BGP Best Current Practices

Philip Smith

<philip@nsrc.org>

LKNOG 5

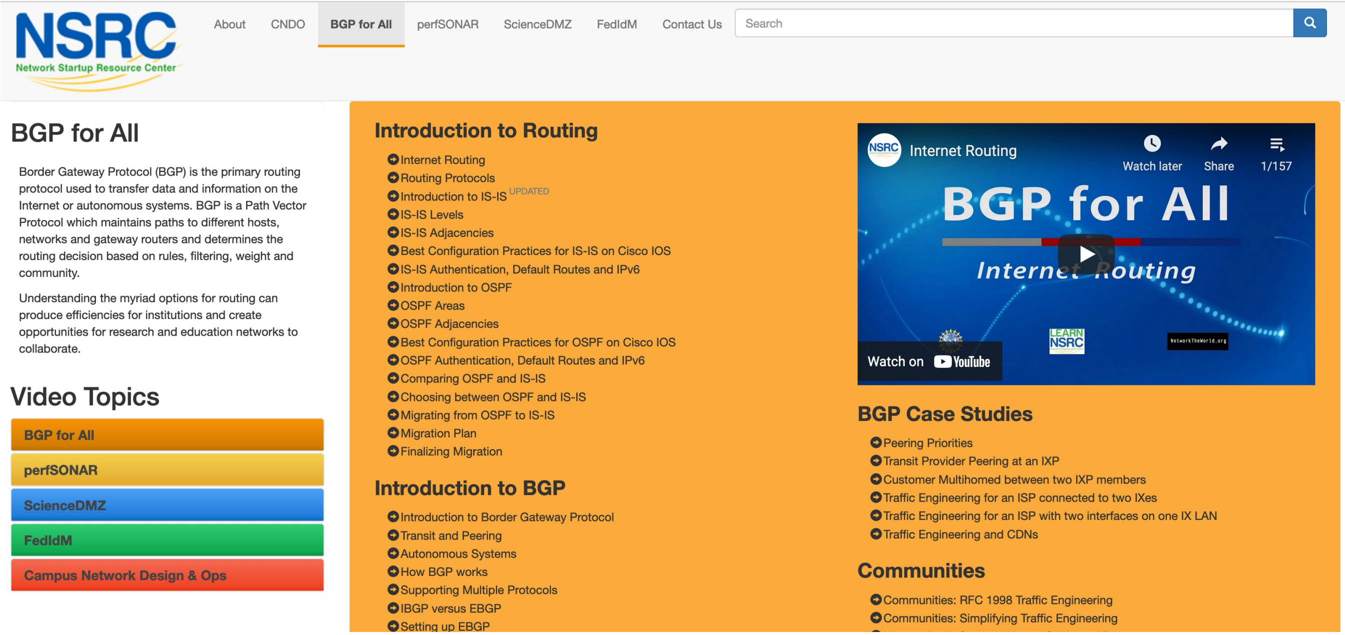
15th July 2021

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

BGP Videos

- NSRC has made a video recording of this presentation, as part of a library of BGP videos for the whole community to use:
  - https://learn.nsrc.org/bgp#bgp_best_practices



BGP for All

Border Gateway Protocol (BGP) is the primary routing protocol used to transfer data and information on the Internet or autonomous systems. BGP is a Path Vector Protocol which maintains paths to different hosts, networks and gateway routers and determines the routing decision based on rules, filtering, weight and community.

Understanding the myriad options for routing can produce efficiencies for institutions and create opportunities for research and education networks to collaborate.

Video Topics

- BGP for All
- perfSONAR
- ScienceDMZ
- FedIdM
- Campus Network Design & Ops

Introduction to Routing

- Internet Routing
- Routing Protocols
- Introduction to IS-IS UPDATED
- IS-IS Levels
- IS-IS Adjacencies
- Best Configuration Practices for IS-IS on Cisco IOS
- IS-IS Authentication, Default Routes and IPv6
- Introduction to OSPF
- OSPF Areas
- OSPF Adjacencies
- Best Configuration Practices for OSPF on Cisco IOS
- OSPF Authentication, Default Routes and IPv6
- Comparing OSPF and IS-IS
- Choosing between OSPF and IS-IS
- Migrating from OSPF to IS-IS
- Migration Plan
- Finalizing Migration

