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

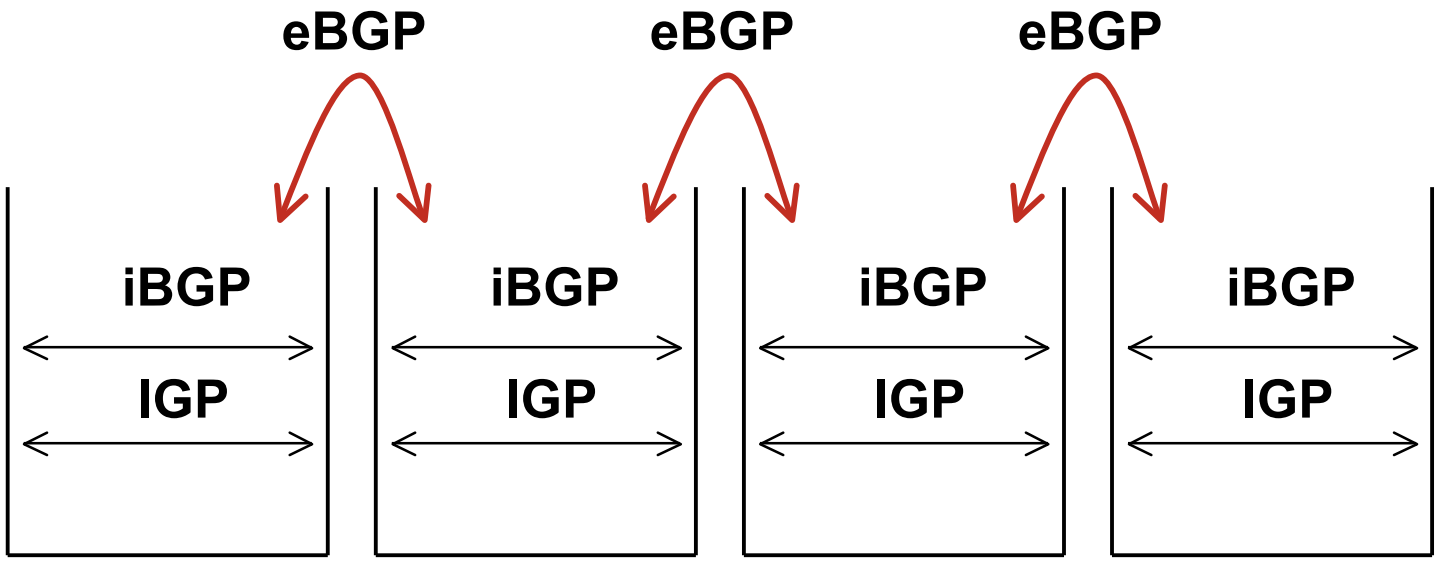
- Internal Routing Protocols (IGPs)
  - Examples are ISIS and OSPF
  - Used for carrying infrastructure addresses
  - **NOT** used for carrying Internet prefixes or customer prefixes
  - ISP design goal is to minimise number of prefixes in IGP to aid scalability and rapid convergence

