattic, Inc,US
2	36992	470605	7.48%	446	1055.17	ETISALAT-MISR,EG
3	9829	437888	6.96%	2218	197.42	BSNL-NIB National Internet Backbone,IN
4	8452	197253	3.13%	1377	143.25	TE-AS TE-AS,EG
5	21669	105729	1.68%	21	5034.71	NJ-STATEWIDE-LIBRARY-NETWORK - New Jersey State Library,US
6	10493	80722	1.28%	7	11531.71	GCN-AS - Grand Central Networks Inc.,US
7	56041	72891	1.16%	525	138.84	CMNET-ZHEJIANG-AP China Mobile communications corporation,CN
8	56046	62134	0.99%	464	133.91	CMNET-JIANGSU-AP China Mobile communications corporation,CN
9	11259	61867	0.98%	58	1066.67	ANGOLATELECOM,AO
10	637	59791	0.95%	242	247.07	DNIC-ASBLK-00616-00665 - DoD Network Information Center,US
11	13118	51281	0.81%	97	528.67	ASN-YARTELECOM PJSC Rostelecom,RU
12	10859	44334	0.70%	26	1705.15	COMPUTER-SCIENCES-CORP-NTIS - Computer Sciences Corp - NTIS,US
13	55021	41747	0.66%	2	20873.50	OTTCOLO - ottcolo inc.,CA
14	22059	39506	0.63%	7	5643.71	-Reserved AS-,ZZ
15	25364	36044	0.57%	16	2252.75	EgyptCyberCenter-AS,EG
16	15475	33634	0.53%	15	2242.27	NOL,EG
17	48159	31888	0.51%	344	92.70	TIC-AS Telecommunication Infrastructure Company,IR
18	25576	30765	0.49%	18	1709.17	AFMIC,EG
19	39891	29648	0.47%	2473	11.99	ALJAWWALSTC-AS Saudi Telecom Company JSC,SA
20	47794	29503	0.47%	131	225.21	ATHEEB-AS Etihad Atheeb Telecom Company,SA
21	246	26506	0.42%	318	83.35	ASIFICS-GW-AS - 754th Electronic Systems Group,US
22	3816	24290	0.39%	978	24.84	COLOMBIA TELECOMUNICACIONES S.A. ESP,CO
23	11105	24278	0.39%	17	1428.12	SFU-AS - Simon Fraser University,CA
24	20030	23998	0.38%	12	1999.83	MCCLI-ARTELCO - Artelco,US
25	2472	23828	0.38%	10	2382.80	FR-DOM-GUYANE Guyane Francaise,FR
26	3709	23263	0.37%	27	861.59	NET-CITY-SA - City of San Antonio,US