Introduction to BGP

- Introduction to Border Gateway Protocol
- Transit and Peering
- Autonomous Systems
- How BGP works
- Supporting Multiple Protocols
- IBGP versus EBGP
- Setting up EBGP

BGP Case Studies

- Peering Priorities
- Transit Provider Peering at an IXP
- Customer Multihomed between two IXP members
- Traffic Engineering for an ISP connected to two IXes
- Traffic Engineering for an ISP with two interfaces on one IX LAN
- Traffic Engineering and CDNs

Communities

- Communities: RFC 1998 Traffic Engineering
- Communities: Simplifying Traffic Engineering

BGP Best Current Practices

- Review of recommendations to ensure the BGP is:
  - Configured optimally
  - Operating optimally
  - Configured securely
  - Operating securely

BGP versus OSPF/IS-IS

- OSPF/IS-IS are used to carry infrastructure reachability prefixes only (Loopbacks, internal point-to-points)
- BGP is used internally (IBGP) and externally (EBGP)
- IBGP is used to carry:
  - Some/all Internet prefixes across backbone
  - Customer prefixes
- EBGP is used to:
  - Exchange prefixes with other ASes
  - Implement routing policy

EBGP Default Behaviour

- Industry standard is described in RFC8212
  - https://tools.ietf.org/html/rfc8212
  - *External BGP (EBGP) Route Propagation Behaviour without Policies*

- Configuring EBGP peering without using filters means:
  - All best paths on the local router are passed to the neighbour
  - All routes announced by the neighbour are received by the local router
  - Can have disastrous consequences (see RFC8212)

EBGP Default Behaviour

- FRR turns on RFC8212 support by default:
  - https://frrouting.org/

```
frr.pfs.lab(config)# router bgp 64512 view LAB
frr.pfs.lab(config-router)# bgp ?
<snip>
ebgp-requires-policy          Require in and out policy for eBGP peers (RFC8212)
<snip>
```

- Recent Cisco IOS-XE (17.3(?) onwards) has the option:

```
asr1k.pfs.lab(config)# router bgp 64512
asr1k.pfs.lab(config-router)# bgp ?
<snip>
  safe-ebgp-policy         Allow propagation of ebgp routes only with policy
                           configured. Command may trigger refresh on all
                           sessions!
<snip>
```

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
- Too many operators are still thinking about “class Cs”, resulting in a proliferation of /24s in the Internet routing table
  - July 2021: 498652 /24s in IPv4 table of 856599 prefixes
- The same is happening for /48s with IPv6
  - July 2021: 57275 /48s in IPv6 table of 124867 prefixes

In Sri Lanka

- Deaggregation happening here too
  - Top 5 prefix announcements

```
                   Sri Lanka Aggregation Savings Summary

       ASN     No of Nets   Savings  Description
      9329            122       107  SLTINT-AS-AP Sri Lanka Telecom Internet, LK
      5087             82        77  LANKA-COM Lanka Communication Services, LK
     45224             60        57  BELLNET-AS-AP Lanka Bell Limited, LK
     18001             68        55  DIALOG-AS Dialog Axiata PLC., LK
    132045             52        47  AIRTEL-AS-ISP Bharti Airtel Lanka Pvt. Limited,
```

  - Significant savings possible for each

Receiving Prefixes

- There are three scenarios for receiving prefixes from other ASes
  - 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 is “joint whois” and queries all RIR databases)

Receiving Prefixes: From Peers

- A peer is an operator with whom you agree to exchange prefixes each originates 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:
    - IRRToolSet: https://github.com/irrtoolset/irrtoolset
    - bgpq3: https://github.com/snar/bgpq3

Receiving Prefixes: From Transit Providers

- Transit Provider is an operator who you pay to give you transit to the WHOLE Internet
- Receiving prefixes from them is not desirable unless for Traffic Engineering
- Ask transit provider to either:
  - Originate a default-route

