BGP Fundamentals

Introduction :

inter-organization connectivity on public networks, such as the Internet, or private dedicated networks. BGP is the only protocol used to exchange networks on the Internet
BGP does not advertise incremental updates or refresh network advertisements like OSPF or ISIS. BGP prefers stability within the network

Inter-Router Communication

BGP uses TCP port 179 to communicate with other routers. TCP allows for handling of fragmentation, sequencing, and reliability (acknowledgement and retransmission) of communication packet

BGP Messages

OPEN

The OPEN message is used to establish a BGP adjacency. Both sides negotiate session capabilities before a BGP peering establishes. The OPEN message contains the BGP version number, ASN of the originating router, Hold Time, BGP Identifier, and other optional parameters that establish the session capabilities . Setting a static BGP RID is a best practice

KEEPALIVE

BGP does not rely on the TCP connection state to ensure that the neighbors are still alive

UPDATE

The Update message advertises any feasible routes, withdraws previously advertised routes, or can do both. An UPDATE message can act as a Keepalive to reduce unnecessary traffic.

NOTIFICATION

A Notification message is sent when an error is detected with the BGP session, such as a hold timer expiring, neighbor capabilities change, or a BGP session reset is requested. This causes the BGP connection to close

Autonomous System Numbers ASN

- ASN 2 byte
- ASN 4 BYTE

ASNs 64,512–65,535 are private ASNs

4,200,000,000–4,294,967,294 are private ASNs. RFC 4893

The Internet Assigned Numbers Authority (IANA) is responsible for assigning all public ASNs

Path Attributes

- Well-known mandatory
- Well-known discretionary
- Optional transitive
- Optional nontransitive

Loop Prevention

The BGP attribute AS_PATH is a well-known mandatory attribute and includes a complete listing of all the ASNs that the prefix advertisement has traversed from its source AS. The AS_PATH is used as a loop prevention mechanism in the BGP protocol

Address Families

RFC 2858 added Multi-Protocol BGP (MP-BGP) MBGP achieves this separation by using the BGP path attributes (PAs) MP_REACH_NLRI and MP_UNREACH_NLRI. These attributes are carried inside BGP update messages and are used to carry network reachability information for different address families.

AFI	SAFI	Network Layer Information
1	1	IPv4 Unicast
1	2	IPv4 Multicast
1	4	IPv4 Unicast with MPLS Label
1	128	MPLS L3VPN IPv4
2	1	IPv6 Unicast
2	4	IPv6 Unicast with MPLS Label
2	128	MPLS L3VPN IPv6
25	65	Virtual Private LAN Service (VPLS)
26	65	Virtual Private Wire Service (VPWS)
25	70	Ethernet VPN (EVPN)

BGP Sessions

Internal BGP (IBGP) Sessions established with an IBGP router that are in the same AS or participate in the same BGP confederation IBGP sessions are considered more secure, and some of BGP's security measures are lowered in comparison to EBGP.

External BPG (EBGP) Sessions established with a BGP router that are in a different AS. EBGP prefixes are assigned an AD of 20 upon installing into the router's RIB

- **IOS:** IOS nodes use the highest IP address of the any *up* loopback interfaces. If there is not an *up* loopback interface, then the highest IP address of any active *up* interfaces becomes the RID when the BGP process initializes.

- **IOS XR:** IOS XR nodes use the IP address of the lowest *up* loopback interface. If there is not any *up* loopback interfaces, then a value of zero (0.0.0.0) is used and prevents any BGP adjacencies from forming.

- **NX-OS:** NX-OS nodes use the IP address of the lowest *up* loopback interface. If there is not any *up* loopback interfaces, then the IP address of the lowest active *up* interface becomes the RID when the BGP process initializes.