50 Most active Prefixes for the past 7 days

RANK	PREFIX	UPDs	%	Origin AS -- AS NAME
1	192.0.119.0/24	118506	1.82%	2635 -- AUTOMATTIC - Automattic, Inc,US
2	192.0.122.0/24	118232	1.82%	2635 -- AUTOMATTIC - Automattic, Inc,US
3	192.0.121.0/24	118143	1.82%	2635 -- AUTOMATTIC - Automattic, Inc,US
4	192.0.120.0/24	117944	1.82%	2635 -- AUTOMATTIC - Automattic, Inc,US
5	209.212.8.0/24	105713	1.63%	21669 -- NJ-STATEWIDE-LIBRARY-NETWORK - New Jersey State Library,US
6	93.181.192.0/19	45426	0.70%	13118 -- ASN-YARTELECOM PJSC Rostelecom,RU
7	131.131.98.0/24	44287	0.68%	10859 -- COMPUTER-SCIENCES-CORP-NTIS - Computer Sciences Corp - NTIS,US
8	162.253.250.0/23	41742	0.64%	55021 -- OTTCOLO - ottcolo inc.,CA
9	172.81.88.0/22	32037	0.49%	10493 -- GCN-AS - Grand Central Networks Inc.,US
10	74.201.42.0/24	24361	0.38%	10493 -- GCN-AS - Grand Central Networks Inc.,US
11	74.201.41.0/24	24312	0.37%	10493 -- GCN-AS - Grand Central Networks Inc.,US
12	64.34.125.0/24	20156	0.31%	22059 -- -Reserved AS-,ZZ
13	76.191.107.0/24	19345	0.30%	22059 -- -Reserved AS-,ZZ
14	155.133.79.0/24	17902	0.28%	200671 -- SKOK-JAWORZNO SKOK Jaworzno,PL
15	197.216.41.0/24	14130	0.22%	11259 -- ANGOLATELECOM,AO
16	168.128.73.0/24	14096	0.22%	132084 -- OPSOURCE-AP 5201 Great America Pkwy # 120,AU
17	185.78.104.0/24	11776	0.18%	34341 -- NCEM Namvaran Consulting Engineers and Managers,IR
18	213.109.33.0/24	10451	0.16%	35745 -- PROVECTOR-AS KSU Provector Mariusz Dziakowicz,PL
19	202.41.70.0/24	9849	0.15%	2697 -- ERX-ERNET-AS Education and Research Network,IN
20	94.73.56.0/21	8870	0.14%	42081 -- SPEEDY-NET-AS Speedy net AD,BG
21	202.41.83.0/24	8615	0.13%	2697 -- ERX-ERNET-AS Education and Research Network,IN
22	196.216.241.0/24	8257	0.13%	37348 -- CAC,EG
23	62.140.96.0/19	7654	0.12%	36992 -- ETISALAT-MISR,EG
24	67.61.206.0/24	7403	0.11%	11492 -- CABLEONE - CABLE ONE, INC.,US
25	62.114.104.0/21	7379	0.11%	36992 -- ETISALAT-MISR,EG
26	62.114.224.0/20	7376	0.11%	36992 -- ETISALAT-MISR,EG
27	62.114.128.0/21	7365	0.11%	36992 -- ETISALAT-MISR,EG
28	62.114.200.0/21	7362	0.11%	36992 -- ETISALAT-MISR,EG
29	62.114.112.0/21	7359	0.11%	36992 -- ETISALAT-MISR,EG
30	62.114.96.0/21	7357	0.11%	36992 -- ETISALAT-MISR,EG
31	62.114.160.0/21	7341	0.11%	36992 -- ETISALAT-MISR,EG

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 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:        202.12.28.0 - 202.12.29.255
netname:        APNIC-AP
descr:          Asia Pacific Network Information Centre
descr:          Regional Internet Registry for the Asia-Pacific
descr:          6 Cordelia Street
descr:          South Brisbane, QLD 4101
descr:          Australia
country:        AU
admin-c:        AIC1-AP
tech-c:         NO4-AP
mnt-by:         APNIC-HM
mnt-irt:        IRT-APNIC-AP
changed:        hm-changed@apnic.net
status:         ASSIGNED PORTABLE
changed:        hm-changed@apnic.net 20110309
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 whois.ripe.net 193.128.0.0
inetnum:        193.128.0.0 - 193.133.255.255
netname:        UK-PIPEX-193-128-133
descr:          Verizon UK Limited
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-lower:      AS1849-MNT
mnt-routes:     AS1849-MNT
mnt-routes:     WCOM-EMEA-RICE-MNT
mnt-irt:        IRT-MCI-GB
source:         RIPE # Filtered
```

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 100.50.0.0/20 block
  - Should only announce this to upstreams
  - Upstreams should only accept this from them
- Configuration on upstream

```
router bgp 100
 neighbor 102.102.10.1 remote-as 101
 neighbor 102.102.10.1 prefix-list customer in
 neighbor 102.102.10.1 prefix-list default out
!
ip prefix-list customer permit 100.50.0.0/20
!
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
 neighbor 102.102.10.1 remote-as 101
 neighbor 102.102.10.1 prefix-list my-peer in
!
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
```

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
 network 101.10.0.0 mask 255.255.224.0
 neighbor 101.5.7.1 remote-as 101
 neighbor 101.5.7.1 prefix-list infilter in
 neighbor 101.5.7.1 prefix-list outfilter out
!
ip prefix-list infilter permit 0.0.0.0/0
!
ip prefix-list outfilter permit 101.10.0.0/19
```

