BGP Best Current Practices

Philip Smith

NSRC

SAFNOG 1

22nd – 23rd April 2014

Johannesburg

Last updated 15 April 2014

Presentation Slides

- Will be available on
  - http://thyme.apnic.net/ftp/seminars/SAFNOG1-BGP-BCP.pdf
  - And on the SAFNOG website
- Feel free to ask questions any time

What is BGP for??

What is an IGP not for?

BGP versus OSPF/ISIS

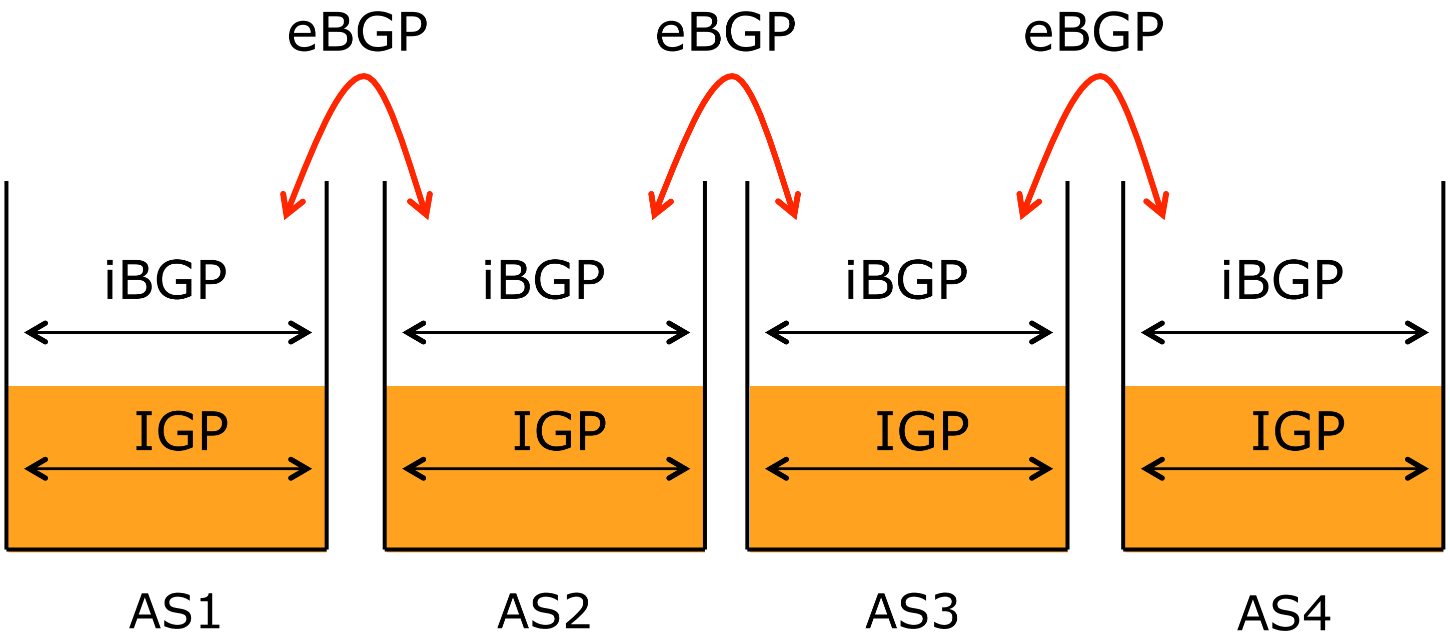
- Internal Routing Protocols (IGPs)
  - examples are ISIS and OSPF
  - used for carrying **infrastructure** addresses
  - **NOT** used for carrying Internet prefixes or customer prefixes
  - design goal is to **minimise** number of prefixes in IGP to aid scalability and rapid convergence

BGP versus OSPF/ISIS

- BGP 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**

BGP Scaling Techniques

- Route Refresh
  - To implement BGP policy changes without hard resetting the BGP peering session
- Route Reflectors
  - Scaling the iBGP mesh
  - A few iBGP speakers can be fully meshed
  - Large networks have redundant per-PoP route-reflectors

BGP Communities

- Another ISP “scaling technique”
- Prefixes are grouped into different “classes” or communities within the ISP network
- Each community can represent a different policy, has a different result in the ISP network
- ISP defined communities can be made available to customers
    - Allows them to manipulate BGP policies as applied to their originated prefixes

Aggregation

Aggregation

- Aggregation means announcing the address block received from the RIR to the other ASes connected to your network
- Subprefixes of this aggregate may be:
  - Used internally in the ISP network
  - Announced to other ASes to aid with multihoming
- Unfortunately too many people are still thinking about class Cs, resulting in a proliferation of /24s in the Internet routing table
  - Apr 2014: 261000 /24s in IPv4 table of 492000 prefixes
- The same is happening for /48s with IPv6
  - Apr 2014: 7200 /48s in IPv6 table of 16700 prefixes

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

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 (April 2014)

- Current Internet Routing Table Statistics

BGP Routing Table Entries	491472
Prefixes after maximum aggregation	193050
Unique prefixes in Internet	242559
Prefixes smaller than registry alloc	171311
/24s announced	261411
ASes in use	46602

Efforts to improve aggregation

- The CIDR Report
  - Initiated and operated for many years by Tony Bates
  - Now combined with Geoff Huston’s routing analysis
    - www.cidr-report.org
    - (covers both IPv4 and IPv6 BGP tables)
  - Results e-mailed on a weekly basis to most operations lists around the world
  - Lists the top 30 service providers who could do better at aggregating
- RIPE Routing WG aggregation recommendations
  - IPv4: RIPE-399 — www.ripe.net/ripe/docs/ripe-399.html
  - IPv6: RIPE-532 — www.ripe.net/ripe/docs/ripe-532.html

