



BGP Best Practices

Philip Smith <pfs@cisco.com>

UKNOF 8

17th September 2007

Goodenough College, London

Deploying BGP

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



The role of IGP and iBGP

Ships in the night?

Or

Good foundations?

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

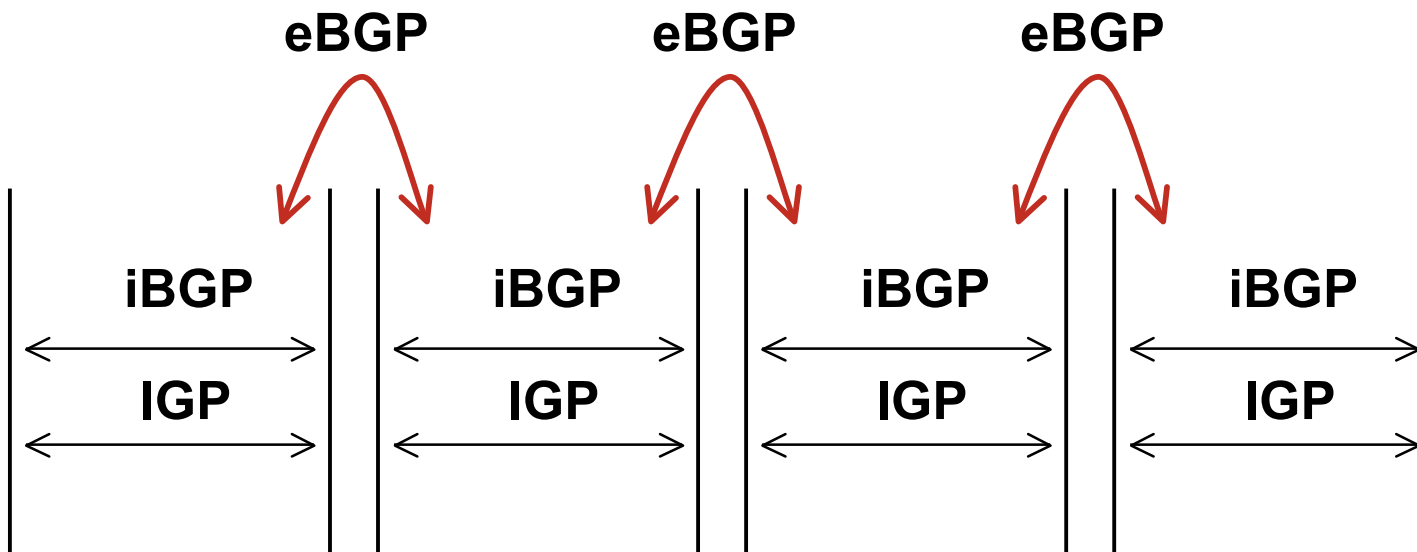
ISP design goal is to **minimise** number of prefixes in IGP to aid scalability and rapid convergence

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
- eBGP is **NOT** the same as iBGP

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
- 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



