BGP

Finding your way through the series of tubes

Who’s this dude?

- My name is Nick Allgood
- I am CCIE #37428
- Working in the industry for over 10 years
- PhD student at UMBC in Baltimore, Maryland (just started)

Layer 1: How do we get where now?

- Two methods of routing to get from source to destination.
- Static Routing – Manually adding a route to a destination
- Dynamic Routing – Using a protocol to inject routes and advertise them dynamically.

- Static Routing Example
  - *Cisco:*
    - `ip route 0.0.0.0 0.0.0.0 192.168.5.1`
  - *Linux:*
    - `route add default gw 192.168.5.1`
  - *Quagga:*
    - `ip route 0.0.0.0/0 192.168.5.1`

Introduction to Dynamic Routing

- Dynamic routing protocols known as Internal Gateway Protocols or IGP's.

- RIP, OSPF, IS-IS are open standards
  - *RIP: Uses hop-count as metric, useful for limited situations*
  - *IS-IS: Uses CLNS for routing, often used in ISP's.*
  - *OSPF: Based on Djikstra's SPF algorithm, most commonly used in enterprises.*

- Routing is a physical implementation of a mathematical graph.

What's the point?

- Static routing has it's uses, but it isn't scalable.
  - *Imagine configuring the static route commands for **EVERY ROUTE** in your network.*
  - *Fun Fact: Still way more common than it should be, and it sucks.*
- With dynamic routing, you create a relationship with adjacent devices and advertise networks.

- OSPF Configuration:
  - *Cisco:*

```
router ospf 1
network 192.168.5.0  0.0.0.255 area 0
```

  - *OpenOSPFD (/etc/ospfd.conf):*

```
area 0.0.0.0 {
interface eth0 { metric 10 }
}
```

  - *Quagga:*

```
router ospf
network 192.168.5.0/24  area 0.0.0.0
```

IGP Diagram

- Filler space to remember to draw/explain IGP fundamentals.

Layer 2: BG-What? An Introduction to BGP

- IGP's address the concern internally for advertising networks, but they are strictly internal and shouldn't be used across the internet. Think of IGP's as the foundation to a house.

- External Gateway Protocol (EGP) was created to allow advertising company owned public IP blocks to the internet. Eventually, EGP was phased out for the Border Gateway Protocol (BGP).

- BGP defined currently defined by IETF RFC 4271 and is an open standard.
  - *Note: While BGP is an open standard, some vendors have proprietary options. (*cough* Cisco *cough*)*

Layer 3: My Brain Hurts: Technical bits about BGP

- Conceptually, BGP is similar to RIP in that it leverages a distance vector metric.
  - *In documentation, it's often called Path-Vector.*
- Utilizes Autonomous System (AS) numbers from ARIN.
  - *An AS is just a fancy term for organizing a company's internet presence.*
  - *Lots of companies do not have a need for an AS.*
  - *Often private AS numbers are used for small deployments.*

- Unlike RIP, instead of using the number of next devices, BGP uses the number of next AS's.
- Uses the AS_PATH entity, which is a list of AS's that the subnet has passed through.

RIP vs BGP Diagram

- Limitations of RIP
- AS Visualizations

Well known Mandatory Attributes

- These are **REQUIRED** in all BGP messages when communicating with each other.

- **ORIGIN**– How the route was placed into BGP
  - *i (IGP)– From an IGP*
  - *e (EGP)– From BGP*
  - *? (Incomplete) – Unknown means*

- **AS_PATH** – The path of AS's to travel to get to the prefix
- **NEXT_HOP** – The IP address of the next device to get to said prefix

Well Known Discretionary Attributes

- **MUST** be supported by devices and passed to next AS, but not be used
- **LOCAL_PREF** – Preference/metric to be used for calculation within an AS.
- **ATOMIC_AGGREGATE** – Used to inform neighbor that this device aggregated routes into a summarization.
  - *i.e – 192.168.0.0/26 & 192.168.0.64/26 could be aggregated to 192.168.0.0/25*

