BGP Best Practices

Philip Smith <pfs@cisco.com>

RIPE NCC Regional Meeting

Manama, Bahrain

14-15 November 2006

Presentation Slides

- **Are available on**

  ftp://ftp-eng.cisco.com

  /pfs/seminars/RIPENCC-Bahrain-BGP-BCP.pdf

  And on the RIPE NCC Bahrain meeting website

BGP Best Practices

How to use BGP on the Internet

BGP versus OSPF/ISIS

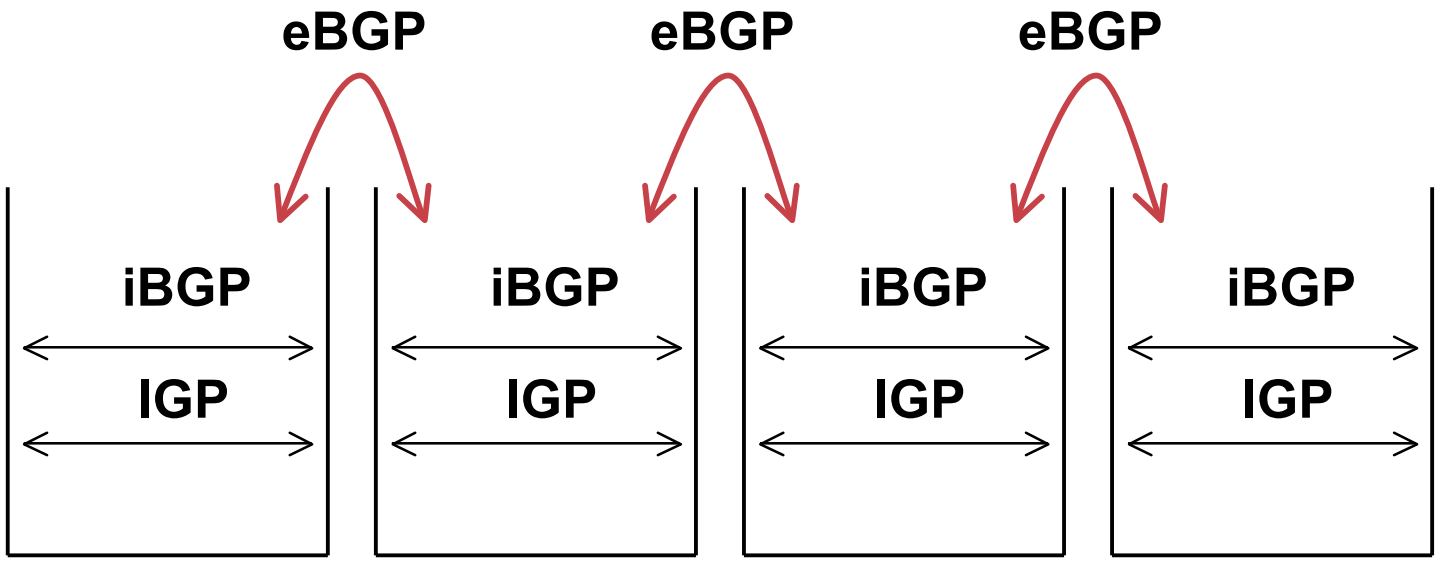
- **Separation of IGP and BGP**
- **Internal Routing Protocols (IGPs)**
  - **Examples are ISIS and OSPF**
  - **Used for carrying infrastructure addresses — infrastructure reachability**
  - **NOT used for carrying Internet prefixes or customer prefixes**
  - **Design goal is to minimise number of prefixes in IGP to aid scalability and speed convergence**

eBGP & iBGP

- BGP used internally (iBGP) and externally (eBGP)
- iBGP used to carry
  - some/all Internet prefixes across ISP backbone
  - ISP's customer prefixes
  - BGP session is run between router loopback interfaces
- eBGP used to
  - exchange prefixes with other ASes
  - implement routing policy
  - BGP session is run on inter-AS point to point links

BGP/IGP model used in ISP networks

- **Model representation**



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

- **Route Flap Damping**

  Is NOT a scaling technique and is now considered HARMFUL

  www.ripe.net/ripe/docs/ripe-378.html

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 means announcing the address block received from the Regional Internet Registry to the other ASes connected to your network**
  - **Aggregate should be generated internally, not on network borders**
- **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**

Announcing an Aggregate

- ISPs who don't and won't aggregate are held in poor regard by community
- The RIRs publish their minimum allocation size
  - Anything from a /20 to a /22 depending on RIR
- No real reason to see anything longer than a /22 prefix in the Internet
  - BUT there are currently >108000 /24s!