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

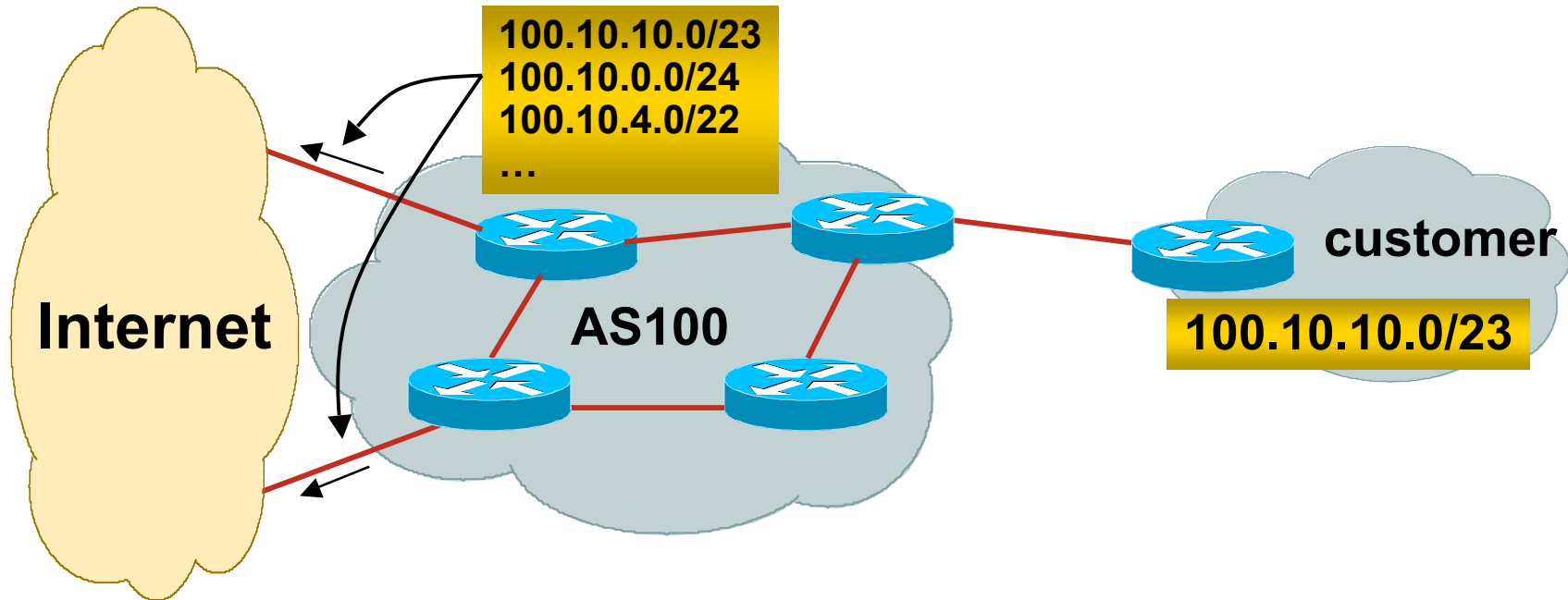
Aggregation

- Address block should be announced to the Internet as an aggregate
- Subprefixes of address block should **NOT** be announced to Internet unless traffic engineering when multihoming
- 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 >120000 /24s!