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

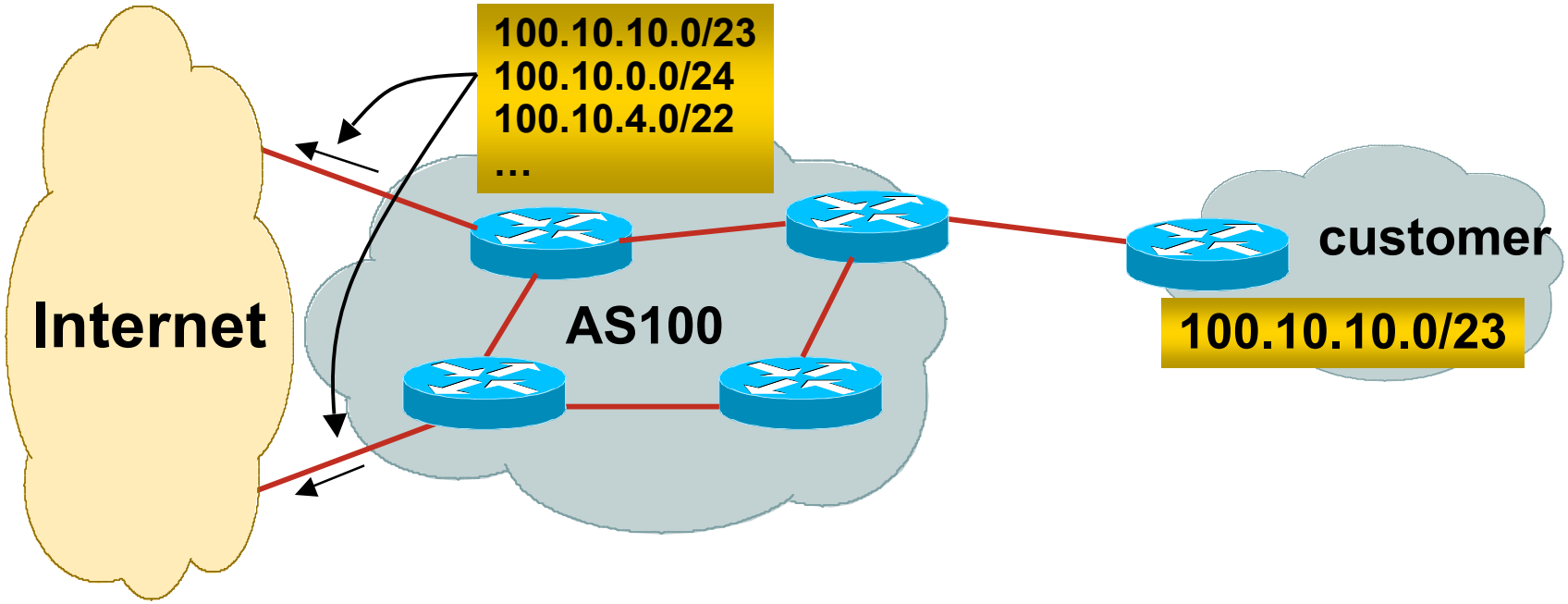
- Address block should be announced to the Internet as an aggregate
- Subprefixes of address block should NOT be announced to Internet unless traffic engineering when multihoming
- Aggregate should be generated internally