    OR

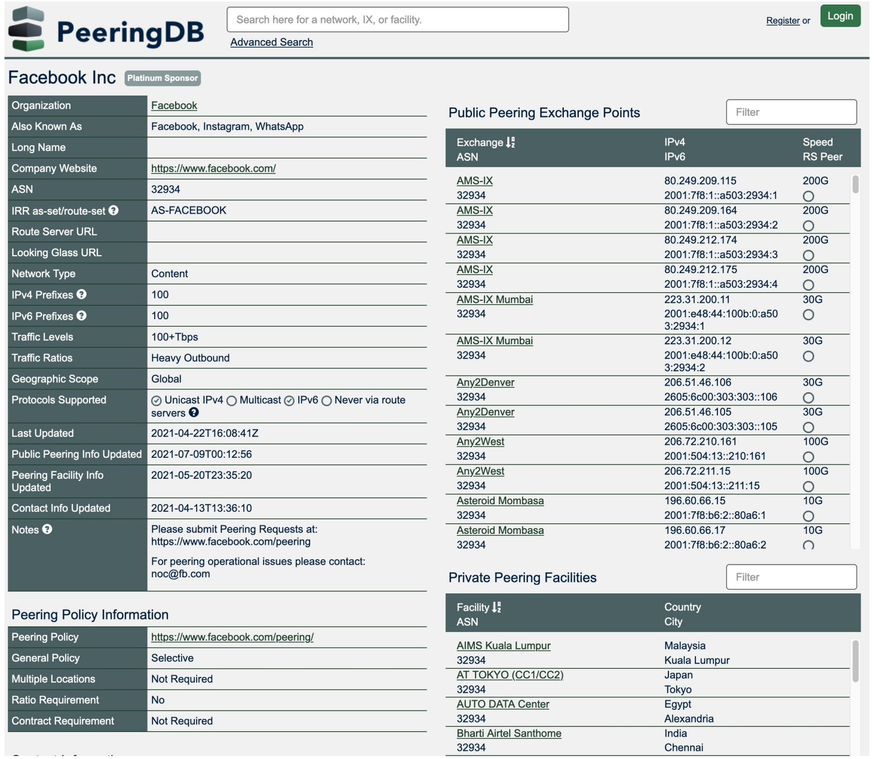
  - Send full table (including default) so you can discard what is not needed for traffic engineering

Receiving Prefixes

- If it is necessary to receive prefixes from any provider, care is required:
  - Don't accept default (unless you need it)
  - Don't accept your own prefixes
- Be careful with special use prefixes for IPv4 and IPv6:
  - http://www.rfc-editor.org/rfc/rfc6890.txt
- Don't accept IPv4 prefixes longer than /24 (historical Class-C)
- Don't accept IPv6 prefixes longer than /48 (minimum for an end-site)
- Don't accept unassigned prefixes:
  - Team Cymru's list of "bogons" & Bogon Route Server
  - http://www.team-cymru.com/bogon-reference.html

The Peering Database

- The Peering Database documents ISPs peering policies
  - https://www.peeringdb.com
- All AS operators are recommended to register in the PeeringDB
  - All operators who are considering peering or are peering must be in the PeeringDB to enhance their peering opportunities



Facebook Inc Platinum Sponsor

Organization	Facebook
Also Known As	Facebook, Instagram, WhatsApp
Long Name	
Company Website	https://www.facebook.com/
ASN	32934
IRR as-set/route-set ?	AS-FACEBOOK
Route Server URL	
Looking Glass URL	
Network Type	Content
IPv4 Prefixes ?	100
IPv6 Prefixes ?	100
Traffic Levels	100+Tbps
Traffic Ratios	Heavy Outbound
Geographic Scope	Global
Protocols Supported	☑ Unicast IPv4 ○ Multicast ☑ IPv6 ○ Never via route servers ?
Last Updated	2021-04-22T16:08:41Z
Public Peering Info Updated	2021-07-09T00:12:56
Peering Facility Info Updated	2021-05-20T23:35:20
Contact Info Updated	2021-04-13T13:36:10
Notes ?	Please submit Peering Requests at: https://www.facebook.com/peering For peering operational issues please contact: noc@fb.com

Peering Policy Information

Peering Policy	https://www.facebook.com/peering/
General Policy	Selective
Multiple Locations	Not Required
Ratio Requirement	No
Contract Requirement	Not Required

