Deploying BGP in ISP (BGP102)

Dawit Birhanu (dawit@cisco.com)

Technical Marketing Engineer, Cisco Systems

Presentation Slides

- Will be available on
  Location will be provided
- Feel free to ask questions any time

Agenda

- The role of IGPs and iBGP
- Aggregation
- Receiving Prefixes
- Origin Validation (RPKI)
- Preparing the Network
- Configuration Tips

cisco

The role of IGPs and iBGP

BGP versus OSPF/ISIS

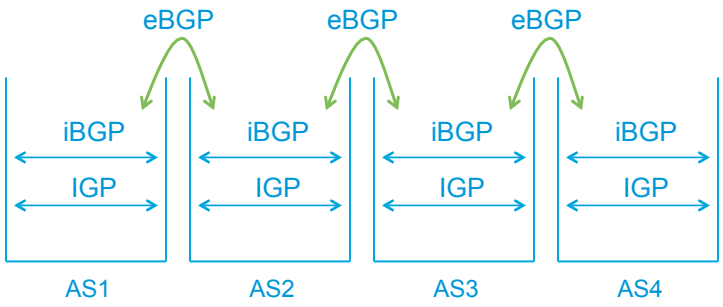
- Internal Routing Protocols (IGPs)
  - examples are ISIS and OSPF
  - used for carrying infrastructure addresses
  - **NOT** used for carrying Internet prefixes or customer prefixes
  - design goal is to minimize number of prefixes in IGP to aid scalability and rapid convergence

BGP versus OSPF/ISIS

- BGP used internally (iBGP) and externally (eBGP)
- iBGP used to carry
  - some/all Internet prefixes across backbone
  - customer prefixes
- eBGP used to
  - exchange prefixes with other ASes
  - implement routing policy

BGP/IGP model used in ISP networks

- Model representation



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

Injecting prefixes into iBGP

- Use iBGP to carry customer prefixes
  - Don't ever use IGP
- Point static route to customer interface if customer is single-homed
  - Enter network into BGP process
  - Ensure that implementation options are used so that the prefix always remains in iBGP, regardless of state of interface
  - i.e. avoid iBGP flaps caused by interface flaps
- Consider eBGP with customer only if:
  - Customer is multi-homed to your network or to other provider, and
  - Customer has its own ASN from one of the RIRs

cisco

Aggregation

Quality or Quantity?

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
- Unfortunately too many people are still thinking about class Cs, resulting in a proliferation of /24s in the Internet routing table

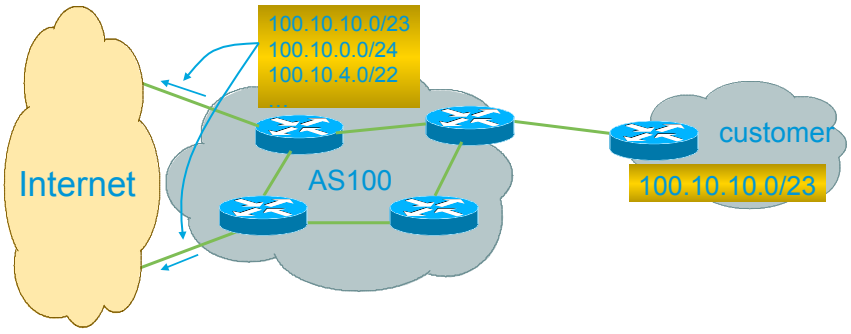
Aggregation

- 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 purposes
  - (see BGP Multihoming Tutorial)
- Aggregate should be generated internally
  - Not on the network borders!

Announcing an Aggregate

- ISPs who don't and won't aggregate are held in poor regard by community
- Registries publish their minimum allocation size
  - Anything from a /20 to a /22 depending on RIR
  - Different sizes for different address blocks
- No real reason to see anything longer than a /22 prefix in the Internet
  - BUT there are currently >185000 /24s!
- But: APNIC changed (Oct 2010) its minimum allocation size on all blocks to /24
  - IPv4 run-out is starting to have an impact

Aggregation – Bad Example

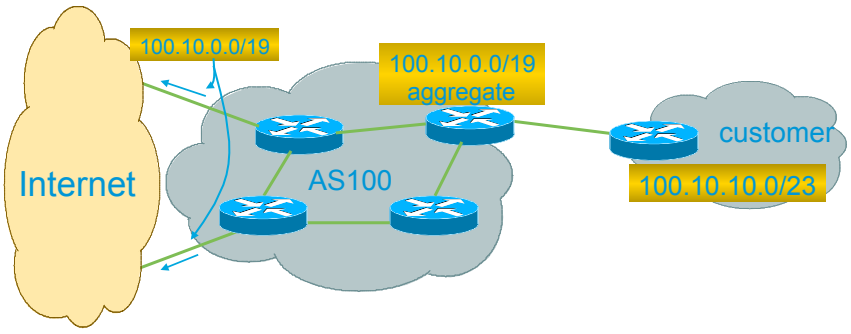


- Customer has /23 network assigned from AS100's /19 address block
- AS100 announces customers' individual networks to the Internet

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 – Good 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 (July 2012)

- Current Internet Routing Table Statistics

BGP Routing Table Entries	420845
*CIDR Aggregated	243337
Prefixes after maximum aggregation	181133
*Unique prefixes in Internet	178173
*Prefixes smaller than registry alloc	149545
/24s announced	224148
ASes in use	41910

"The New Swamp"

- Swamp space is name used for areas of poor aggregation
  - The original swamp was 192.0.0.0/8 from the former class C block
    - Name given just after the deployment of CIDR
  - The new swamp is creeping across all parts of the Internet
    - Not just RIR space, but "legacy" space too

"The New Swamp" RIR Space – February 1999

RIR blocks contribute 88% of the Internet Routing Table

Block	Networks	Block	Networks	Block	Networks	Block	Networks
24/8	165	*79/8*	*0*	*118/8*	*0*	*201/8*	*0*
41/8	*0*	*80/8*	*0*	*119/8*	*0*	202/8	2276
58/8	*0*	*81/8*	*0*	*120/8*	*0*	203/8	3622
59/8	*0*	*82/8*	*0*	*121/8*	*0*	204/8	3792
60/8	*0*	*83/8*	*0*	*122/8*	*0*	205/8	2584
61/8	3	*84/8*	*0*	*123/8*	*0*	206/8	3127
62/8	87	*85/8*	*0*	*124/8*	*0*	207/8	2723
63/8	20	*86/8*	*0*	*125/8*	*0*	208/8	2817
64/8	*0*	*87/8*	*0*	*126/8*	*0*	209/8	2574
65/8	*0*	*88/8*	*0*	*173/8*	*0*	210/8	617
66/8	*0*	*89/8*	*0*	*174/8*	*0*	*211/8*	*0*
67/8	*0*	*90/8*	*0*	*186/8*	*0*	212/8	717
68/8	*0*	*91/8*	*0*	*187/8*	*0*	213/8	1
69/8	*0*	*96/8*	*0*	*189/8*	*0*	216/8	943
70/8	*0*	*97/8*	*0*	*190/8*	*0*	*217/8*	*0*
71/8	*0*	*98/8*	*0*	192/8	6275	*218/8*	*0*
72/8	*0*	*99/8*	*0*	193/8	2390	*219/8*	*0*
73/8	*0*	*112/8*	*0*	194/8	2932	*220/8*	*0*
74/8	*0*	*113/8*	*0*	195/8	1338	*221/8*	*0*
75/8	*0*	*114/8*	*0*	196/8	513	*222/8*	*0*
76/8	*0*	*115/8*	*0*	198/8	4034		

