BGP Best Current Practices

ISP/IXP Workshops
Recommended IOS Releases

Which IOS??
Which IOS?

- IOS is a feature rich and highly complex router control system
- ISPs should choose the IOS variant which is most appropriate for the intended application
- There is an exclusive service provider train in IOS
  This is 12.0S, supporting 7200, 7500, 10000 and 12000
  Images also available for 2500, 2600, 3600 and 4500, but are completely unsupported
- There is a service provider image in most IOS releases
  This is the image with –p– in its name, for example:
  c7200-p-mz.122-8.T1 and c2600-p-mz.121-14
  The –p– image is IP-only plus ISIS/CLNS
Which IOS?

- **12.n** – for example 12.2
  
  This means the IOS is a mainline image
  NO new features
  ONLY bug fixes
  The aim is stability!

- **12.nT** – for example 12.2T
  
  This means the IOS is the technology release
  NEW features
  Bug fixes
  Avoid unless you need the feature!
12.3 IOS release images

- 12.3 is the old “mainline” train
  Originated from 12.2T, currently at 12.3(15)
  Bug fix release only – aiming for stability
  Supports more platforms and has more features than 12.2 or 12.1

- 12.3T was the old “technology train”
  new features introduced in IOS 12.3
  Last release was 12.3(14)T2

- Available on CCO, supported by TAC
12.4 IOS release images

• 12.4 is the current “mainline” train
  Originated from 12.3T, currently at 12.4(1a)
  Bug fix release only – aiming for stability
  Supports more platforms and has more features than 12.3

• 12.4T is the current “technology train”
  new features introduced in IOS 12.4
  As yet there is no 12.4T release

• Available on CCO, supported by TAC
IOS images for ISPs

- **12.0S** is the release for all ISPs
  
  For 7200, 7500, 10000 and GSR/12000 only
  
  Currently at 12.0(31)S

- **12.2S** is a new ISP release
  
  For 7x00 series (x = 2, 3, 5 and 6)
  
  Combines 12.0S and 12.1E enhancements
  
  Currently at 12.2(25)S4

- **Available on CCO, supported by TAC**
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
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 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**
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
Configuring Aggregation – Cisco IOS

- ISP has 101.10.0.0/19 address block
- To put into BGP as an aggregate:
  
  ```
  router bgp 100
  network 101.10.0.0 mask 255.255.224.0
  ip route 101.10.0.0 255.255.224.0 null0
  ```

- The static route is a “pull up” route
  more specific prefixes within this address block ensure connectivity to ISP’s customers
  “longest match lookup”
Aggregation

• Address block should be announced to the Internet as an aggregate

• Subprefixes of address block should NOT be announced to Internet unless special circumstances (more later)

• Aggregate should be generated internally

Not on the network borders!
Announcing Aggregate – Cisco IOS

• Configuration Example
  
  router bgp 100

  network 101.10.0.0 mask 255.255.224.0

  neighbor 102.102.10.1 remote-as 101

  neighbor 102.102.10.1 prefix-list out-filter out

  !

  ip route 101.10.0.0 255.255.224.0 null0

  !

  ip prefix-list out-filter permit 101.10.0.0/19

  ip prefix-list out-filter deny 0.0.0.0/0 le 32
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 >90000 /24s!
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 – 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 – 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?
The Internet Today (June 2005)

• Current Internet Routing Table Statistics

  BGP Routing Table Entries          165559
  Prefixes after maximum aggregation  95442
  Unique prefixes in Internet         79379
  Prefixes smaller than registry alloc 78267
  /24s announced                       90447
    only 5717 /24s are from 192.0.0.0/8
  ASes in use                          19872
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
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 four 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 Customers

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

  ```
pfs-pc$ whois -h whois.apnic.net 202.12.29.0
  netname: APNIC-AP-AU-BNE
  descr: APNIC Pty Ltd - Brisbane Offices + Servers
  descr: Level 1, 33 Park Rd
  descr: PO Box 2131, Milton
  descr: Brisbane, QLD.
  country: AU
  admin-c: HM20-AP
  tech-c: NO4-AP
  mnt-by: APNIC-HM
  changed: hm-changed@apnic.net 20030108
  status: ASSIGNED PORTABLE
  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:

```bash
$ whois -h whois.ripe.net 193.128.2.0
inetnum: 193.128.2.0 - 193.128.2.15
descr: Wood Mackenzie
country: GB
admin-c: DB635-RIPE
tech-c: DB635-RIPE
status: ASSIGNED PA
mnt-by: AS1849-MNT
changed: davids@uk.uu.net 20020211
source: RIPE