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

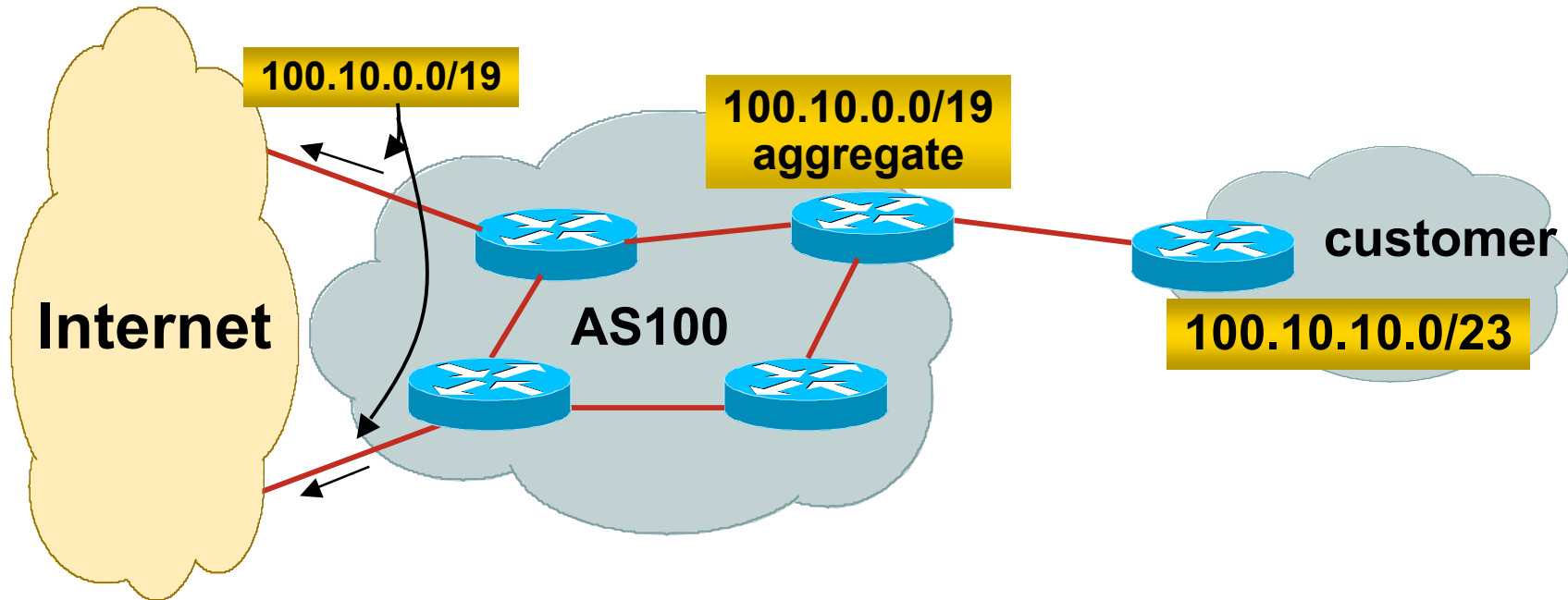
Load on Internet backbone routers as network is reinserted into routing table

Some ISP's suppress the flaps

Internet may take 10-20 min or longer to be visible

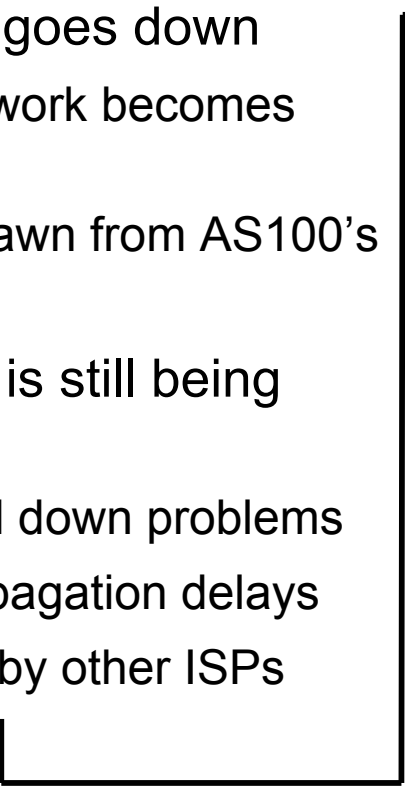
Where is the Quality of Service???

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?

The Internet Today (September 2007)

- Current Internet Routing Table Statistics

BGP Routing Table Entries	230291
Prefixes after maximum aggregation	120032
Unique prefixes in Internet	111045
Prefixes smaller than registry alloc	122198
/24s announced	121356
only 5708 /24s are from 192.0.0.0/8	
ASes in use	26164

BGP Report (bgp.potaroo.net)

- 199336 total announcements in October 2006
- 129795 prefixes

After aggregating including full AS PATH info

i.e. **including** each ASN's **traffic engineering**

35% saving possible

- 109034 prefixes

After aggregating by Origin AS

i.e. **ignoring** each ASN's **traffic engineering**

10% saving possible

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

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

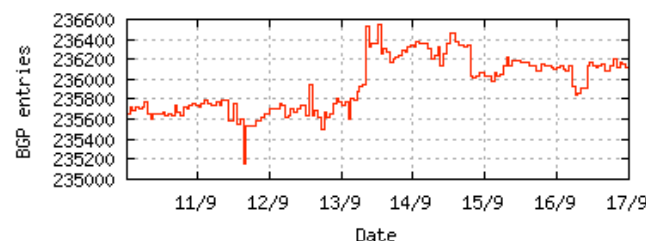
Very effectively challenges the traffic engineering excuse