SAJID Abdallah Naif

BGP Fundamentals 2

Every path's attributes impact the desirability of the route when a router selects the best path. A BGP router advertises only the best path to the neighboring routers

BGP Best-Path Calculation

BGP recalculates the best path for a prefix upon

- BGP next-hop reachability change
- Failure of an interface connected to an EBGP peer
- Redistribution change
- Reception of new paths for a route

These attributes are processed in the order listed:

1. Weight
2. Local Preference
3. Local originated (network statement, redistribution, aggregation)
4. AIGP
5. Shortest-AS Path
6. Origin Type
7. Lowest MED
8. EBGP over IBGP
9. Lowest IGP Next-Hop
10. If both paths are external (EBGP), prefer the first (oldest)
11. Prefer the route that comes from the BGP peer with the lower RID
12. Prefer the route with the minimum cluster list length
13. Prefer the path that comes from the lowest neighbor address

Basic Configuration on IOS ,NX OS and IOS XR

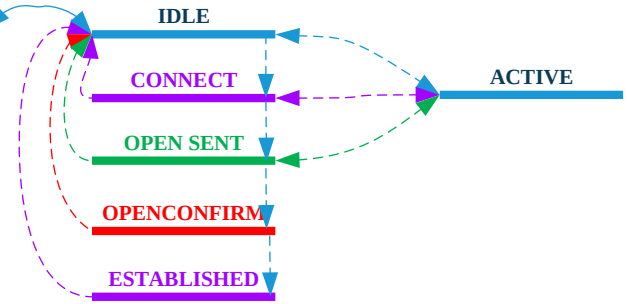
```
R1 (Default IPv4 Address-Family Enabled)
router bgp 65100
 neighbor 10.1.12.2 remote-as 65100
```

```
R2 (Default IPv4 Address-Family Disabled)
router bgp 65100
 no bgp default ipv4-unicast
 neighbor 10.1.12.1 remote-as 65100
 !
 address-family ipv4
  neighbor 10.1.12.1 activate
 exit-address-family
```

```
IOS XR
router bgp 65100
 bgp router-id 192.168.1.1
 address-family ipv4 unicast
 !
 neighbor 10.1.12.2
  remote-as 65100
  address-family ipv4 unicast
```

```
NX-OS
router bgp 65100
 address-family ipv4 unicast
 neighbor 10.1.12.2 remote-as 65100
  address-family ipv4 unicast
```

BGP Neighbor States



-Idle: BGP detects a start event, tries to initiate a TCP connection to the BGP peer, and also listens for a new connect from a peer router If an error causes BGP to go back to the Idle state for a second time
the BGP Process is administratively down.
the BGP Process is awaiting the next retry attempt.
the BGP is just configure on new neighbor Already established BGP Peering is reset.
-Connect : Connect BGP initiates the TCP connection If the 3-way TCP handshake completes, the established BGP Session If successful, it will continue to the OpenSent state. if fails, it will continue to the active state. if BGP reset is send it will move back to the idle state
-Active : Active In this state, BGP starts a new 3-way TCP handshake. If a connection is established,an Open message is sent, the Hold Timer is set to 4 minutes, and the state moves to OpenSent. If this attempt for TCP connection fails, the state moves back to the Connect state and resets the ConnectRetryTimer
-OpenSent : In this state, an Open message has been sent from the originating router and is awaiting an Open message from the other router the matching open message has not been received from peer. BGP will be waiting for an Open message from the remote BGP neighbor.
-OpenConfirm BGP waits for a Keepalive or Notification message. Upon receipt of a neighbor's Keepalive. the state is moved to Established. If the hold timer expires. a stop event occurs or a Notification message is received, and the state moved to idle .
-Established

BGP Tables

Adj-RIB-in: Contains the NLRIs in original form before inbound route policies are processed. The table is purged after all route policies are processed to save memory

Loc-RIB: Contains all the NLRIs that originated locally or were received from other BGP peers. After NLRIs pass the validity and next-hop reachability check, the BGP best path algorithm selects the best NLRI for a specific prefix. The Loc-RIB table is the table used for presenting routes to the ip routing table

Adj-RIB-out: Contains the NLRIs after outbound route policies have processed

- **Connected Network**: The next-hop BGP attribute is set to 0.0.0.0, the origin attribute is set to *i* (IGP), and the BGP weight is set to 32,768.
- **Static Route or Routing Protocol**: The next-hop BGP attribute is set to the next-hop IP address in the RIB, the origin attribute is set to *i* (IGP), the BGP weight is set to 32,768; and the MED is set to the IGP metric.