route: 193.128.0.0/14
descr: PIPEX-BLOCK1
origin: AS1849
notify: routing@uk.uu.net
mnt-by: AS1849-MNT
changed: beny@uk.uu.net 20020321
source: RIPE
```

**ASSIGNED PA** – means that it is Provider Aggregatable address space and can only be used for connecting to the ISP who assigned it.
Receiving Prefixes from customer: Cisco IOS

• For Example:
  downstream has 100.50.0.0/20 block
  should only announce this to upstreams
  upstreams should only accept this from them

• Configuration on upstream
  router bgp 100
  neighbor 102.102.10.1 remote-as 101
  neighbor 102.102.10.1 prefix-list customer in
  !
  ip prefix-list customer permit 100.50.0.0/20
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 peer: Cisco IOS

• For Example:
  peer has 220.50.0.0/16, 61.237.64.0/18 and 81.250.128.0/17 address blocks

• Configuration on local router
  router bgp 100
    neighbor 102.102.10.1 remote-as 101
    neighbor 102.102.10.1 prefix-list my-peer in
  !
  ip prefix-list my-peer permit 220.50.0.0/16
  ip prefix-list my-peer permit 61.237.64.0/18
  ip prefix-list my-peer permit 81.250.128.0/17
  ip prefix-list my-peer deny 0.0.0.0/0 le 32
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
  - special circumstances – see later
- 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

• Downstream Router Configuration

  router bgp 100

  network 101.10.0.0 mask 255.255.224.0

  neighbor 101.5.7.1 remote-as 101

  neighbor 101.5.7.1 prefix-list infilt in

  neighbor 101.5.7.1 prefix-list outfilt out

  !

  ip prefix-list infilt permit 0.0.0.0/0

  !

  ip prefix-list outfilt permit 101.10.0.0/19
Receiving Prefixes:
From Upstream/Transit Provider

• Upstream Router Configuration

```
router bgp 101
neighbor 101.5.7.2 remote-as 100
neighbor 101.5.7.2 default-originate
neighbor 101.5.7.2 prefix-list cust-in in
neighbor 101.5.7.2 prefix-list cust-out out
!
ip prefix-list cust-in permit 101.10.0.0/19
!
ip prefix-list cust-out permit 0.0.0.0/0
```
Receiving Prefixes: From Upstream/Transit Provider

- If necessary to receive prefixes from any provider, care is required
  - don’t accept RFC1918 etc prefixes
  - 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”
Receiving Prefixes

router bgp 100
  network 101.10.0.0 mask 255.255.224.0
  neighbor 101.5.7.1 remote-as 101
  neighbor 101.5.7.1 prefix-list in-filter in

  ip prefix-list in-filter deny 0.0.0.0/0           ! Block default
  ip prefix-list in-filter deny 0.0.0.0/8 le 32
  ip prefix-list in-filter deny 10.0.0.0/8 le 32
  ip prefix-list in-filter deny 101.10.0.0/19 le 32 ! Block local prefix
  ip prefix-list in-filter deny 127.0.0.0/8 le 32
  ip prefix-list in-filter deny 169.254.0.0/16 le 32
  ip prefix-list in-filter deny 172.16.0.0/12 le 32
  ip prefix-list in-filter deny 192.0.2.0/24 le 32
  ip prefix-list in-filter deny 192.168.0.0/16 le 32
  ip prefix-list in-filter deny 224.0.0.0/3 le 32   ! Block multicast
  ip prefix-list in-filter deny 0.0.0.0/0 ge 25     ! Block prefixes >/24
  ip prefix-list in-filter permit 0.0.0.0/0 le 32
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
Prefixes into iBGP
Injecting prefixes into iBGP

- Use iBGP to carry customer prefixes
  don’t use IGP
- Point static route to customer interface
- Use BGP network statement
- As long as static route exists (interface active), prefix will be in BGP
Router Configuration: network statement

• Example:

interface loopback 0
  ip address 215.17.3.1 255.255.255.255
!
interface Serial 5/0
  ip unnumbered loopback 0
  ip verify unicast reverse-path
!
ip route 215.34.10.0 255.255.252.0 Serial 5/0
!
router bgp 100
  network 215.34.10.0 mask 255.255.252.0
Injecting prefixes into iBGP

• interface flap will result in prefix withdraw and reannounce
  use “ip route...permanent”

• many ISPs use redistribute static rather than network statement
  only use this if you understand why
Router Configuration: redistribute static

- **Example:**

  ```
  ip route 215.34.10.0 255.255.252.0 Serial 5/0
  
  router bgp 100
   redistribute static route-map static-to-bgp
  <snip>
  
  route-map static-to-bgp permit 10
   match ip address prefix-list ISP-block
   set origin igp
  <snip>
  
  ip prefix-list ISP-block permit 215.34.10.0/22 le 30
  ```
Injecting prefixes into iBGP

• Route-map ISP-block can be used for many things:
  setting communities and other attributes
  setting origin code to IGP, etc

• Be careful with prefix-lists and route-maps
  absence of either/both means all statically routed prefixes go into iBGP
Scaling the network

How to get out of carrying all prefixes in IGP
Why use BGP rather than IGP?

- IGP has Limitations:
  - The more routing information in the network
  - Periodic updates/flooding “overload”
  - Long convergence times
  - Affects the core first

Policy definition
- Not easy to do
Preparing the Network

- We want to deploy BGP now…
- BGP will be used therefore an ASN is required
- If multihoming to different ISPs is intended in the near future, a public ASN should be obtained:
  - 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

• 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
IGP

- Decide on IGP: OSPF or ISIS 😊
- Assign loopback interfaces and /32 addresses 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 is 110, static distance is 1
  Smallest distance wins
Prepparing the Network
IGP (cont)

• Be prudent deploying IGP – keep the Link State Database Lean!
  
  Router loopbacks go in IGP
  
  Backbone 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 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
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 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, the 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?
  Techniques such as peer templates?
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
  
  Use `distance bgp 200 200 200`

  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 routes 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 either by IGP or by 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
Preparing the Network Configuration Summary

- IGP essential networks are in IGP
- Customer networks are now in iBGP
  iBGP deployed over the backbone
  Full or Partial or Upstream Edge only
- BGP distance is greater than any IGP
- Now ready to deploy eBGP
BGP Best Current Practices

ISP/IXP Workshops