  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

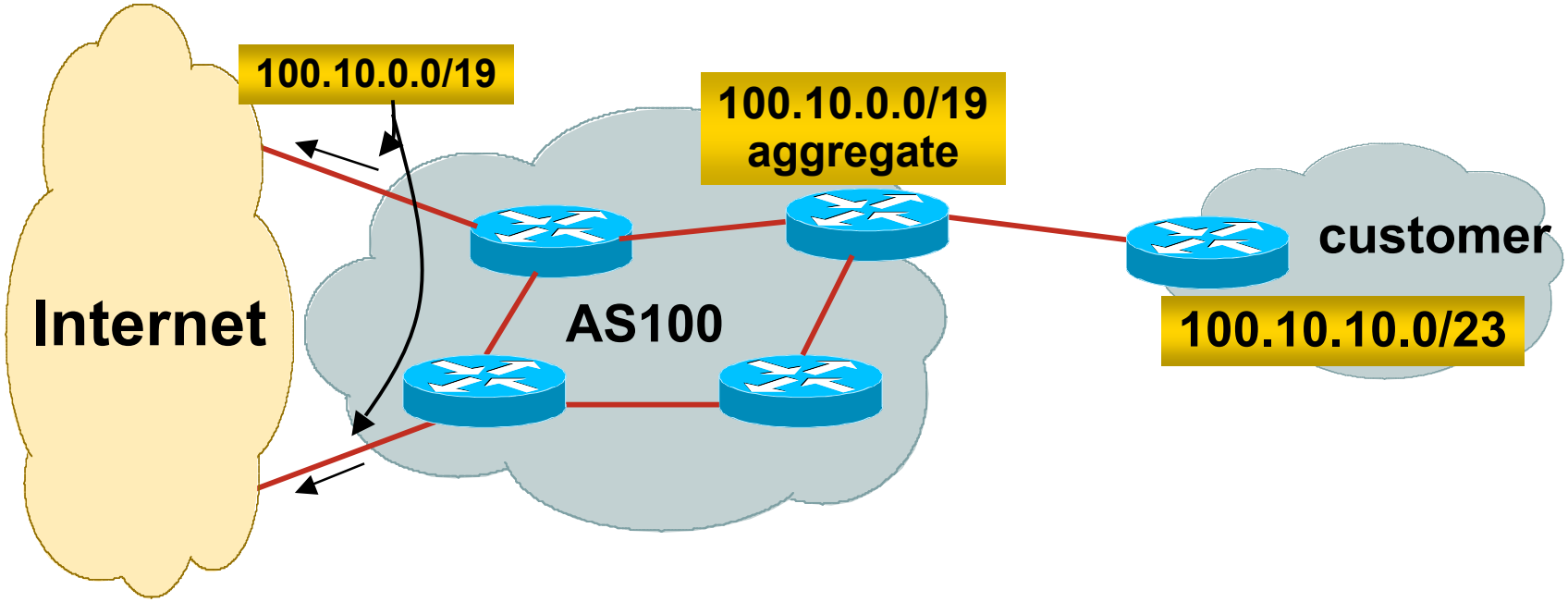


- Customer has /23 network assigned from AS100's /19 address block
- AS100 announces customers' individual networks to the Internet

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???

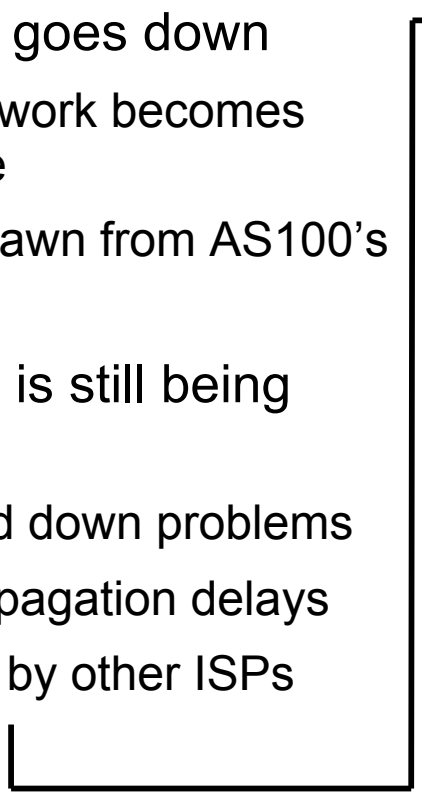
Aggregation – Example



- 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

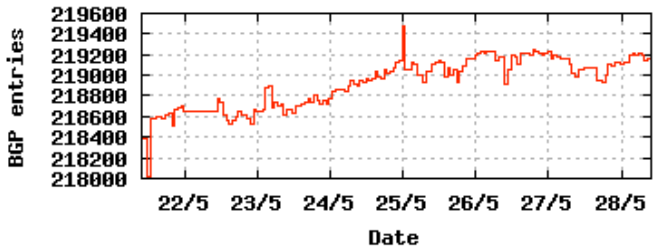
Efforts to Improve Aggregation
The CIDR Report

- 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

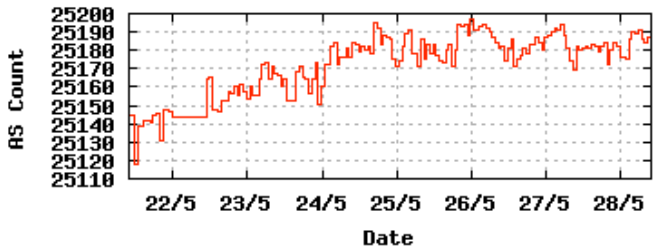
Date	Prefixes	CIDR Aggregated
21-05-07	218385	140025
22-05-07	218650	139831
23-05-07	218653	139850
24-05-07	218776	139698
25-05-07	219469	139898
26-05-07	219203	139943
27-05-07	219232	139870
28-05-07	219115	140020