Receiving Prefixes

Receiving Prefixes

- There are three scenarios for receiving prefixes from other ASNs
  - Customer talking BGP
  - Peer talking BGP
  - Upstream/Transit talking BGP
- Each has different filtering requirements and need to be considered separately

Receiving Prefixes: From Customers

- ISPs should only accept prefixes which have been assigned or allocated to their downstream customer
- If ISP has assigned address space to its customer, then the customer IS entitled to announce it back to his ISP
- If the ISP has NOT assigned address space to its customer, then:
  - Check in the five RIR databases to see if this address space really has been assigned to the customer
  - The tool: whois –h jwhois.apnic.net x.x.x.0/24
    - (jwhois queries all RIR databases)

Receiving Prefixes: From 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 – see BGP Multihoming presentations
- Ask upstream/transit provider to either:
  - originate a default-route

    OR
  - announce one prefix you can use as default

Receiving Prefixes: From Upstream/Transit Provider

- If necessary to receive prefixes from any provider, care is required.
  - Don’t accept default (unless you need it)
  - Don’t accept your own prefixes
- Special uses prefixes for IPv4 and IPv6:
  - http://www.rfc-editor.org/rfc/rfc6890.txt
- For IPv4:
  - Don’t accept prefixes longer than /24 (?)
    - /24 was the historical class C
- For IPv6:
  - Don’t accept prefixes longer than /48 (?)
    - /48 is the ‘minimum block delegated to a site’

Receiving Prefixes: From Upstream/Transit Provider

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

Receiving 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 IPv4 /32 and IPv6 /128 address
- Consider the DMZ nets:
  - Use unnumbered interfaces?
  - Use next-hop-self on iBGP neighbours
  - Or carry the DMZ IPv4 /30s and IPv6 /127s in the iBGP
  - Basically keep the DMZ nets out of the IGP!

iBGP: Next-hop-self

- BGP speaker announces external network to iBGP peers using router’s local address (loopback) as next-hop
- Used by many ISPs on edge routers
  - Preferable to carrying DMZ point-to-point link addresses in the IGP
  - Reduces size of IGP to just core infrastructure
  - Alternative to using unnumbered interfaces
  - Helps scale network
  - Many ISPs consider this “best practice”

Limiting AS Path Length

- Some BGP implementations have problems with long AS_PATHS
  - Memory corruption
  - Memory fragmentation
- Even using AS_PATH prepends, it is not normal to see more than 20 ASes in a typical AS_PATH in the Internet today
  - The Internet is around 5 ASes deep on average
  - Largest AS_PATH is usually 16-20 ASNs

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

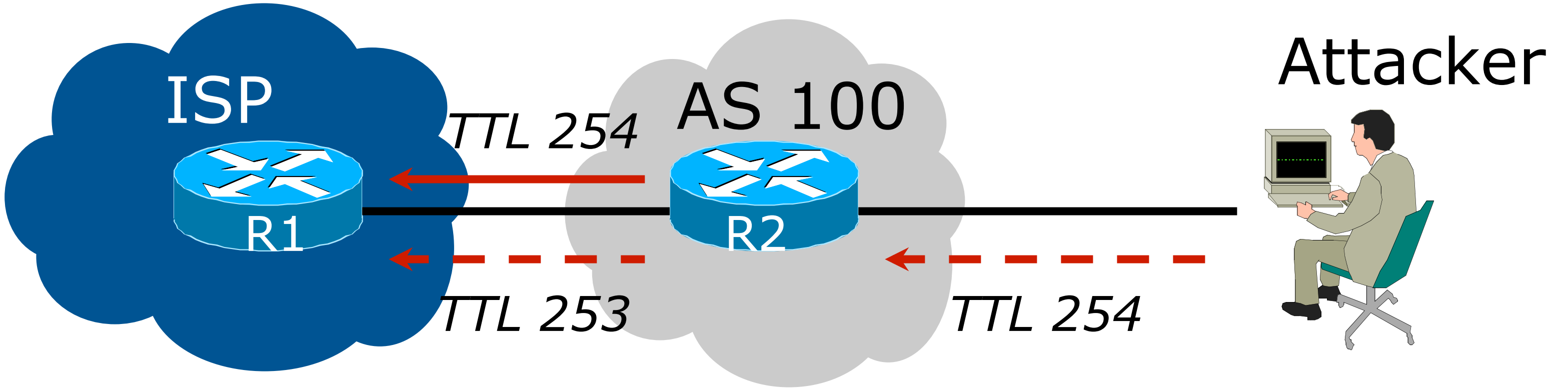
```
*>i193.105.15.0       2516 3257 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 50404 50404 50404 50404
   50404 50404 50404 50404 50404 50404 50404 i
```

This example shows 100 prepends (for no obvious reason)

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

BGP TTL "hack"

- Implement RFC5082 on BGP peerings
  - (Generalised TTL Security Mechanism)
  - Neighbour sets TTL to 255
  - Local router expects TTL of incoming BGP packets to be 254
  - No one apart from directly attached devices can send BGP packets which arrive with TTL of 254, so any possible attack by a remote miscreant is dropped due to TTL mismatch



BGP TTL "hack"

- TTL Hack:
  - Both neighbours must agree to use the feature
  - TTL check is much easier to perform than MD5
  - (Called BTSH – BGP TTL Security Hack)
- Provides "security" for BGP sessions
  - In addition to packet filters of course
  - MD5 should still be used for messages which slip through the TTL hack
  - See 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
  - 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

BGP Best Current Practices

The End