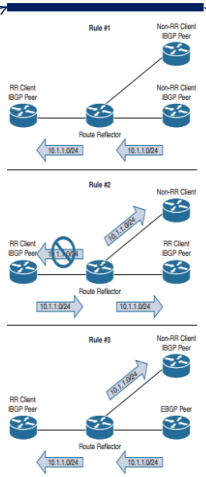
SAJID Abdallah Naif

BGP Fundamentals 3

Route Reflectors

RFC 1966 introduces the concept that an IBGP peering can be configured so that it reflects routes to another IBGP peer. The router reflecting routes is known as a *route reflector (RR),* and the router receiving reflected routes is a *route reflector client*

Rule #1: If a RR receives a NLRI from a non-RR client, the RR advertises the NLRI to a RR client. It does not advertise the NLRI to a non-route-reflector client.
Rule #2: If a RR receives a NLRI from a RR client, it advertises the NLRI to RR client(s) and non-RR client(s). Even the RR client that sent the advertisement receives a copy of the route, but it discards the NLRI because it sees itself as the route originator.
Rule #3: If a RR receives a route from an EBGP peer, it advertises the route to RR client(s) and non-RR client(s).



Loop Prevention in Route Reflectors

ORIGINATOR_ID, an optional nontransitive BGP attribute is created by the first route reflector and sets the value to the RID of the router that injected/advertised the route into the AS. If the ORIGINATOR_ID is already populated on an NLRI, it should not be overwritten. If a router receives a NLRI with its RID in the Originator attribute, the NLRI is discarded.
CLUSTER_LIST, a nontransitive BGP attribute, is updated by the route reflector. This attribute is appended (not overwritten) by the route reflector with its cluster-id. By default this is the BGP identifier. The cluster-id can be set with the BGP configuration command **bgp cluster-id** *cluster-id* on IOS and IOS XR nodes. NX-OS devices use the command **cluster-id** *cluster-id.* If a route reflector receives a NLRI with its cluster-id in the Cluster List attribute, the NLRI is discarded

Out-of-Band Route Reflectors

IBGP Scalability

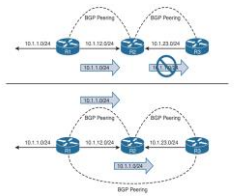
IBGP
The need for BGP within an AS typically occurs when the multiple routing policies exist, or when transit connectivity is provided between autonomous systems

Advertising the full BGP table into an IGP is not a viable solution for the following reasons:

Scalability: IPv4 networks and continues to increase in size. IGPs cannot scale to that level of routes

Custom Routing: The path could be longer, which would normally be deemed Subo-ptimal from an IGP protocol's perspective

Path Attributes: All the BGP path attributes cannot be maintained within IGP protocols.



IBGP Full Mesh Requirement:

IBGP peers do not prepend their ASN to the AS_PATH, because the NLRIs would fail the validity check and would not install the prefix into the IP routing table RFC 4271 states that all BGP routers within a single AS must be fully meshed to provide a complete loop-free routing table and prevent traffic blackholing best practice is **Peering via Loopback Addresses** because is more efficient and preferable

EBGP

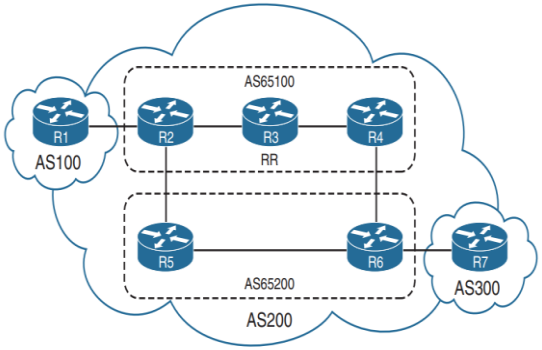
EBGP peerings are the core component of the BGP protocol on the Internet. EBGP is the exchange of network prefixes between autonomous systems. The following behaviors are different on EBGP sessions when compared to IBGP sessions:

■ Time to Live (TTL) on BGP packets is set to one. (TTL on IBGP packets is set to 255, which allows for multihop sessions).
■ The advertising router modifies the BGP next-hop to the IP address sourcing the BGP connection.
■ The advertising router prepends its ASN to the existing AS_PATH.
■ The receiving router verifies that the AS_PATH does not contain an ASN that matches the local routers. BGP discards the NLRI if it fails the AS_PATH loop prevention check.