Status Summary

Table History

Date	Prefixes	CIDR Aggregated
10-09-07	235597	154635
11-09-07	235740	154322
12-09-07	235653	154483
13-09-07	235774	150682
14-09-07	236349	151825
15-09-07	236070	151835
16-09-07	236109	154346
17-09-07	236118	152308

Plot: [BGP Table Size](#)



AS Summary

26286	Number of ASes in routing system
11095	Number of ASes announcing only one prefix
1941	Largest number of prefixes announced by an AS AS4538 : ERX-CERNET-BKB China Education and Research Network Center
89157120	Largest address span announced by an AS (/32s) AS721 : DISA-ASNBLK - DoD Network Information Center

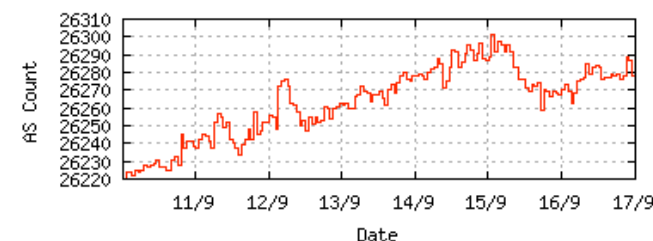
Plot: [AS count](#)

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)



AS Report

<http://www.cidr-report.org/cgi-bin/as-report?as=AS4134&view=2.0>

Google

AS Report

Announced Prefixes

Rank	AS	Type	Originate Addr Space (pfx)	Transit Addr space (pfx)	Description
4	AS4134		ORG+TRN Originate: 66535360 /6.01	Transit: 36472256 /6.88	CHINANET-BACKBONE No.31,Jin-rong Street

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	%
7	AS4134	CHINANET-BACKBONE No.31,Jin-rong Street	1077	809	73	341	736	68.34%