Optional Transitive Attributes

- Not required to be supported, but **MUST** be passed to the next AS
- **AGGREGATOR** – Device that performed the ATOMIC_AGGREGATE
- **COMMUNITY** – A type of "tag" used to give extra flexibility when filtering and other purposes.
  - *Fun Fact: If you're getting DDoS'd and you have the agreement with your provider, you often simply have to tag the offending subnet with a community and re-advertise it to the provider (iBGP) to mitigate.*

Optional Non-transitive Attributes

- Not required to be supported, not required to pass to next AS
- Multi-Exit Discriminator(MED) – Metric advertised to other AS's. Used to influence routing decisions **INBOUND** to an AS
- ORIGINATOR_ID - Identifies the router that originated the path
- CLUSTER_LIST – Used with route-reflectors for loop prevention.

Layer 4: Are we there yet? Establishing a BGP peering

- BGP peers to neighboring devices using TCP 179. Neighbors do not have to be directly adjacent.
- Unlike with IGP's, you must manually configure the peering's to establish a neighbor relationship.
- As part of the configuration, you must include the AS number of the remote system along with the peer's IP address.
- BGP peering's to a different AS number are known as external BGP (eBGP).
- BGP peering's to the same AS number are known as internal BGP (iBGP).

BGP Message Types

- **OPEN** – Used by both speakers to identify and begin to establish a peering..
- **KEEPALIVE** – Used between neighbors to ensure reachability. (heartbeat)
  - *Also used to acknowledge valid OPEN messages.*
- **UPDATE** – Used to exchange routing information
- **NOTIFICATION** – Used when an error has occurred. Session is closed immediately.

BGP Peering Configuration

- Cisco

```
router bgp 65535
address-family ipv4 unicast
neighbor 172.16.1.2 remote-as 65530
```

- Quagga

```
router bgp 65535
neighbor 172.16.1.2 remote-as 65530
```

- OpenBGPD (/etc/bgpd.conf)

```
AS 65535
neighbor 172.16.1.2 {
remote-as 65530
}
```

Verifying a Peering

- State should be **ESTABLISHED** or have a number at the end
  - ***ACTIVE*** *is BAD!*

- Cisco / Quagga
  - *show ip bgp summary / show ip bgp neighbor 172.16.1.2*

- OpenBGPD
  - *bgpctl show summary / bgpctl show neighbor 172.16.1.2*

Valid Peering Example

```
Neighbor        V           AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
172.16.1.2      4        65535  291896  291897 106464135    0    0 26w2d            3
```

Layer 5: Can you see me now? Advertising routes into BGP

- BGP does not advertise routes automatically once a peering is established.

- You must first inject the prefixes into the protocol itself.
  - *Manually*
  - *Redistribution*

- Once injected into the protocol, you then must specify the prefixes to advertise on a per-neighbor basis inbound and outbound.
  - *route-maps*
  - *prefix-lists*

- You should **ONLY** advertise prefixes that have been assigned to you!!!

Cisco Configuration

```
router bgp 65535
address-family ipv4 unicast
neighbor 172.16.1.2 remote-as 65530
! Manually
network 192.168.99.0 mask 255.255.255.0
! Redistribution
redistribute connected
!
neighbor 172.16.1.2 route-map TO-ISP out
neighbor 172.16.1.2 route-map FROM-ISP in
neighbor 172.16.1.2 prefix-list PFX-TO-ISP out
neighbor 172.16.1.2 prefix-list PFX-FROM-ISP in
```

Quagga Configuration

```
router bgp 65535
neighbor 172.16.1.2 remote-as 65530
! Manually
network 192.168.99.0/24
! Redistribution
redistribute connected
!
neighbor 172.16.1.2 route-map TO-ISP out
neighbor 172.16.1.2 route-map FROM-ISP in
neighbor 172.16.1.2 prefix-list PFX-TO-ISP out
neighbor 172.16.1.2 prefix-list PFX-FROM-ISP in
```

OpenBGPD Configuration(/etc/bgpd.conf)

```
AS 65535
# Advertise our space
network 192.168.99.0/24
network 172.16.55.0/24
# Neighbor Configuration
neighbor 172.16.1.2 {
remote-as 65530
}
# Filtering
deny to 172.16.1.2
allow to 172.16.1.2 from prefix 192.168.99.0/24 prefixlen = 24
allow to 172.16.1.2 from prefix 172.16.55.0/24 prefixlen = 24
```