"The New Swamp" RIR Space – February 2010

RIR blocks contribute about 87% of the Internet Routing Table

Block	Networks	Block	Networks	Block	Networks	Block	Networks
24/8	3328	79/8	1119	118/8	1349	201/8	4136
41/8	3448	80/8	2335	119/8	1694	202/8	11354
58/8	1675	81/8	1709	120/8	531	203/8	11677
59/8	1575	82/8	1358	121/8	1756	204/8	5744
60/8	888	83/8	1357	122/8	2687	205/8	3037
61/8	2890	84/8	1341	123/8	2400	206/8	3951
62/8	2418	85/8	2492	124/8	2259	207/8	4635
63/8	3114	86/8	780	125/8	2514	208/8	6498
64/8	*6601*	87/8	1466	126/8	106	209/8	5536
65/8	3966	88/8	1068	173/8	1994	210/8	4977
66/8	*7782*	89/8	3168	174/8	1089	211/8	3130
67/8	3771	90/8	377	186/8	1223	212/8	3550
68/8	3221	91/8	4555	187/8	1501	213/8	3442
69/8	5280	96/8	778	189/8	3063	216/8	7645
70/8	2008	97/8	725	190/8	6945	217/8	3136
71/8	1327	98/8	1312	192/8	6952	218/8	1512
72/8	4050	99/8	288	193/8	6820	219/8	1303
73/8	4	112/8	883	194/8	5177	220/8	2108
74/8	*5074*	113/8	890	195/8	5325	221/8	980
75/8	1164	114/8	996	196/8	1857	222/8	1058
76/8	1034	115/8	1616	198/8	4504		
77/8	1964	116/8	1755	199/8	4372		
78/8	1397	117/8	1611	200/8	8884		

"The New Swamp" Summary

- RIR space shows creeping deaggregation

  It seems that an RIR /8 block averages around 5000 prefixes (and upwards) once fully allocated

- Food for thought:

  The 120 RIR /8s combined will cause:

  635000 prefixes with 5000 prefixes per /8 density

  762000 prefixes with 6000 prefixes per /8 density

  Plus 12% due to "non RIR space deaggregation"

  → Routing Table size of 853440 prefixes

"The New Swamp" Summary

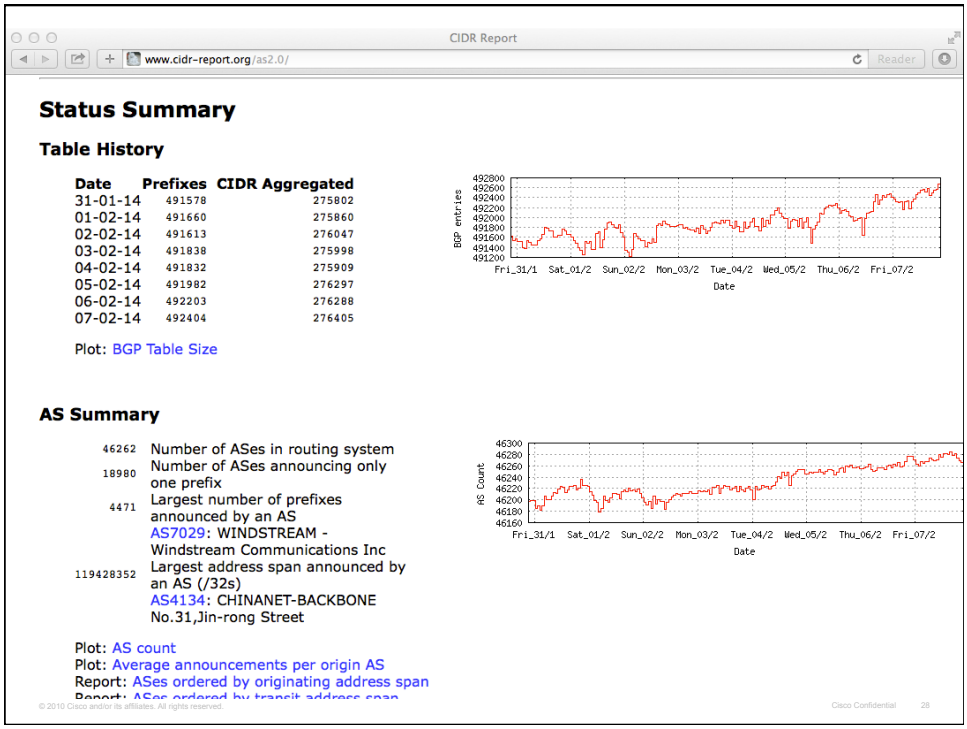
- Rest of address space is showing similar deaggregation too ☹
- What are the reasons?
  - Main justification is traffic engineering
- Real reasons are:
  - Lack of knowledge
  - Laziness
  - Deliberate & knowing actions

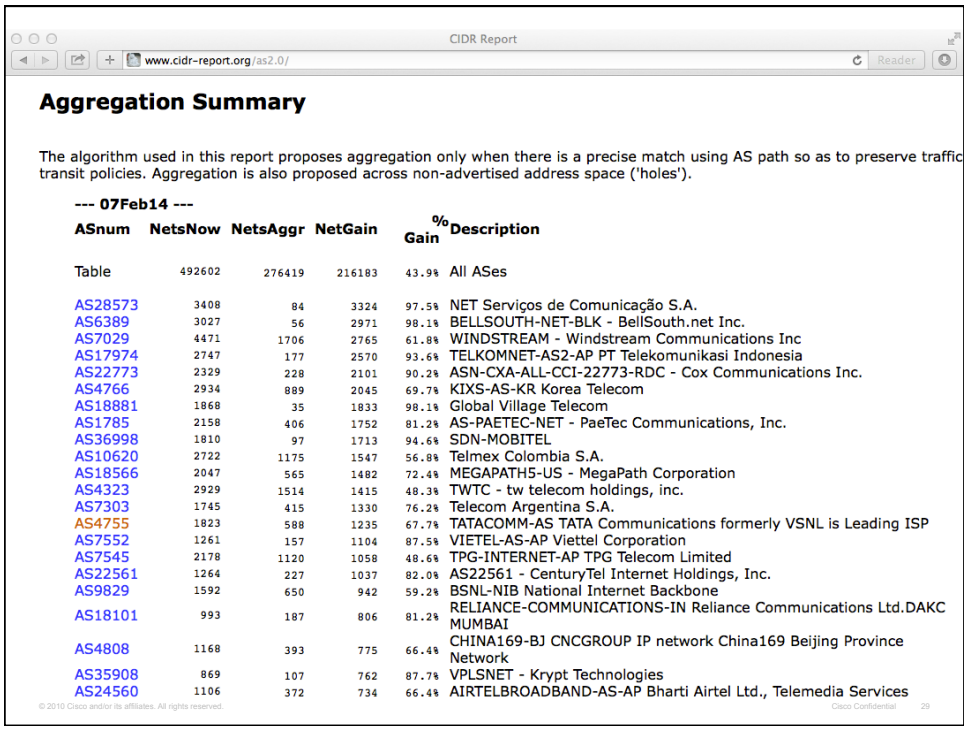
Efforts to improve aggregation

- The CIDR Report
  - Initiated and operated for many years by Tony Bates and revised by Philip Smith
  - Now combined with Geoff Huston's routing analysis
    - **www.cidr-report.org**
  - 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 recommendation
  - **RIPE-399 — http://www.ripe.net/ripe/docs/ripe-399.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