Prefix	AS Path	Aggregation Suggestion
58.30.0.0/15	65056 4637 4134	
58.32.0.0/13	65056 4637 4134	
58.40.0.0/15	65056 4637 4134	
58.42.0.0/15	65056 4637 4134	+ Announce - aggregate of 58.42.0.0/16 (65056 4637 4134) and 58.43.0.0/16 (65056 4637 4134)
58.42.0.0/17	65056 4637 4134	- Withdrawn - aggregated with 58.42.128.0/17 (65056 4637 4134)
58.42.128.0/17	65056 4637 4134	- Withdrawn - aggregated with 58.42.0.0/17 (65056 4637 4134)
58.43.0.0/16	65056 4637 4134	- Withdrawn - aggregated with 58.42.0.0/16 (65056 4637 4134)
58.44.0.0/14	65056 4637 4134	
58.48.0.0/13	65056 4637 4134	
58.48.0.0/14	65056 4637 4134	- Withdrawn - matching aggregate 58.48.0.0/13 65056 4637 4134
58.52.0.0/14	65056 4637 4134	- Withdrawn - matching aggregate 58.48.0.0/13 65056 4637 4134
58.56.0.0/15	65056 4637 4134	
58.58.0.0/15	65056 4637 4134	+ Announce - aggregate of 58.58.0.0/16 (65056 4637 4134) and 58.59.0.0/16 (65056 4637 4134)
58.58.0.0/16	65056 4637 4134	- Withdrawn - aggregated with 58.59.0.0/16 (65056 4637 4134)
58.59.0.0/17	65056 4637 4134	- Withdrawn - aggregated with 58.59.128.0/17 (65056 4637 4134)
58.59.128.0/17	65056 4637 4134	- Withdrawn - aggregated with 58.59.0.0/17 (65056 4637 4134)
58.59.128.0/19	65056 4637 4134	- Withdrawn - matching aggregate 58.59.128.0/17 65056 4637 4134
58.59.160.0/19	65056 4637 4134	- Withdrawn - matching aggregate 58.59.128.0/17 65056 4637 4134
58.59.192.0/19	65056 4637 4134	- Withdrawn - matching aggregate 58.59.128.0/17 65056 4637 4134
58.59.224.0/19	65056 4637 4134	- Withdrawn - matching aggregate 58.59.128.0/17 65056 4637 4134
58.60.0.0/14	65056 4637 4134	
58.60.0.0/15	65056 4637 4134	- Withdrawn - matching aggregate 58.60.0.0/14 65056 4637 4134
58.62.0.0/15	65056 4637 4134	- Withdrawn - matching aggregate 58.60.0.0/14 65056 4637 4134
58.66.0.0/15	65056 4637 4134	+ Announce - aggregate of 58.66.0.0/16 (65056 4637 4134) and 58.67.0.0/16 (65056 4637 4134)
58.66.0.0/17	65056 4637 4134	- Withdrawn - aggregated with 58.66.128.0/17 (65056 4637 4134)
58.66.128.0/18	65056 4637 4134	- Withdrawn - aggregated with 58.66.192.0/18 (65056 4637 4134)
58.66.192.0/18	65056 4637 4134	- Withdrawn - aggregated with 58.66.128.0/18 (65056 4637 4134)
58.67.0.0/17	65056 4637 4134	- Withdrawn - aggregated with 58.67.128.0/17 (65056 4637 4134)
58.67.128.0/17	65056 4637 4134	- Withdrawn - aggregated with 58.67.0.0/17 (65056 4637 4134)
58.82.0.0/17	65056 4637 4134	

AS Report

<http://www.cidr-report.org/cgi-bin/as-report?as=AS18566&view=2.0>

Google

AS Report

Announced Prefixes

Rank	AS	Type	Originate Addr Space (pfx)	Transit Addr space (pfx)	Description
145	AS18566	ORIGIN	Originate: 2288128 /10.87	Transit: 0 /0.00	COVAD - Covad Communications Co.

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	AS18566	COVAD - Covad Communications Co.	1025	928	4	101	924	90.15%

Prefix	AS Path	Aggregation Suggestion
64.105.0.0/16	65056 4637 2828 18566	
64.105.0.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.4.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.6.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.8.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.10.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.14.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.16.0/24	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.17.0/24	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.18.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.20.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.22.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.24.0/21	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.32.0/21	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.40.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.42.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.44.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.46.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.48.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.50.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.52.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.54.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.56.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.58.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.60.0/23	65056 4637 3356 18566	
64.105.62.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.64.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.66.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.68.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566
64.105.70.0/23	65056 4637 2828 18566 - Withdrawn	- matching aggregate 64.105.0.0/16 65056 4637 2828 18566

Importance of Aggregation

- Size of routing table

 - Memory is no longer a problem

 - 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

The BGP Instability Report is updated daily. This report was generated on 16 September 2007 07:44 (UTC+1000)