BGP Path Selection Algorithm

- Highest WEIGHT (Cisco only)
- Highest LOCAL_PREF
- Local route generated by this device
- Shortest AS_PATH
- Lowest ORIGIN code (igp > egp > ?)
- Lowest MED
- eBGP over iBGP
- Lowest IGP cost
- Oldest BGP route installed
- Lowest BGP router ID. (IP address)

Verifying BGP Routes

- Cisco / Quagga
  - *show ip bgp*

- OpenBGPD
  - *bgpctl show rib*

BGP Route Table Example

```
   Network          Next Hop            Metric LocPrf Weight Path
*> 192.168.99.0/24  10.1.1.254               0    120      0 6059 i
*> 172.16.55.0/24   10.1.1.254                    120      0 6059 174 174 4826 38803 56203 i
```

“*” – Preferred BGP route

“>” – Installed in main route table

Layer 6: Don’t tread on me! More technical bits and security

- "In C++ it's harder to shoot yourself in the foot, but when you do, you blow off your whole leg.” - Bjarne Stroustrup
  - *BGP is very similar to this, except it’s not really hard to shoot yourself in the foot.*

- You are responsible for filtering all prefixes into your network, this include RFC 1918 addresses.

- If you do not implement filtering, you **WILL** become a transit AS at some point.

- **DO NOT** rely on your upstream provider to filter.

The dangers of a transit AS

- A transit AS is exactly as it sounds, your AS could be used as a ‘next hop’ AS from somewhere else to get to a destination.

- This is a huge security risk, your AS should never be in the transit path

- In addition to security, being a transit AS can be a huge tax on computing resources
  - *-Generally providers throttle the bandwidth based on what you pay.*
  - *It’s a very common (and recommended) practice to also throttle bandwidth locally.*

Transit AS Diagram

- Self Explanatory, but do it anyway

More Security Concerns

- Another security concern is if someone advertises public prefixes that belong to you.
- If this happens, then traffic destined to your public prefixes will be either black holed or worse, intercepted.
- This is often accidental but has been known to be used as an attack vector. Often due to incorrect filters on the upstream provider.
- Very difficult to troubleshoot, often have to use BGP looking glass to check paths.

Advertising wrong prefixes and BGP looking glass

- Draw up and explain what happens when some jerk advertises prefixes you own.

- Show a quick demo of Level3's BGP looking glass.

Diagram about RBHT

- ISP creates a dummy IP address as a next-hop that points to NULL on their BGP devices
- DDoS detected on one specific DMZ'd host, disrupting your entire service
- Create a static route of the host being attacked to NULL.
- Create a route-map to advertise this specific prefix with a BGP community and redistribute into iBGP.
- Provider will create a route-map that matches the BGP community used and then will have a policy to NULL route that prefix until the DDoS stops.

- *The needs of the many outweigh the needs of the one* -- Spock

Layer 7: The Application Layer

Questions ?

Thank you!

- Thanks for putting up with me!
- I will be here all day if you wish to learn/discuss more after the session.
- Feel free to email me at nick.allgood@gmail.com

References

- OpenBGPD, OpenOSPFD
  - https://www.openbsd.org/papers/linuxtag06-network.pdf
- Quagga
  - http://www.nongnu.org/quagga/
- Cisco
  - http://www.cisco.com/c/en/us/td/docs/ios-xml/ios/iproute_ospf/configuration/12-4t/iro-12-4t-book/iro-cfg.html
  - http://www.cisco.com/c/en/us/td/docs/ios/12_2/ip/configuration/guide/fipr_c/1cfbgp.html

More References

- Remote Black Hole Triggering (RBHT)
  https://www.cisco.com/c/dam/en/us/products/collateral/security/ios-network-foundation-protection-nfp/prod_white_paper0900aecd80313fac.pdf

- Level3 Looking Glass:
  - *http://lookingglass.level3.net/bgp/lg_bgp_main.php*

- BGP RFC 4271
  - *https://tools.ietf.org/html/rfc4271*