CIDR Report

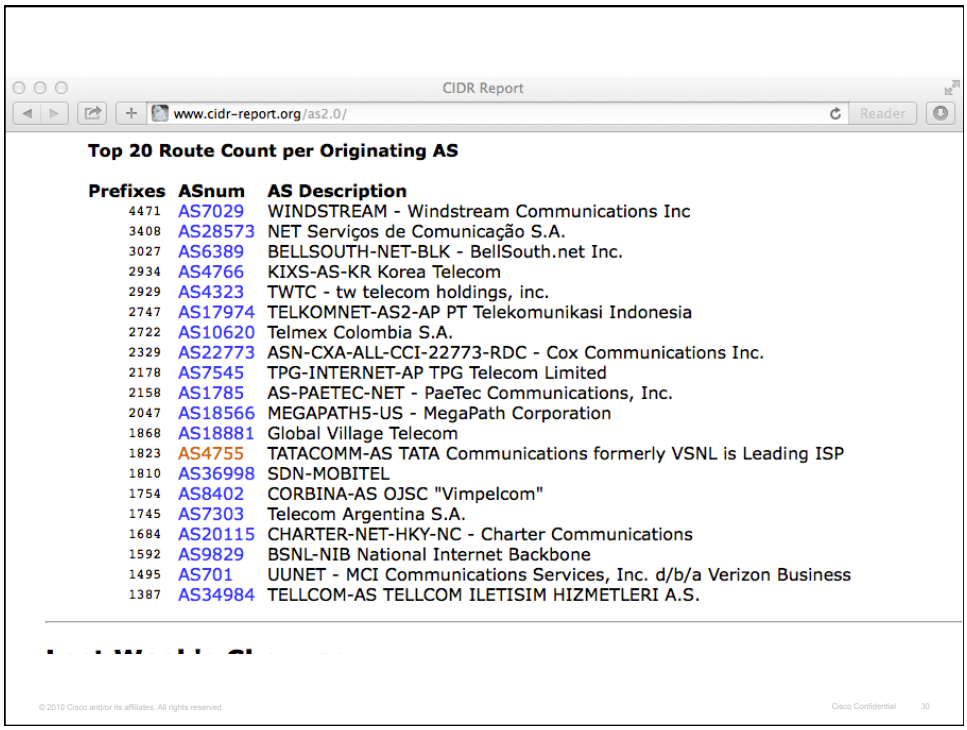
www.cidr-report.org/as2.0/

Aggregation Summary

The algorithm used in this report proposes aggregation only when there is a precise match using AS path so as to preserve traffic transit policies. Aggregation is also proposed across non-advertised address space ('holes').

--- 07Feb14 ---

ASnum	NetsNow	NetsAggr	NetGain	% Gain	Description
Table	492602	276419	216183	43.9%	All ASes
AS28573	3408	84	3324	97.5%	NET Serviços de Comunicação S.A.
AS6389	3027	56	2971	98.1%	BELLSOUTH-NET-BLK - BellSouth.net Inc.
AS7029	4471	1706	2765	61.8%	WINDSTREAM - Windstream Communications Inc
AS17974	2747	177	2570	93.6%	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
AS22773	2329	228	2101	90.2%	ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
AS4766	2934	889	2045	69.7%	KIXS-AS-KR Korea Telecom
AS18881	1868	35	1833	98.1%	Global Village Telecom
AS1785	2158	406	1752	81.2%	AS-PAETEC-NET - PaeTec Communications, Inc.
AS36998	1810	97	1713	94.6%	SDN-MOBITEL
AS10620	2722	1175	1547	56.8%	Telmex Colombia S.A.
AS18566	2047	565	1482	72.4%	MEGAPATH5-US - MegaPath Corporation
AS4323	2929	1514	1415	48.3%	TWTC - tw telecom holdings, inc.
AS7303	1745	415	1330	76.2%	Telecom Argentina S.A.
AS4755	1823	588	1235	67.7%	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
AS7552	1261	157	1104	87.5%	VIETEL-AS-AP Viettel Corporation
AS7545	2178	1120	1058	48.6%	TPG-INTERNET-AP TPG Telecom Limited
AS22561	1264	227	1037	82.0%	AS22561 - CenturyTel Internet Holdings, Inc.
AS9829	1592	650	942	59.2%	BSNL-NIB National Internet Backbone
AS18101	993	187	806	81.2%	RELIANCE-COMMUNICATIONS-IN Reliance Communications Ltd.DAKC MUMBAI
AS4808	1168	393	775	66.4%	CHINA169-BJ CNCGROUP IP network China169 Beijing Province Network
AS35908	869	107	762	87.7%	VPLSNET - Krypt Technologies
AS24560	1106	372	734	66.4%	AIRTELBROADBAND-AS-AP Bharti Airtel Ltd., Telemedia Services