50 Most active ASes for the past 31 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	4538	1087398	2.42%	2447	444.38	ERX-CERNET-BKB China Education and Research Network Center
2	18566	624495	1.39%	1025	609.26	COVAD - Covad Communications Co.
3	6197	524322	1.17%	1037	505.61	BATI-ATL - BellSouth Network Solutions, Inc
4	9498	489154	1.09%	1028	475.83	BBIL-AP BHARTI BT INTERNET LTD.
5	7011	467460	1.04%	942	496.24	FRONTIER-AND-CITIZENS - Frontier Communications of America, Inc.
6	4766	452304	1.01%	820	551.59	KIXS-AS-KR Korea Telecom
7	9583	434283	0.97%	1180	368.04	SIFY-AS-IN Sify Limited
8	22773	427537	0.95%	775	551.66	CCINET-2 - Cox Communications Inc.
9	852	351549	0.78%	606	580.11	ASN852 - Telus Advanced Communications
10	6198	340020	0.76%	586	580.24	BATI-MIA - BellSouth Network Solutions, Inc
11	16852	242967	0.54%	406	598.44	BROADWING-FOCAL - Broadwing Communications Services, Inc.
12	19916	215813	0.48%	568	379.95	ASTRUM-0001 - OLM LLC
13	4323	211624	0.47%	1364	155.15	TWTC - Time Warner Telecom, Inc.
14	174	208930	0.46%	997	209.56	COGENT Cogent/PSI
15	4668	198875	0.44%	517	384.67	LGNET-AS-KR LG CNS
16	2907	193024	0.43%	328	588.49	ERX-SINET-AS National Center for Science Information Systems
17	9929	191544	0.43%	369	519.09	CNCNET-CN China Netcom Corp.
18	5668	176188	0.39%	664	265.34	AS-5668 - CenturyTel Internet Holdings, Inc.
19	855	173387	0.39%	578	299.98	CANET-ASN-4 - Bell Aliant
20	15290	173303	0.39%	289	599.66	ALLST-15290 - Allstream Corp. Corporation Allstream
21	4755	171617	0.38%	1423	120.60	VSNL-AS Videsh Sanchar Nigam Ltd. Autonomous System
22	1785	164939	0.37%	347	475.33	USLEC-ASN-1785 - USLEC Corp.

50 Most active Prefixes for the past 31 days

RANK	PREFIX	UPDs	%	Origin AS -- AS NAME
1	221.135.22.0/24	49747	0.11%	9583 -- SIFY-AS-IN Sify Limited
2	221.135.113.0/24	30527	0.07%	9583 -- SIFY-AS-IN Sify Limited
3	12.108.254.0/24	30379	0.07%	26829 -- YKK-USA - YKK USA,INC
4	62.24.238.0/24	29989	0.07%	13285 -- OPALTELECOM-AS Opal Telecom
5	117.58.192.0/19	29754	0.07%	7491 -- PI-PH-AS-AP PI-PHILIPINES
6	209.163.125.0/24	29432	0.06%	14390 -- CORENET - Coretel America, Inc.
7	80.243.64.0/20	28616	0.06%	21332 -- NTC-AS New Telephone Company
8	202.56.250.0/24	20811	0.05%	9498 -- BBIL-AP BHARTI BT INTERNET LTD.
9	203.101.87.0/24	19365	0.04%	9498 -- BBIL-AP BHARTI BT INTERNET LTD.
10	193.46.60.0/24	18783	0.04%	43403 -- SVIAZ-PLUS-AS LLC "Sviaz Plus"
11	12.106.30.0/24	16055	0.04%	22072 --
12	210.18.10.0/24	15792	0.03%	9583 -- SIFY-AS-IN Sify Limited
13	208.46.32.0/24	15102	0.03%	27289 -- CLEOCOMMUNICATIONS - CLEO COMMUNICATIONS INC
14	221.135.253.0/24	14243	0.03%	9583 -- SIFY-AS-IN Sify Limited
15	221.135.80.0/24	13994	0.03%	9583 -- SIFY-AS-IN Sify Limited
16	64.95.193.0/24	12307	0.03%	30707 --
17	210.214.173.0/24	12142	0.03%	9583 -- SIFY-AS-IN Sify Limited
18	210.214.220.0/24	12101	0.03%	9583 -- SIFY-AS-IN Sify Limited
19	210.214.177.0/24	12045	0.03%	9583 -- SIFY-AS-IN Sify Limited
20	210.214.172.0/24	12041	0.03%	9583 -- SIFY-AS-IN Sify Limited
21	210.214.221.0/24	12021	0.03%	9583 -- SIFY-AS-IN Sify Limited
22	210.214.210.0/24	11962	0.03%	9583 -- SIFY-AS-IN Sify Limited
23	210.214.211.0/24	11942	0.03%	9583 -- SIFY-AS-IN Sify Limited
24	89.4.131.0/24	11863	0.03%	24731 -- ASN-NESMA National Engineering Services and Marketing Company Ltd. (NESMA)
25	221.135.77.0/24	11794	0.03%	9583 -- SIFY-AS-IN Sify Limited
26	89.4.130.0/24	11754	0.03%	24731 -- ASN-NESMA National Engineering Services and Marketing Company Ltd. (NESMA)
27	203.63.26.0/24	10633	0.02%	9747 -- EZINTERNET-AS-AP EZInternet Pty Ltd