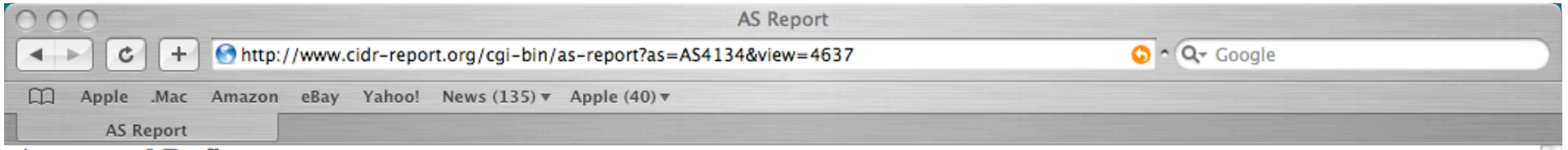
Plot: BGP Table Size

AS Summary

25190	Number of ASes in routing system
10666	Number of ASes announcing only one prefix
1483	Largest number of prefixes announced by an AS AS7018: ATT-INTERNET4 - AT&T WorldNet Services
89890048	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)



Announced Prefixes

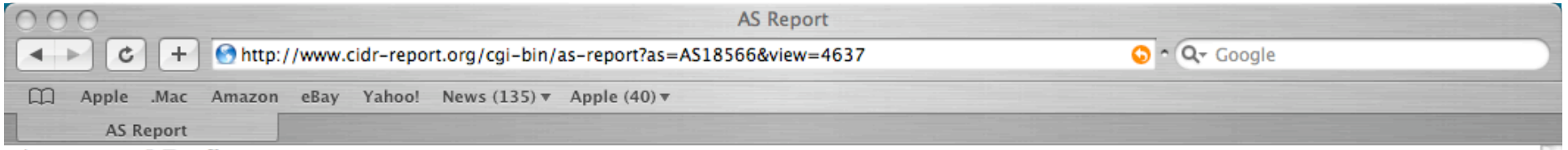
```
Rank  AS       Type    Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
4     AS4134   ORG+TRN Originate:   56476672 /6.25   Transit:   30243328 /7.15  CHINANET-BACKBONE No.31,Jin-rong Street
```

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       %
    4 AS4134  CHINANET-BACKBONE No.31,Jin-rong Street    1257   1005     64    316    941  74.86%