Public Peering Exchange Points

Exchange ↓ ASN	IPv4 IPv6	Speed RS Peer
AMS-IX 32934	80.249.209.115 2001:7f8:1::a503:2934:1	200G ○
AMS-IX 32934	80.249.209.164 2001:7f8:1::a503:2934:2	200G ○
AMS-IX 32934	80.249.212.174 2001:7f8:1::a503:2934:3	200G ○
AMS-IX 32934	80.249.212.175 2001:7f8:1::a503:2934:4	200G ○
AMS-IX Mumbai 32934	223.31.200.11 2001:e48:44:100b:0:a503:2934:1	30G ○
AMS-IX Mumbai 32934	223.31.200.12 2001:e48:44:100b:0:a503:2934:2	30G ○
Any2Denver 32934	206.51.46.106 2605:6c00:303:303::106	30G ○
Any2Denver 32934	206.51.46.105 2605:6c00:303:303::105	30G ○
Any2West 32934	206.72.210.161 2001:504:13::210:161	100G ○
Any2West 32934	206.72.211.15 2001:504:13::211:15	100G ○
Asteroid Mombasa 32934	196.60.66.15 2001:7f8:b6:2::80a6:1	10G ○
Asteroid Mombasa 32934	196.60.66.17 2001:7f8:b6:2::80a6:2	10G ○

Private Peering Facilities

Facility ↓ ASN	Country City
AIMS Kuala Lumpur 32934	Malaysia Kuala Lumpur
AT TOKYO (CC1/CC2) 32934	Japan Tokyo
AUTO DATA Center 32934	Egypt Alexandria
Bharti Airtel Santhome 32934	India Chennai

Internet Routing Registry

- Many major transit providers and several content providers pay attention to what is contained in the Internet Routing Registry
  - There are many IRR instances operating, the most commonly used being those hosted by the Regional Internet Registries, RADB, and some transit providers
- Best practice for any AS holder is to document their routing policy in the IRR
  - A route-object is the absolute minimum requirement

- Some network operators now using RPKI and ROAs to securely indicate the origin AS of their routes
  - Takes priority over IRR entries
  - IRR contains a lot of outdated (unmaintained) information

Internet Routing Registry

- Which IRR database to use?
  - Members of a Regional Internet Registry are recommended to use their RIR's Internet Routing Registry instance
    - Usually managed via the RIR's member portal giving easy access for creation and update of objects
    - Provided as part of the RIR's services to its members
  - Operators who do not belong to any RIR generally use:
    - Their upstream transit provider's Routing Registry (if provided)
    - The RADB
      - https://www.radb.net
      - Note: Placing objects in the RADB requires an annual subscription fee

Route Origin Authorisation

- Essential first step to secure the global routing system
- Answers this question:
  - How do we know that an AS is permitted to originate the prefix it is originating?
- Uses Resource Public Key Infrastructure (RPKI)
- Prevents route hijacking and mis-origination
- ROAs signed by RIR members via member portal
  - Digital object containing address prefixes and AS number
- Allows routers to validate prefix announcements received from EBGP peers
  - Drop *invalid*, accept *valid*, low priority for those with no ROA
- More details
  - http://www.bgp4all.com/pfs/_media/workshops/02-rpki.pdf

Configuration Tips

Of passwords, tricks and templates

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

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”

BGP Community Behaviour

- IBGP
  - Propagate communities to all IBGP speaking routers
    - Be aware: some vendors have this turned off by default
  - Use communities for internal BGP scaling
    - Tagging services and destinations
- EBGP
  - Only send communities to influence your BGP peer’s policies
    - Don’t just send all communities you have
    - Don’t send those you’ve learned from other operators
  - Only accept communities that you need to influence your BGP policies
    - Otherwise remove/overwrite

Vendor Community Policy implementation

- Be aware that each vendor has differing policy language behaviours for:
  - Treatment of well known communities
  - Setting communities
  - Removing communities
  - Replacing communities
- Consult:
  - Vendor documentation
  - https://www.rfc-editor.org/rfc/rfc8651.txt for discussion of some of the issues operators need to be aware of

Limiting AS Path Length

- Some announcements have ridiculous lengths of AS-paths
  - This example is an error in one IPv6 implementation

```
*> 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 shows 100 prepends (for no obvious reason)