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** - look the address up!!

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/

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 when multihoming

- 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 RFC1918 etc prefixes

<ftp://ftp.rfc-editor.org/in-notes/rfc3330.txt>

don't accept your own prefixes

don't accept default (unless you need it)

don't accept prefixes longer than /24

- Check Project Cymru's list of "bogons"

<http://www.cymru.com/Documents/bogon-list.html>

<http://www.cymru.com/BGP/bogon-rs.html>

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



Configuration Tips

Of passwords, tricks and templates

iBGP and IGP 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
 - The Internet is around 5 ASes deep on average
 - Largest AS_PATH is usually 16-20 ASNs

Limiting AS Path Length

- Some announcements have ridiculous lengths of AS-paths:

```
*> 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

```
*> 194.146.180.0/22        2497 3257 29686 16327 16327  
16327 16327 16327 16327 16327 16327 16327 16327  
16327 16327 16327 16327 16327 16327 16327 16327  
16327 16327 16327 i
```

This example shows 20 prepends (for no obvious reason)

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

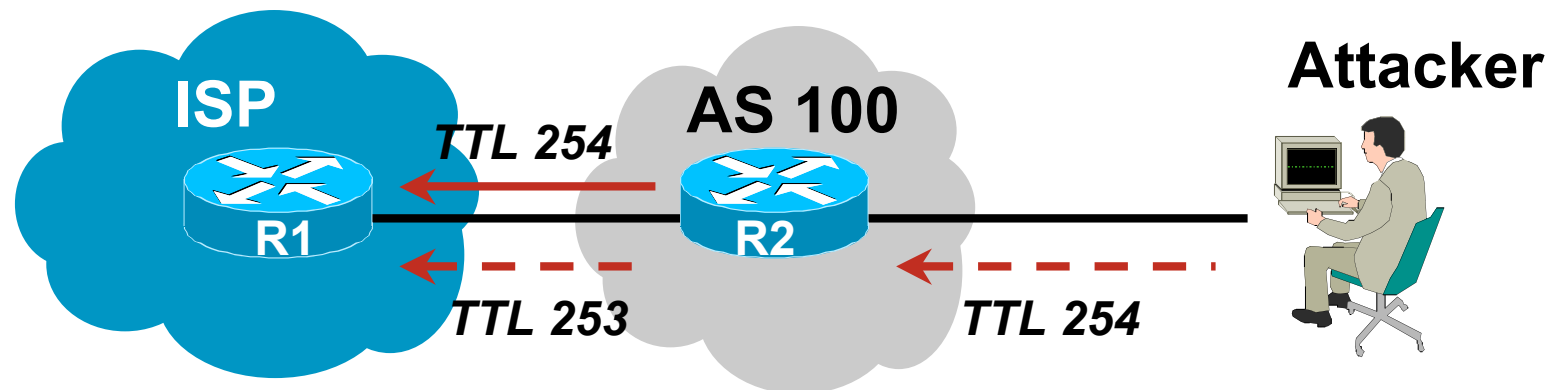
BGP TTL “hack”

- Implement RFC3682 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



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 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
www.cymru.com/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
 - 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

- 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
- 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

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!
- 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