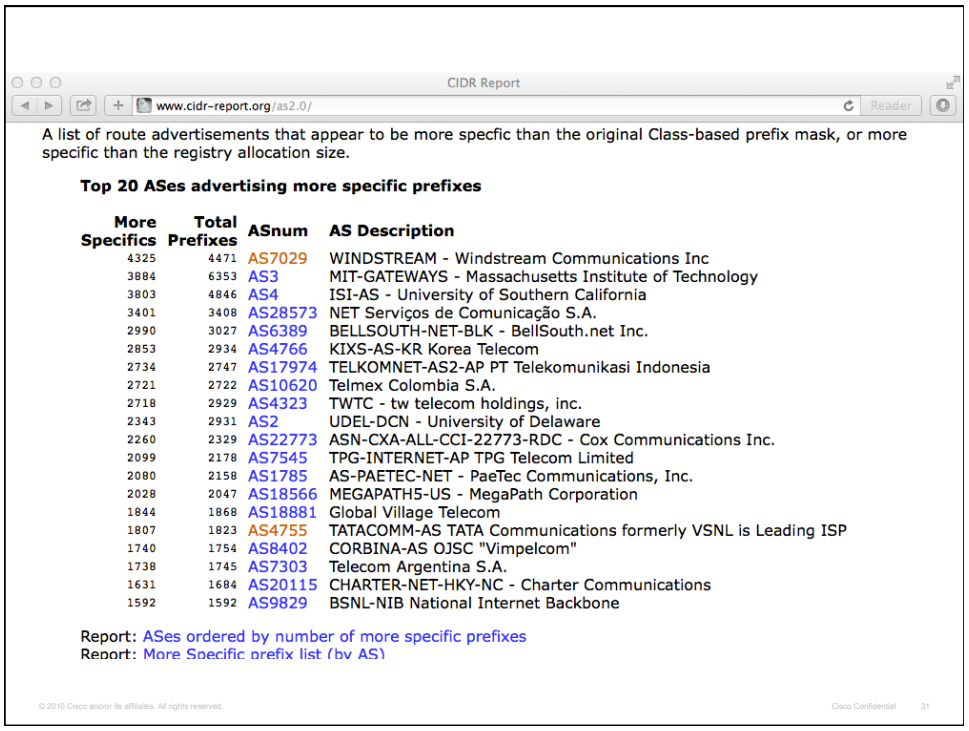


CIDR Report

www.cidr-report.org/as2.0/

Top 20 Route Count per Originating AS

Prefixes	ASnum	AS Description
4471	AS7029	WINDSTREAM - Windstream Communications Inc
3408	AS28573	NET Serviços de Comunicação S.A.
3027	AS6389	BELLSOUTH-NET-BLK - BellSouth.net Inc.
2934	AS4766	KIXS-AS-KR Korea Telecom
2929	AS4323	TWTC - tw telecom holdings, inc.
2747	AS17974	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
2722	AS10620	Telmex Colombia S.A.
2329	AS22773	ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
2178	AS7545	TPG-INTERNET-AP TPG Telecom Limited
2158	AS1785	AS-PAETEC-NET - PaeTec Communications, Inc.
2047	AS18566	MEGAPATH5-US - MegaPath Corporation
1868	AS18881	Global Village Telecom
1823	AS4755	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
1810	AS36998	SDN-MOBITEL
1754	AS8402	CORBINA-AS OJSC "Vimpelcom"
1745	AS7303	Telecom Argentina S.A.
1684	AS20115	CHARTER-NET-HKY-NC - Charter Communications
1592	AS9829	BSNL-NIB National Internet Backbone
1495	AS701	UUNET - MCI Communications Services, Inc. d/b/a Verizon Business
1387	AS34984	TELLCOM-AS TELLCOM ILETISIM HIZMETLERI A.S.

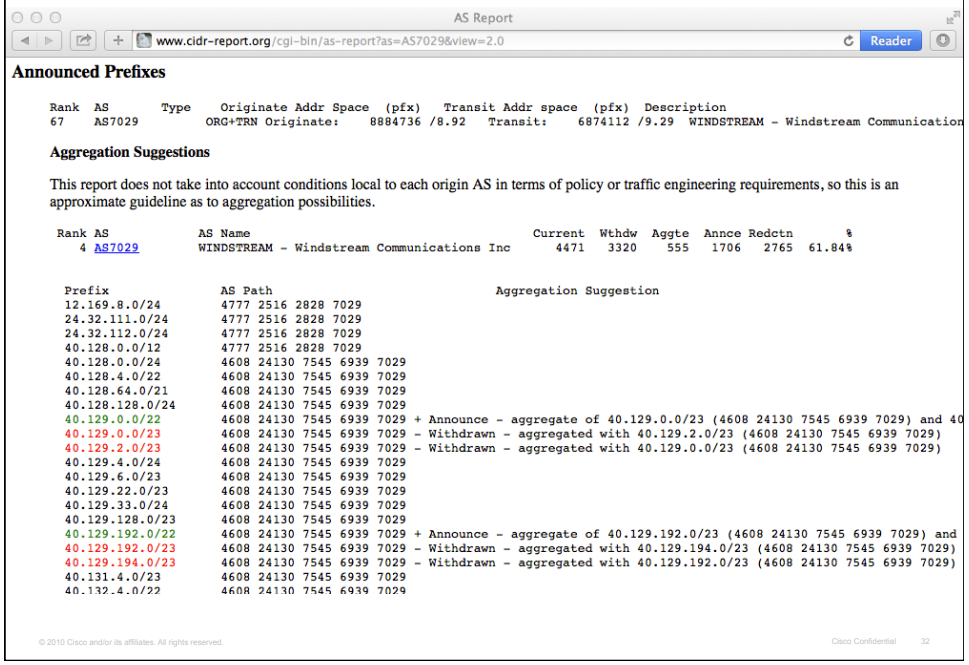


A list of route advertisements that appear to be more specfic than the original Class-based prefix mask, or more specific than the registry allocation size.

Top 20 ASes advertising more specific prefixes

More Specifics	Total Prefixes	ASnum	AS Description
4325	4471	AS7029	WINDSTREAM - Windstream Communications Inc
3884	6353	AS3	MIT-GATEWAYS - Massachusetts Institute of Technology
3803	4846	AS4	ISI-AS - University of Southern California
3401	3408	AS28573	NET Serviços de Comunicação S.A.
2990	3027	AS6389	BELLSOUTH-NET-BLK - BellSouth.net Inc.
2853	2934	AS4766	KIXS-AS-KR Korea Telecom
2734	2747	AS17974	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
2721	2722	AS10620	Telmex Colombia S.A.
2718	2929	AS4323	TWTC - tw telecom holdings, inc.
2343	2931	AS2	UDEL-DCN - University of Delaware
2260	2329	AS22773	ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
2099	2178	AS7545	TPG-INTERNET-AP TPG Telecom Limited
2080	2158	AS1785	AS-PAETEC-NET - PaeTec Communications, Inc.
2028	2047	AS18566	MEGAPATH5-US - MegaPath Corporation
1844	1868	AS18881	Global Village Telecom
1807	1823	AS4755	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
1740	1754	AS8402	CORBINA-AS OJSC "Vimpelcom"
1738	1745	AS7303	Telecom Argentina S.A.
1631	1684	AS20115	CHARTER-NET-HKY-NC - Charter Communications
1592	1592	AS9829	BSNL-NIB National Internet Backbone

Report: ASes ordered by number of more specific prefixes
Report: More Specific prefix list (by AS)



Announced Prefixes

```
Rank  AS       Type    Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
67    AS7029         ORG+TRN Originate:    8884736 /8.92   Transit:    6874112 /9.29  WINDSTREAM - Windstream Communication
```