AS 4134: CHINANET-BACKBONE No.31,Jin-rong Street
  Prefix  (AS Path)                    Aggregation Action
58.30.0.0/15          4608 1221 4637 4134
58.32.0.0/13          4608 1221 4637 4134
58.40.0.0/15          4608 1221 4637 4134
58.42.0.0/15          4608 1221 4637 4134  + Announce - aggregate of 58.42.0.0/16 (4608 1221 4637 4134) and 58.43.0.0/16 (4608
58.42.0.0/17          4608 1221 4637 4134  - Withdrawn - aggregated with 58.42.128.0/17 (4608 1221 4637 4134)
58.42.128.0/17        4608 1221 4637 4134  - Withdrawn - aggregated with 58.42.0.0/17 (4608 1221 4637 4134)
58.43.0.0/16          4608 1221 4637 4134  - Withdrawn - aggregated with 58.42.0.0/16 (4608 1221 4637 4134)
58.44.0.0/14          4608 1221 4637 4134
58.48.0.0/13          4608 1221 4637 4134
58.48.0.0/14          4608 1221 4637 4134  - Withdrawn - matching aggregate 58.48.0.0/13 4608 1221 4637 4134
58.52.0.0/14          4608 1221 4637 4134  - Withdrawn - matching aggregate 58.48.0.0/13 4608 1221 4637 4134
58.56.0.0/15          4608 1221 4637 4134
58.58.0.0/15          4608 1221 4637 4134  + Announce - aggregate of 58.58.0.0/16 (4608 1221 4637 4134) and 58.59.0.0/16 (4608
58.58.0.0/16          4608 1221 4637 4134  - Withdrawn - aggregated with 58.59.0.0/16 (4608 1221 4637 4134)
58.59.0.0/17          4608 1221 4637 4134  - Withdrawn - aggregated with 58.59.128.0/17 (4608 1221 4637 4134)
58.59.128.0/17        4608 1221 4637 4134  - Withdrawn - aggregated with 58.59.0.0/17 (4608 1221 4637 4134)
58.59.128.0/19        4608 1221 4637 4134  - Withdrawn - matching aggregate 58.59.128.0/17 4608 1221 4637 4134
58.59.160.0/19        4608 1221 4637 4134  - Withdrawn - matching aggregate 58.59.128.0/17 4608 1221 4637 4134
58.59.192.0/19        4608 1221 4637 4134  - Withdrawn - matching aggregate 58.59.128.0/17 4608 1221 4637 4134
58.59.224.0/19        4608 1221 4637 4134  - Withdrawn - matching aggregate 58.59.128.0/17 4608 1221 4637 4134
58.60.0.0/14          4608 1221 4637 4134
58.60.0.0/15          4608 1221 4637 4134  - Withdrawn - matching aggregate 58.60.0.0/14 4608 1221 4637 4134
58.62.0.0/15          4608 1221 4637 4134  - Withdrawn - matching aggregate 58.60.0.0/14 4608 1221 4637 4134
58.66.0.0/17          4608 1221 4637 4134
58.66.128.0/18        4608 1221 4637 4134
58.67.0.0/17          4608 1221 4637 4134
58.82.0.0/17          4608 1221 4637 4134
58.82.192.0/19        4608 1221 4637 4134
```



Announced Prefixes

Rank	AS	Type	Originate Addr Space	(pfx)	Transit Addr space	(pfx)	Description
144	AS18566	ORIGIN	Originate: 2268160	/10.89	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	%
2	AS18566	COVAD - Covad Communications Co.	1010	979	0	31	979	96.93%

AS18566: COVAD - Covad Communications Co.

Prefix	(AS Path)	Aggregation Action
64.105.0.0/16	4608 1221 4637 3356 18566	
64.105.0.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.4.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.6.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.8.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.10.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.14.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.16.0/24	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.17.0/24	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.18.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.20.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.22.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.24.0/21	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.32.0/21	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.40.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.42.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.44.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.46.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.48.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.50.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.52.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.54.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.56.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.58.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.60.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.62.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.64.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.66.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566
64.105.68.0/23	4608 1221 4637 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4608 1221 4637 3356 18566

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**

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 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
  - 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

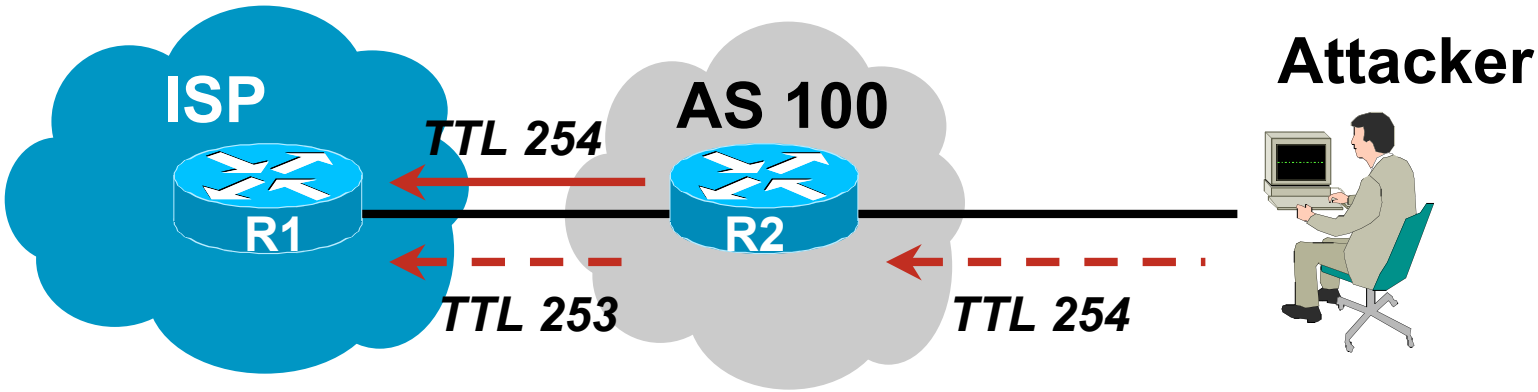
```
*> 194.146.180.0/22        2497 3257 29686 16327 16327
16327 16327 16327 16327 16327 16327 16327 16327
16327 16327 16327 16327 16327 16327 16327 16327
16327 16327 16327 i
```

  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 Project 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