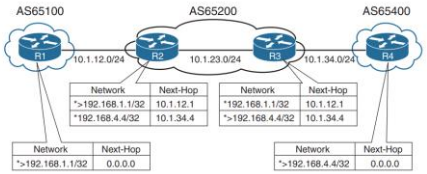
Confederations

RFC 3065 introduced the concept of BGP confederations as an alternative solution to IBGP full mesh scalability issues A confederation consists of sub-ASs known as a Member-AS that combine into a larger AS known as an AS Confederation. Member ASs normally use ASNs from the private ASN range (64512-65535). EBGP peers from the confederation have no knowledge that they are peering with a confederation, and they reference the confederation identifier in their configuration

Confederations share behaviors from both IBGP sessions and EBGP sessions. The changes are as follows:
■ The AS_PATH attribute contains a subfield called AS_CONFED_SEQUENCE.
The AS_CONFED_SEQUENCE is displayed in parentheses before any external ASNs in the AS_PATH. As the route passes from Member-AS to Member-AS, the AS_CONFED_SEQUENCE is appended to contain the Member-AS ASNs. The AS_CONFED_SEQUENCE attribute is used to prevent loops, but it is not used (counted) when choosing shortest AS_PATH.
■ Route reflectors can be used within the Member-AS like normal IBGP peerings.
■ The BGP MED attribute is transitive to all other Member-ASs, but does not leave the confederation.
■ The LOCAL_PREF attribute is transitive to all other Member-ASs, but does not leave the confederation.
■ IOS XR nodes do not require a route policy when peering with a different
Member-AS, even though the **remote-as** is different.
■ The next-hop address for external confederation routes does not change as the route is exchanged between Member-AS to Member-AS.
■ The AS_CONFED_SEQUENCE is removed from the AS_PATH when the route is advertised outside of the confederation



EBGP and IBGP Topologies



Next-Hop Manipulation
Configuring the **next-hop-self** address-family feature modifies the next-hop address in all external NLRIs using the IP address of the BGP neighbor

Combining EBGP sessions with IBGP sessions can cause confusion in terminology and concepts The most common issue involves the failure of the next-hop accessibility.
IBGP peers do not modify the next-hop address if the NLRI has a next-hop address other than 0.0.0.0. The next-hop address must be resolvable in the global RIB for it to be valid and advertised to other BGP peers.To correct the issue,by two possible technique
■ IGP advertisement. Remember to use the passive interface to prevent an accidental
adjacency from forming. Most IGPs do not provide the filtering capability like BGP.
■ Advertising the networks into BGP

BY/KH Abdallah Nail

BGP Fundamentals 4

Route Summarization

Summarizing prefixes conserves router resource(s) and accelerates best path calculation by reducing the size of the table. Summarization also provides the benefit(s) of stability by reducing routing churn by hiding route flaps from downstream routers

The two techniques for BGP summarization are the following:
- **Static:** Create a static route to Null 0 for the prefix, and then advertise the network via a network statement. The downfall to this technique is that the summary route will always be advertised even if the networks are not available.
- **Dynamic:** Configure an aggregation network range. When viable routes that match the network range enter the BGP table, an aggregate route is created. On the originating router, the aggregated prefix sets the next-hop to Null 0. The route to Null 0 is automatically created by BGP as a loop-prevention mechanism.