Aggregation Suggestions

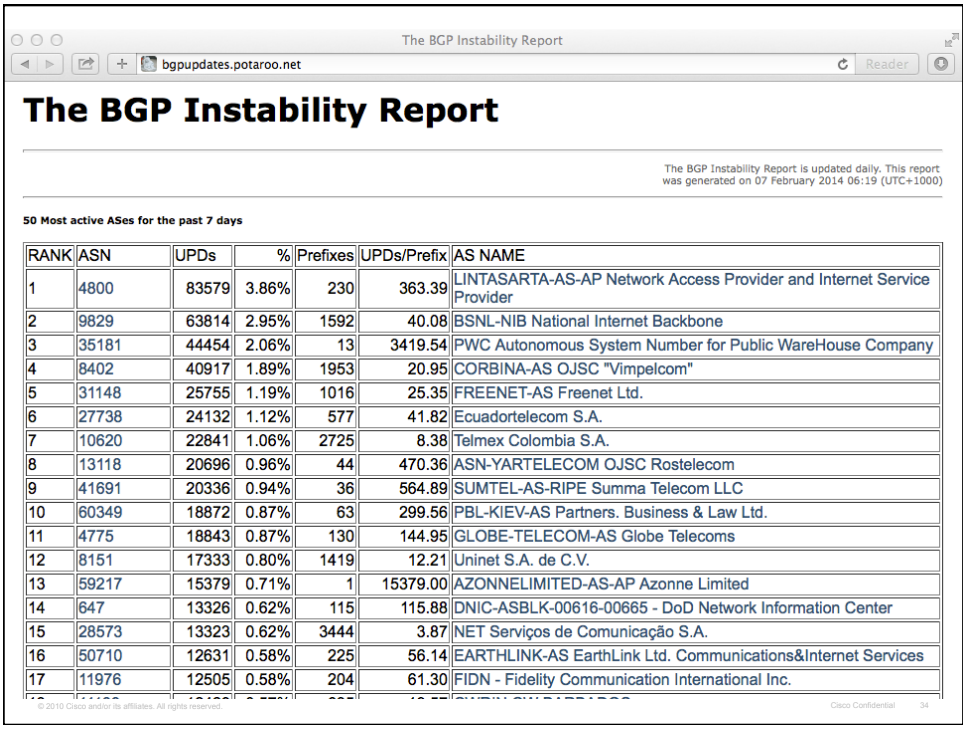
This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

```
 Rank AS           AS Name                                    Current  Wthdw  Aggte  Annce Redctn       %
    4 AS7029       WINDSTREAM - Windstream Communications Inc    4471   3320    555   1706   2765  61.84%

  Prefix              AS Path                              Aggregation Suggestion
  12.169.8.0/24       4777 2516 2828 7029
  24.32.111.0/24      4777 2516 2828 7029
  24.32.112.0/24      4777 2516 2828 7029
  40.128.0.0/12       4777 2516 2828 7029
  40.128.0.0/24       4608 24130 7545 6939 7029
  40.128.4.0/22       4608 24130 7545 6939 7029
  40.128.64.0/21      4608 24130 7545 6939 7029
  40.128.128.0/24     4608 24130 7545 6939 7029
  40.129.0.0/22       4608 24130 7545 6939 7029 + Announce - aggregate of 40.129.0.0/23 (4608 24130 7545 6939 7029) and 40
  40.129.0.0/23       4608 24130 7545 6939 7029 - Withdrawn - aggregated with 40.129.2.0/23 (4608 24130 7545 6939 7029)
  40.129.2.0/23       4608 24130 7545 6939 7029 - Withdrawn - aggregated with 40.129.0.0/23 (4608 24130 7545 6939 7029)
  40.129.4.0/24       4608 24130 7545 6939 7029
  40.129.6.0/23       4608 24130 7545 6939 7029
  40.129.22.0/23      4608 24130 7545 6939 7029
  40.129.33.0/24      4608 24130 7545 6939 7029
  40.129.128.0/23     4608 24130 7545 6939 7029
  40.129.192.0/22     4608 24130 7545 6939 7029 + Announce - aggregate of 40.129.192.0/23 (4608 24130 7545 6939 7029) and
  40.129.192.0/23     4608 24130 7545 6939 7029 - Withdrawn - aggregated with 40.129.194.0/23 (4608 24130 7545 6939 7029)
  40.129.194.0/23     4608 24130 7545 6939 7029 - Withdrawn - aggregated with 40.129.192.0/23 (4608 24130 7545 6939 7029)
  40.131.4.0/23       4608 24130 7545 6939 7029
  40.132.4.0/22       4608 24130 7545 6939 7029
```

Importance of Aggregation

- Size of routing table
  - Router Memory is not so much of a problem as it was in the 1990s
  - Routers can be specified to carry 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
  - **http://bgpupdates.potaroo.net/instability/bgpupd.html**



The BGP Instability Report

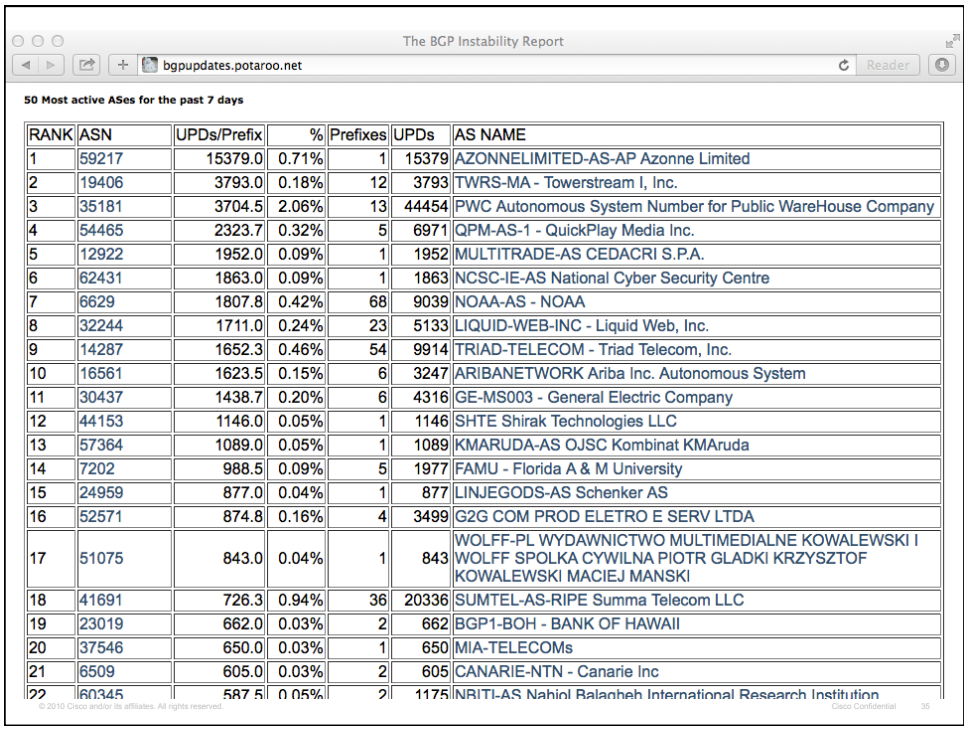
bgpupdates.potaroo.net

The BGP Instability Report

The BGP Instability Report is updated daily. This report was generated on 07 February 2014 06:19 (UTC+1000)

50 Most active ASes for the past 7 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	4800	83579	3.86%	230	363.39	LINTASARTA-AS-AP Network Access Provider and Internet Service Provider
2	9829	63814	2.95%	1592	40.08	BSNL-NIB National Internet Backbone
3	35181	44454	2.06%	13	3419.54	PWC Autonomous System Number for Public WareHouse Company
4	8402	40917	1.89%	1953	20.95	CORBINA-AS OJSC "Vimpelcom"
5	31148	25755	1.19%	1016	25.35	FREENET-AS Freenet Ltd.
6	27738	24132	1.12%	577	41.82	Ecuadortelecom S.A.
7	10620	22841	1.06%	2725	8.38	Telmex Colombia S.A.
8	13118	20696	0.96%	44	470.36	ASN-YARTELECOM OJSC Rostelecom
9	41691	20336	0.94%	36	564.89	SUMTEL-AS-RIPE Summa Telecom LLC
10	60349	18872	0.87%	63	299.56	PBL-KIEV-AS Partners. Business & Law Ltd.
11	4775	18843	0.87%	130	144.95	GLOBE-TELECOM-AS Globe Telecoms
12	8151	17333	0.80%	1419	12.21	Uninet S.A. de C.V.
13	59217	15379	0.71%	1	15379.00	AZONNELIMITED-AS-AP Azonne Limited
14	647	13326	0.62%	115	115.88	DNIC-ASBLK-00616-00665 - DoD Network Information Center
15	28573	13323	0.62%	3444	3.87	NET Serviços de Comunicação S.A.
16	50710	12631	0.58%	225	56.14	EARTHLINK-AS EarthLink Ltd. Communications&Internet Services
17	11976	12505	0.58%	204	61.30	FIDN - Fidelity Communication International Inc.