The Internet Today (November 2006)

- **Current Internet Routing Table Statistics**

  From my Routing Report: http://thyme.apnic.net

BGP Routing Table Entries	**202457**
Prefixes after maximum aggregation	**109985**
Unique prefixes in Internet	**98204**
Prefixes smaller than registry alloc	**102061**
/24s announced	**108212**
only 5754 /24s are from 192.0.0.0/8	
ASes in use	**23532**

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**
    - **http://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**
  - **Website allows searches and computations of aggregation to be made on a per AS basis**

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 whois.apnic.net x.x.x.0/24

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

- 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 required for multihoming/traffic engineering**
- **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)
- Check Rob Thomas' list of "bogons"
  - http://www.cymru.com/Documents/bogon-list.html
- Or get a BGP feed from the Bogon Route Server
  - http://www.cymru.com/BGP/bogon-rs.html

Configuration Tips

Of templates, passwords, tricks, and more 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
- Keep IGP routing table small
- 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!

Next-hop-self

- **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**
  - **BGP speaker announces external network using local address (loopback) as next-hop**

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**
    - **http://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!
- Use passwords on iBGP session
  - Not being paranoid, VERY necessary

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**

(cont…)

eBGP Template
Example continued

- **Use password agreed between you and your peer on eBGP session**
- **Use intelligent maximum-prefix tracking**

  Router will warn you if there are sudden increases in BGP table size, bringing down eBGP if desired
- **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!!

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

BGP TTL “hack”