Receiving Prefixes: From Upstream/Transit Provider

- Upstream Router Configuration

```
router bgp 101
 neighbor 101.5.7.2 remote-as 100
 neighbor 101.5.7.2 default-originate
 neighbor 101.5.7.2 prefix-list cust-in in
 neighbor 101.5.7.2 prefix-list cust-out out
!
ip prefix-list cust-in permit 101.10.0.0/19
!
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 101.5.7.1 remote-as 101
 neighbor 101.5.7.1 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 addr
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 proto
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 & Expmnt
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 permit 2001::/32             ! Teredo
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 permit 2002::/16             ! 6to4
ipv6 prefix-list v6in-filter deny 2002::/16 le 128        ! 6to4 subnets
ipv6 prefix-list v6in-filter deny 2020:3030::/32 le 128   ! Local Prefix
ipv6 prefix-list v6in-filter deny 3ffe::/16 le 128        ! Old 6bone
ipv6 prefix-list v6in-filter permit 2000::/3 le 48        ! Global Unicast
ipv6 prefix-list v6in-filter deny ::/0 le 128
```

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 215.17.3.1 255.255.255.255
!
interface Serial 5/0
 ip unnumbered loopback 0
 ip verify unicast reverse-path
!
ip route 215.34.10.0 255.255.252.0 Serial 5/0
!
router bgp 100
 network 215.34.10.0 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 215.34.10.0 255.255.252.0 Serial 5/0
!
router bgp 100
 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 215.34.10.0/22 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

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 15
```

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

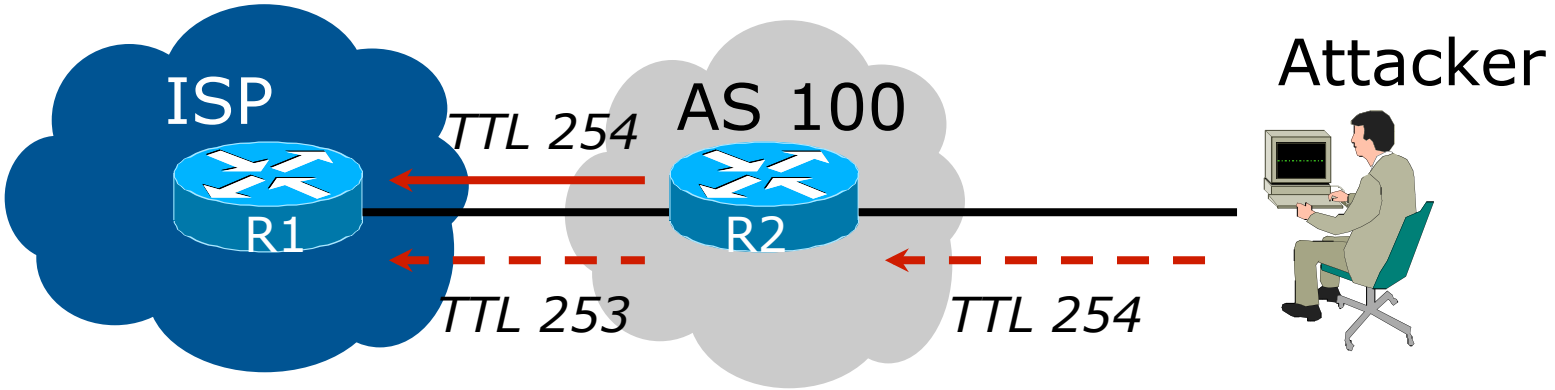
- Allow configuration of the maximum number of prefixes a BGP router will receive from a peer
- Two level control:
  - Warning threshold: log warning message
  - Maximum: tear down the BGP peering, manual intervention required to restart

```
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

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 version 3 BGP 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!!

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
- It’s all about scaling – if your network won’t scale, then it won’t be successful

BGP Best Current Practices

ISP Workshops