The BGP Instability Report

bgpupdates.potaroo.net

50 Most active ASes for the past 7 days

RANK	ASN	UPDs/Prefix	%	Prefixes	UPDs	AS NAME
1	59217	15379.0	0.71%	1	15379	AZONNELIMITED-AS-AP Azonne Limited
2	19406	3793.0	0.18%	12	3793	TWRS-MA - Towerstream I, Inc.
3	35181	3704.5	2.06%	13	44454	PWC Autonomous System Number for Public WareHouse Company
4	54465	2323.7	0.32%	5	6971	QPM-AS-1 - QuickPlay Media Inc.
5	12922	1952.0	0.09%	1	1952	MULTITRADE-AS CEDACRI S.P.A.
6	62431	1863.0	0.09%	1	1863	NCSC-IE-AS National Cyber Security Centre
7	6629	1807.8	0.42%	68	9039	NOAA-AS - NOAA
8	32244	1711.0	0.24%	23	5133	LIQUID-WEB-INC - Liquid Web, Inc.
9	14287	1652.3	0.46%	54	9914	TRIAD-TELECOM - Triad Telecom, Inc.
10	16561	1623.5	0.15%	6	3247	ARIBANETWORK Ariba Inc. Autonomous System
11	30437	1438.7	0.20%	6	4316	GE-MS003 - General Electric Company
12	44153	1146.0	0.05%	1	1146	SHTE Shirak Technologies LLC
13	57364	1089.0	0.05%	1	1089	KMARUDA-AS OJSC Kombinat KMAruda
14	7202	988.5	0.09%	5	1977	FAMU - Florida A & M University
15	24959	877.0	0.04%	1	877	LINJEGODS-AS Schenker AS
16	52571	874.8	0.16%	4	3499	G2G COM PROD ELETRO E SERV LTDA
17	51075	843.0	0.04%	1	843	WOLFF-PL WYDAWNICTWO MULTIMEDIALNE KOWALEWSKI I WOLFF SPOLKA CYWILNA PIOTR GLADKI KRZYSZTOF KOWALEWSKI MACIEJ MANSKI
18	41691	726.3	0.94%	36	20336	SUMTEL-AS-RIPE Summa Telecom LLC
19	23019	662.0	0.03%	2	662	BGP1-BOH - BANK OF HAWAII
20	37546	650.0	0.03%	1	650	MIA-TELECOMs
21	6509	605.0	0.03%	2	605	CANARIE-NTN - Canarie Inc
22	60345	587.5	0.05%	2	1175	NRITI-AS Nahiol Balagheh International Research Institution

Aggregation Summary

- Aggregation on the Internet could be MUCH better
  - 35% saving on Internet routing table size is quite feasible
  - Tools are available
    - Commands on the routers are not hard
    - CIDR-Report webpage

Receiving Prefixes

Receiving Prefixes

- There are three scenarios for receiving prefixes from other 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 the five RIR databases to see if this address space really has been assigned to the customer