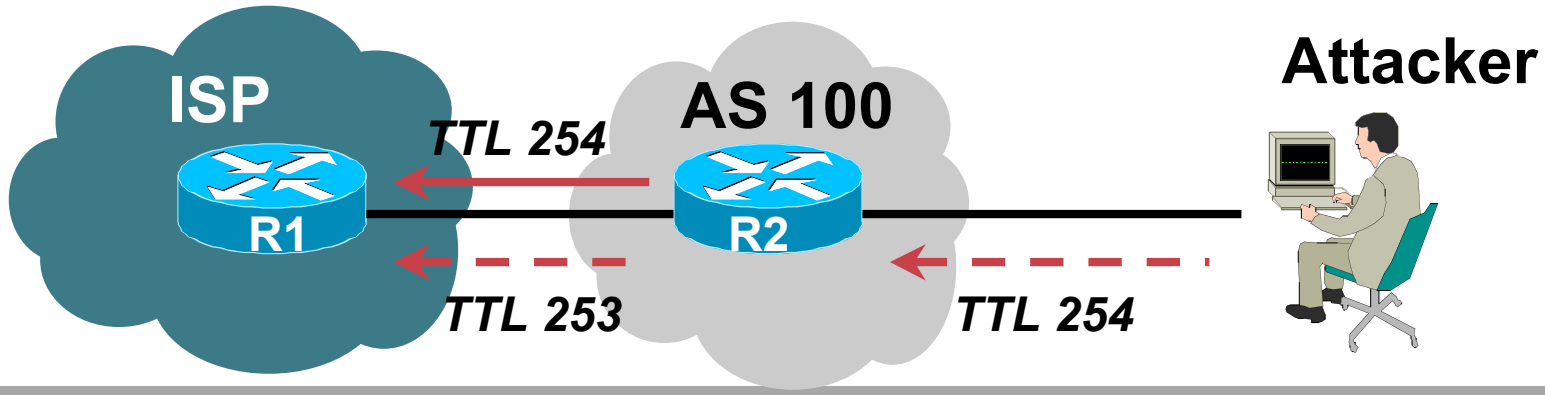
- **Implement RFC3682 on BGP peerings**

  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

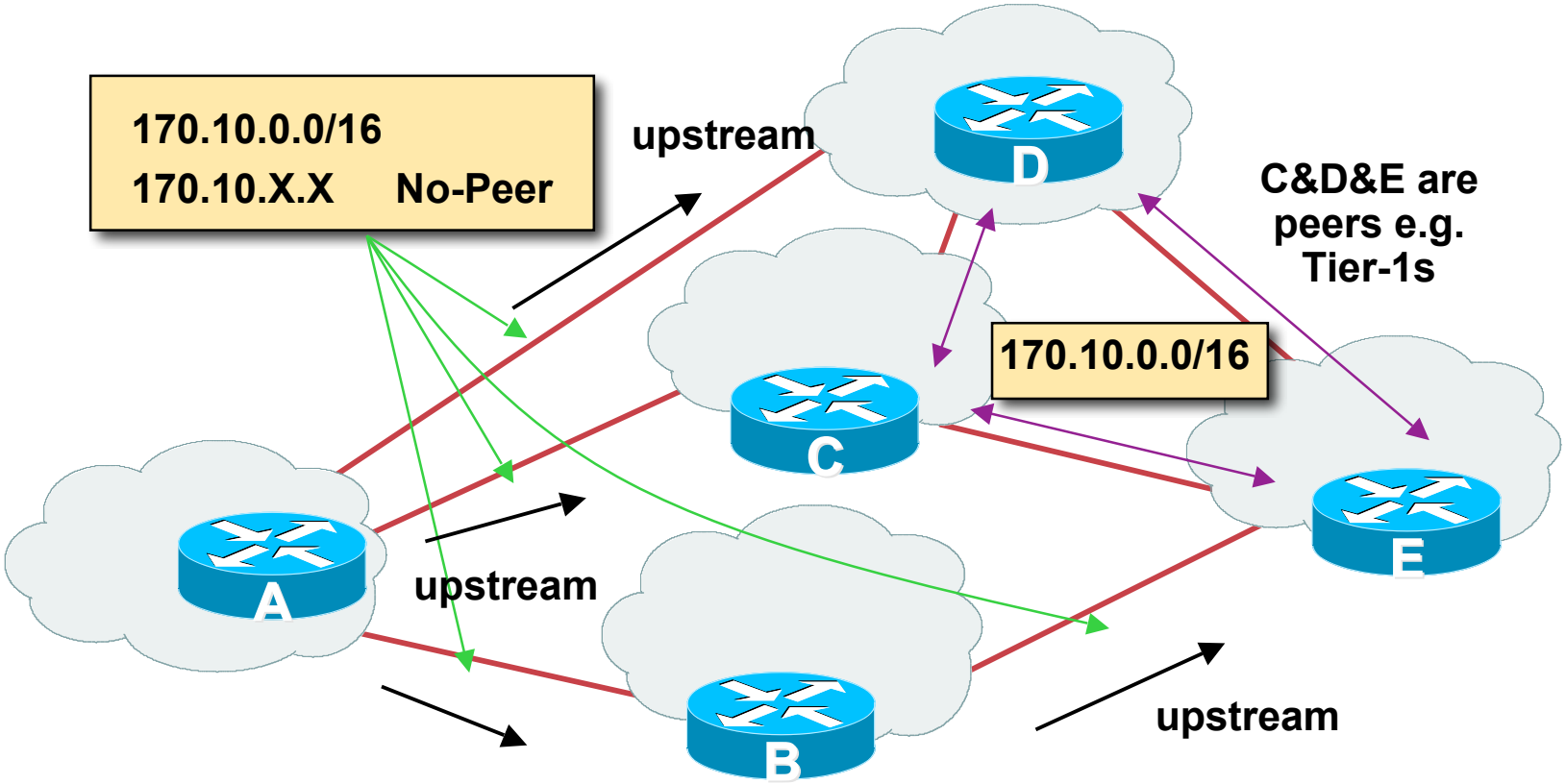
  See http://www.nanog.org/mtg-0302/hack.html for more details



BGP Futures

What is around the corner…?

No-Peer Community



- **Sub-prefixes marked with no-peer community are not sent to bi-lateral peers**

  They are only sent to upstream providers

32-bit Autonomous System Number (ASN)

- **32 bit ASNs are coming soon**
  - **16 bit ASN space is running out — will be exhausted by October 2010**
  - **Represented as "65.4321" — i.e. two 16-bit integers**
  - **With AS 23456 reserved for the transition**

www.ietf.org/internet-drafts/draft-ietf-idr-as4bytes-12.txt

www.ietf.org/internet-drafts/draft-michaelson-4byte-as-representation-02.txt

www.ietf.org/internet-drafts/draft-rekhter-as4octet-ext-community-01.txt

www.apnic.net/docs/policy/proposals/prop-032-v002.html

Concern 1: De-aggregation

- **RIR space shows creeping deaggregation**
  - **It seems that an RIR /8 block averages around 6000 prefixes once fully allocated**
  - **So their existing 74 /8s will eventually cause 444000 prefix announcements**
- **Food for thought:**
  - **Remaining 59 unallocated /8s and the 74 RIR /8s combined will cause:**
  - **798000 prefixes with 6000 prefixes per /8 density**
  - **Plus 12% due to “non RIR space deaggregation”**
  - **→ Routing Table size of 893760 prefixes**

Concern 2: BGP Updates

- BGP Flapping was the “bad guy” of the mid-90s
- BGP Updates is the “bad guy” of today & tomorrow
  - Work by Geoff Huston: bgpupdates.potaroo.net
- 10 providers cause 10% of all the BGP updates on the Internet today
  - All causing more than 2600 updates per day
  - (Connexion by Boeing produces 1450 updates per day)
  - Seeing total of 700k updates per day
  - In 5 years time this will be 2.8M updates per day
- What will this mean for the routers??

BGP Best Practices

Philip Smith <pfs@cisco.com>

RIPE NCC Regional Meeting

Manama, Bahrain

14-15 November 2006