```
*>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
```

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

Private-AS – Removal

- Private, Documentation, and Unassigned ASNs MUST be removed from all announcements to the public Internet
  - Include configuration to remove these ASNs in the EBGP template
- As with private, reserved and unassigned address space, these ASNs must not be leaked to or used on the public Internet

ASN:	Usage:
0 and 65535	(reserved)
64496-64511	(documentation – RFC5398)
64512-65534	(private use only)
23456	(represent 32-bit range in 16-bit world)
65536-65551	(documentation – RFC5398)
4200000000-4294967295	(private use only)

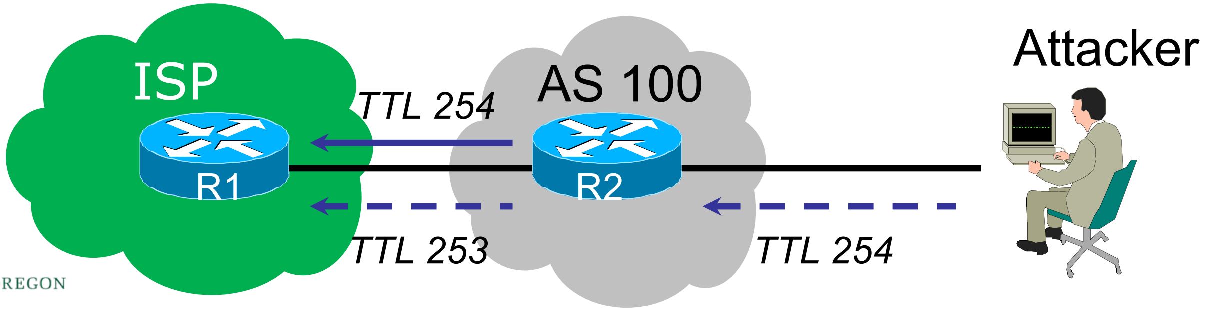
BGP Maximum Prefix Tracking

- Allow configuration of the maximum number of prefixes a BGP router will receive from a peer
  - – If you expect N prefixes, set the “maximum prefixes” to be 2xN
    - Then router will warn/tear down BGP session if the limit is exceeded
  - – If you are receiving the full BGP table, it is still a good idea to set a limit
    - Prevents against major accidental leaks
  - – Cisco IOS CLI example:

```
neighbor <x.x.x.x> maximum-prefix <max> [restart N] [<threshold>] [warning-only]
```

BGP TTL “hack”

- Implement RFC5082 on EBGP peerings
  - (Generalised TTL Security Mechanism)
  - Neighbour sets TTL to 255
  - Local router expects TTL of incoming BGP packets to be 254
  - Nothing 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



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 in following slides
  - – Also see Team Cymru's BGP templates
    - http://www.team-cymru.com/community-services.html

IBGP Template Example (1)

- IBGP between loopbacks!
- Next-hop-self
  - Keep DMZ and external point-to-point out of IGP
- Always send communities in IBGP
  - Otherwise BGP policy accidents will happen
  - (Default on some vendor implementations, optional on others)
- Hardwire BGP to version 4
  - Prevents accidental configuration of BGP version 3 which is still supported in some implementations

IBGP Template Example (2)

- 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 (1)

- BGP damping
  - Do NOT use it unless you understand the impact
  - Do NOT use the vendor defaults without thinking
- Remove private/unassigned ASNs from announcements
  - Common omission today
- Adhere to RFC8212
  - Make sure there is inbound and outbound filtering applied to all EBGP sessions!
- Use password agreed between you and peer on EBGP session

EBGP Template Example (2)

- 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!!

Routing Security

- Implement the recommendations in https://www.manrs.org
  - Prevent propagation of incorrect routing information
    - Filter BGP peers, in & out!
  - Prevent traffic with spoofed source addresses
    - BCP38 – Unicast Reverse Path Forwarding
  - Facilitate communication between network operators
    - NOC to NOC Communication
    - Up-to-date details in Route and AS Objects, and PeeringDB
  - Facilitate validation of routing information
    - Route Origin Authorisation using RPKI

MANRS

Questions?

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center