  The tool: whois

Receiving Prefixes: From Customers

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

```
$ whois -h whois.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 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
    - www.isc.org/sw/IRRToolSet/
- Alternatively, you can use origin-AS validation
  - Recommended if (or when) your routers support it
  - Enables you to automatically validate that the origin AS in the AS path is valid using RIRs registries
  - Discussed in the next section

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 Tutorial
- Ask upstream/transit provider to either:
  - originate a default-route
    - *OR*
  - announce one prefix you can use as default

Receiving Prefixes: From Upstream/Transit Provider

- If necessary to receive prefixes from any provider, care is required.
  - Don't accept default (unless you need it)
  - Don't accept your own prefixes
- For IPv4:
  - Don't accept private (RFC1918) and certain special use prefixes:
    - **http://www.rfc-editor.org/rfc/rfc5735.txt**
  - Don't accept prefixes longer than /24 (?)
- For IPv6:
  - Don't accept certain special use prefixes:
    - **http://www.rfc-editor.org/rfc/rfc5156.txt**
  - Don't accept prefixes longer than /48 (?)

Receiving Prefixes: From Upstream/Transit Provider

- Check Team Cymru's list of "bogons"
  - www.team-cymru.org/Services/Bogons/http.html
- 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 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

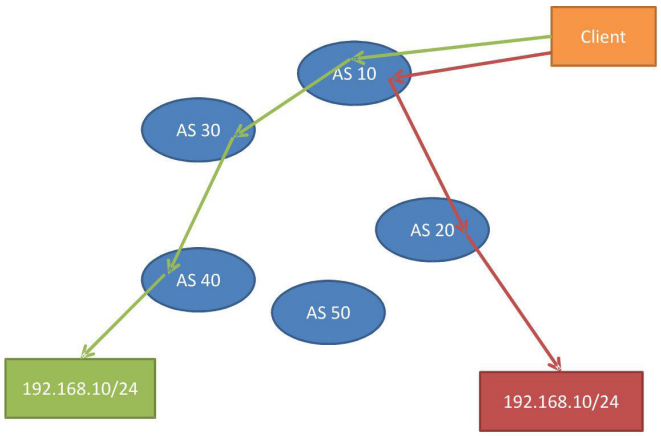
cisco

BGP Origin-AS Validation

Security issue for BGP route distribution

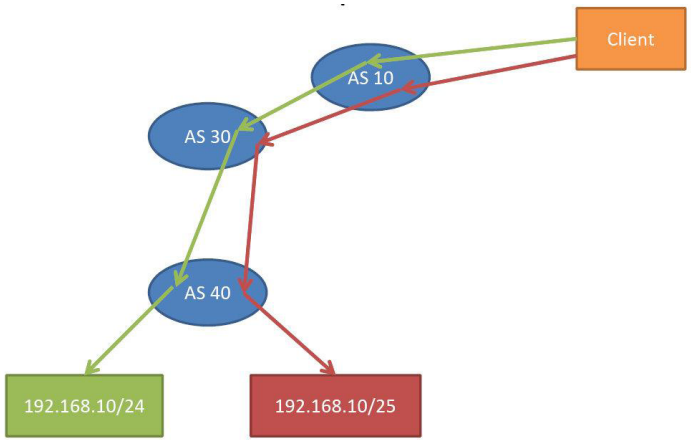
- Any AS can inject any prefixes in BGP, leading to prefix hijacking done by whichever mistake or malicious
- The manifestation of prefix hijacking are
  - an AS announcing someone else's prefix
  - as AS announcing a more specific of someone else's prefix
- The actual incidents are:
  - http://www.networkworld.com/news/2009/011509-bgp-attacks.html
- Need a mechanism to differentiate between invalid and legitimate routes for a BGP destination

Same Prefix: Shorter AS_PATH length wins



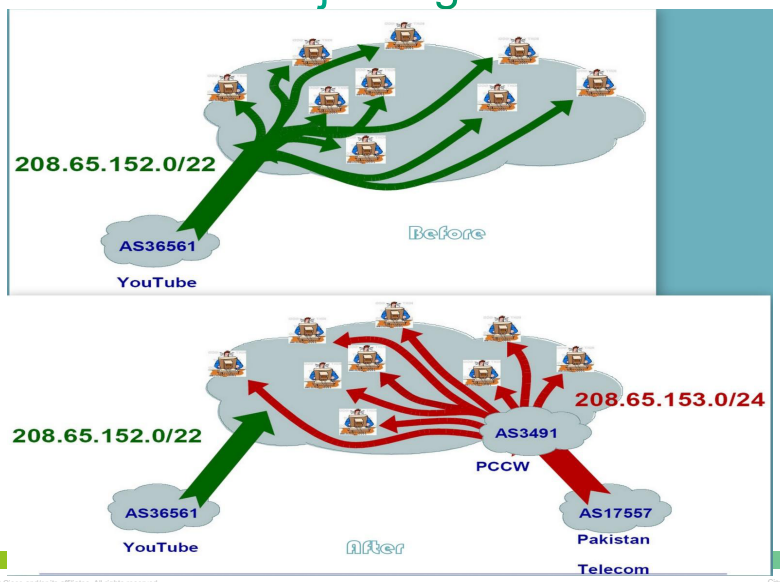
Source: nanog 46 preso

Same Prefix: More specific wins



Source: nanog 46 preso

Youtube Prefix Hijacking



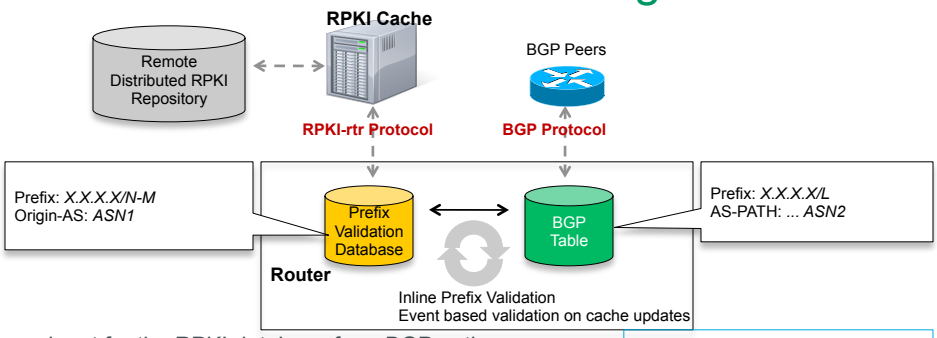
Standardization: IETF

- IETF Security Inter Domain Routing WG
  Focus on Inter Provider Internet Security
- Origin-AS Validation
  - http://datatracker.ietf.org/wg/sidr/
  - draft-ietf-sidr-pfx-validate-10.txt
  - draft-ietf-sidr-rpki-rtr-26.txt
  - RFC6483

RPKI (Resource Public Key Infrastructure)

- RPKI is a globally distributed database containing, among other things, information mapping BGP (Internet) prefixes to their authorized origin-AS numbers
- Routers running BGP can connect to the RPKI to validate the origin-AS of BGP paths

RPKI Database and BGP Design



- Input for the RPKI database for a BGP path:
  - BGP prefix/mask-length (X.X.X.X/N or X:X::X/N)
  - Origin-AS
- If the BGP prefix/mask-length has no covering ROAs in the RPKI database, the validity of path is "unknown"
- If the BGP prefix is covered by one or more ROAs in the RPKI database,
  - If any of the covering ROAs maps to the input origin-AS, the validity of the BGP route is "valid"
  - If none of the covering ROAs map to the input origin-AS, the validity of the BGP route is "invalid"

ROA: Route Origin Authorization

Origin-AS Validity Check Example

BGP Prefix / Origin-AS	RPKI Database ROAs	
10.0.1/24 AS 300	10/8-20 AS 100	Does not cover BGP prefix
	10.0/16-24 AS 200	Cover BGP prefix
	10.0/16-32 AS 300	Cover BGP prefix / **Origin AS matches**

valid

BGP Prefix / Origin-AS	RPKI Database ROAs	
10.0.1/24 AS 400	10/8-20 AS 100	Does not cover BGP prefix
	10.0/16-24 AS 200	Cover BGP prefix
	10.0/16-32 AS 300	Cover BGP prefix

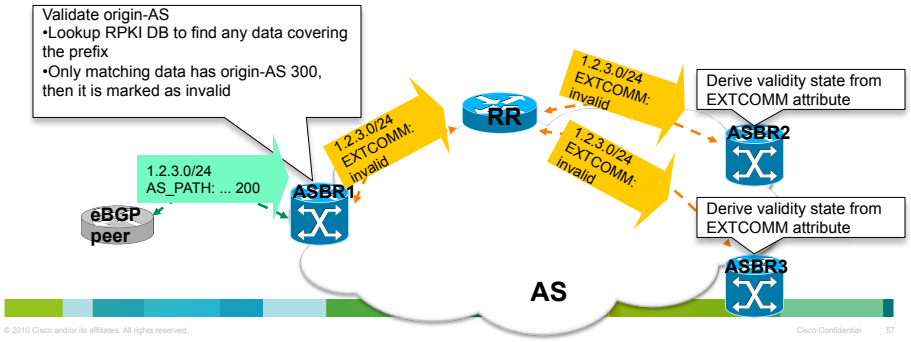
invalid

BGP Prefix / Origin-AS	RPKI Database ROAs	
20.0.1/24 AS 500	10/8-20 AS 100	Does not cover BGP prefix
	10.0/16-24 AS 200	Does not cover BGP prefix

unknown

iBGP Signaling of Origin-AS Validity State

- When a BGP route is received from outside AS, ASBRs should check this received path for origin-AS validity
- ASBRs that validates the origin-AS should signal the validity state of the route to its iBGP peers through a non-transitive BGP extended community attribute
- Upon receiving validity state information via extended community, iBGP peers can derive the validity state without having to lookup RPKI database



cisco

Preparing the Network

Before we begin …

Preparing the Network

- We will deploy BGP across the network before we try and multihome
- BGP will be used therefore an ASN is required
- If multihoming to different ISPs, public ASN needed:
  - Either go to upstream ISP who is a registry member, or
  - Apply to the RIR yourself for a one off assignment, or
  - Ask an ISP who is a registry member, or
  - Join the RIR and get your own IP address allocation too
    - (this option strongly recommended)!

Preparing the Network Initial Assumptions

- The network is not running any BGP at the moment
  - single statically routed connection to upstream ISP
- The network is not running any IGP at all
  - Static default and routes through the network to do “routing”

Preparing the Network
First Step: IGP

- Decide on an IGP: OSPF or ISIS ☺
- Assign loopback interfaces and /32 address to each router which will run the IGP
  - Loopback is used for OSPF and BGP router id anchor
  - Used for iBGP and route origination
- Deploy IGP (e.g. OSPF)
  - IGP can be deployed with NO IMPACT on the existing static routing
  - e.g. OSPF distance might be 110m static distance is 1
  - Smallest distance wins

Preparing the Network
IGP (cont)

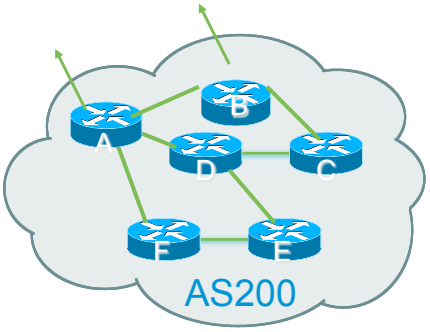
- Be prudent deploying IGP – keep the Link State Database Lean!
  - Router loopbacks go in IGP
  - WAN point to point links go in IGP
  - (In fact, any link where IGP dynamic routing will be run should go into IGP)
  - Summarise on area/level boundaries (if possible) – i.e. think about your IGP address plan

Preparing the Network IGP (cont)

- Routes which don't go into the IGP include:
  - Dynamic assignment pools (DSL/Cable/Dial)
  - Customer point to point link addressing
    - (using next-hop-self in iBGP ensures that these do NOT need to be in IGP)
  - Static/Hosting LANs
  - Customer assigned address space
  - Anything else not listed in the previous slide

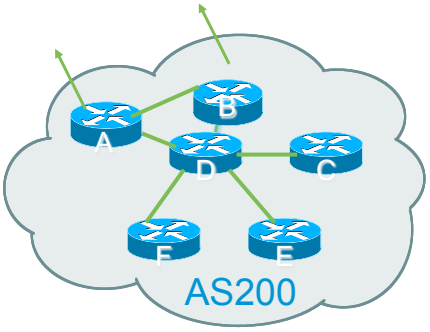
Preparing the Network Second Step: iBGP

- Second step is to configure the local network to use iBGP
- iBGP can run on
  - all routers, or
  - a subset of routers, or
  - just on the upstream edge
- *iBGP must run on all routers which are in the transit path between external connections*



Preparing the Network
Second Step: iBGP (Transit Path)

- *iBGP must run on all routers which are in the transit path between external connections*
- Routers C, E and F are not in the transit path
  - Static routes or IGP will suffice
- Router D is in the transit path
  - Will need to be in iBGP mesh, otherwise routing loops will result



Preparing the Network
Layers

- Typical SP networks have three layers:
  - Core – the backbone, usually the transit path
  - Distribution – the middle, PoP aggregation layer
  - Aggregation – the edge, the devices connecting customers

Preparing the Network Aggregation Layer

- iBGP is optional
  - Many ISPs run iBGP here, either partial routing (more common) or full routing (less common)
  - Full routing is not needed unless customers want full table
  - Partial routing is cheaper/easier, might usually consist of internal prefixes and, optionally, external prefixes to aid external load balancing
    - Communities and peer-groups make this administratively easy
- Many aggregation devices can't run iBGP
  - Static routes from distribution devices for address pools
  - IGP for best exit

Preparing the Network Distribution Layer

- Usually runs iBGP
  - Partial or full routing (as with aggregation layer)
- But does not have to run iBGP
  - IGP is then used to carry customer prefixes (does not scale)
  - IGP is used to determine nearest exit
- Networks which plan to grow large should deploy iBGP from day one
  - Migration at a later date is extra work
  - No extra overhead in deploying iBGP.
  - Indeed IGP benefits

Preparing the Network Core Layer

- Core of network is usually the transit path
- iBGP necessary between core devices
  - Full routes or partial routes:
    - Transit ISPs carry full routes in core
    - Edge ISPs carry partial routes only
- Core layer includes AS border routers

Preparing the Network iBGP Implementation

Decide on:

- Best iBGP policy
  - Will it be full routes everywhere, or partial, or some mix?
- iBGP scaling technique
  - Community policy?
  - Route-reflectors?
  - Configuration templates such as neighbor groups, sessions groups?

Preparing the Network iBGP Implementation

- Then deploy iBGP:
  - Step 1: Introduce iBGP mesh on chosen routers
    - make sure that iBGP distance is greater than IGP distance (it usually is)
  - Step 2: Install "customer" prefixes into iBGP
    - Check! Does the network still work?
  - Step 3: Carefully remove the static routing for the prefixes now in IGP and iBGP
    - Check! Does the network still work?
  - Step 4: Deployment of eBGP follows

Preparing the Network iBGP Implementation

Install "customer" prefixes into iBGP?

- Customer assigned address space
  - Network statement/static route combination
  - Use unique community to identify customer assignments
- Customer facing point-to-point links
  - Redistribute connected through filters which only permit point-to-point link addresses to enter iBGP
  - Use a unique community to identify point-to-point link addresses (these are only required for your monitoring system)
- Dynamic assignment pools & local LANs
  - Simple network statement will do this
  - Use unique community to identify these networks

Preparing the Network iBGP Implementation

Carefully remove static routes?

- Work on one router at a time:
  - Check that static route for a particular destination is also learned by the iBGP
  - If so, remove it
  - If not, establish why and fix the problem
  - (Remember to look in the RIB, not the FIB!)
- Then the next router, until the whole PoP is done
- Then the next PoP, and so on until the network is now dependent on the IGP and iBGP you have deployed

Preparing the Network Completion

- Previous steps are NOT flag day steps
  - Each can be carried out during different maintenance periods, for example:
  - Step One on Week One
  - Step Two on Week Two
  - Step Three on Week Three
  - And so on
  - And with proper planning will have NO customer visible impact at all

Configuration Tips

iBGP and IGPs Reminder!

- Make sure loopback is configured on router
  - iBGP between loopbacks, NOT real interfaces
- Make sure IGP carries loopback /32 address
- Consider the DMZ nets:
  - Use unnumbered interfaces?
  - Use next-hop-self on iBGP neighbours
  - Or carry the DMZ /30s 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 /30 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
- July 26, 2012 Internet AS path report for AS6447 (http://bgp.potaroo.net/as6447/) shows that
  - Average AS path length is 3.8
  - Maximum AS path length is 13
  - Maximum prepended AS path length is 34

Limiting AS Path Length

- Some announcements have ridiculous lengths of AS-paths:

```
*> 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 is an error in one IPv6 implementation

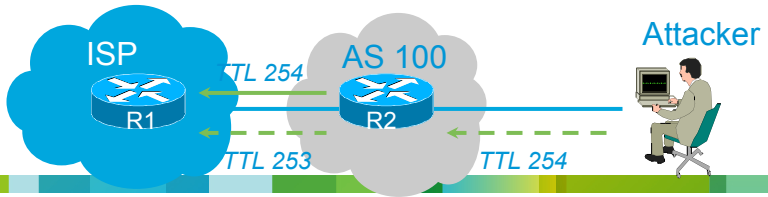
```
*>  96.27.246.0/24      2497 1239 12026 12026 12026
12026 12026 12026 12026 12026 12026 12026 12026 12026
12026 12026 12026 12026 12026 12026 12026 12026 12026
12026 i
```

  This example shows 21 prepends (for no obvious reason)

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

Generalized TTL Security Mechanism (GTSM)

- Also known as BGP TTL Security "Hack" (BTSH)
- Implement RFC5082 on BGP peerings
  - 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
  - Some implementations drop it in HW without any CPU impact



Generalized TTL Security Mechanism

- GTSM:
  - Both neighbours must agree to use the feature
  - TTL check is much easier to perform than MD5
- 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 www.nanog.org/mtg-0302/hack.html 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/ReadingRoom/Documents/

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 accidents will happen
- Hardwire BGP to version 4, if there is a version configuration option
  - Yes, this is being paranoid!

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 – particularly, from man-in-the-middle type of attack
- Powerful preventative tool, especially when combined with filters and the TTL "hack"

eBGP Template Example

- 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
- Use password agreed between you and peer on eBGP session
- Use TTL security (GTSM) if both peers support it

eBGP Template Example continued

- 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!
- Either make BGP admin distance higher than that of any IGP, or make sure to block your own prefixes inbound,
  - 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

Thank you.