In both methods of route aggregation, a new network prefix with a shorter prefix length is advertised into BGP. Because the aggregated prefix is a new route, the summarizing router is the originator for the new aggregate route

Aggregate-Address

summary-only – suppress all less specific, by default the aggregate does not do that

Flexible Route Suppression
Some traffic engineering designs require "leaking" routes, which is the advertisement of a subset of more specific routes in addition to performing the summary

Leaking Suppressed Routes The **summary-only** keyword suppresses all the more specific routes of an aggregate address from being advertised. After a route is suppressed, it is still possible to advertise the suppressed route to a specific neighbor **unsuppress-map**

Selective Prefix Suppression Selective prefix suppression explicitly lists the networks that should not be advertised along with the summary route to neighbor routers **suppress-map**

Atomic Aggregate

The Atomic Aggregate attribute indicates that a loss of path information has occurred.

Route Aggregation with AS_SET

To keep the BGP path information history, the optional **as-set** keyword may be used with the **aggregate-address** command. As the router generates the aggregate route, BGP attributes from the summarized routes are copied over to it The AS_SET is displayed within brackets

Route Aggregation with Selective Advertisement of AS-SET

Using the AS-SET feature with network aggregation combines all the attributes of the original prefixes into the aggregated prefixes. This might cause issues with your routing policy. be aware about that you can use the **advertise-map** option allows for conditionally matching and denying attributes that should be permitted or denied in the aggregated route

Default Route Advertisement

Advertising a default route into the BGP table requires the default route to exist in the RIB and the BGP configuration command **default-information originate** to be used. The redistribution of a default route or use of a network 0.0.0.0/0 does not work without the **default-information originate** command

Default Route Advertisement per Neighbor

Some network topologies restrict the size of the BGP advertisements to a neighbor because the remote router does not have enough processing power or memory for the full BGP routing table **neighbor** *ip-address* **default-originate**

BGP Communities

BGP communities provide additional capability for tagging routes and for modifying BGP routing policy on upstream and downstream routers. BGP communities can be appended, removed, or modified selectively on each attribute as the route travels from router to router. *BGP communities* are an optional transitive

A BGP community can be displayed as a full 32-bit number (0-4,294,967,295) or as two 16-bit numbers (0-65535):(0-65535) commonly referred to as *new-format Private BGP communities* follow the convention that the first 16-bits represent the AS of the community origination, and the second 16-bits represent a pattern defined by the originating AS In 2006, RFC 4360 expanded BGP communities' capabilities by providing an extended format. *Extended BGP communities* provide structure for various classes of information and are commonly used for VPN Services

IOS XR and NX-OS devices display BGP communities in new-format by default, and IOS nodes display communities in decimal format by default
ip bgp-community new-format
IOS and NX-OS devices do not advertise BGP communities to peers by default
IOS XR advertises BGP communities to IBGP peers by default

Well known

- no-advertise – do not send beyond local router (0xFFFFFF02)
- no-export – do not send beyond local AS (0xFFFFFF01)
- local-as – do not send to ebgp sub-AS peers within confed (0xFFFFFF03)
- internet – permit any – overwrite all communities and allow prefix to be announced everywhere
- gshut – gracefull shutdown, like overload bit in ISIS, „go around me" signal to all BGP speakers

Remove Private AS

remove private AS feature:
- Removes only private ASNs on routes advertised to EBGP peers.
- If the AS-Path for the route has only private ASNs, the private ASNs are removed.
- If the AS-Path for the route has a private ASN between public ASNs, it is assumed that this is a design choice, and the private ASN is not removed
- If the AS-Path contains confederations (AS_CONFED_SEQ), BGP removes the private AS numbers only if they are included after the AS_CONFED_SEQ (Confederation AS-Path) of the path.

Allow AS

The *Allow AS* feature allows for routes to be received and processed even if the router detects its own ASN in the AS-Path

Local AS

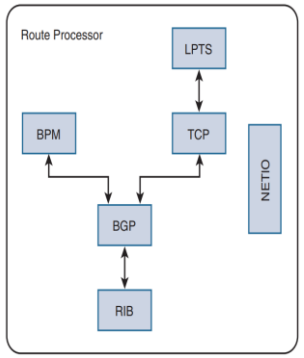
The *LocalAS* feature is configured on a per peer basis, and allows for BGP sessions to establish using an alternate ASN than the ASN that the BGP process is running on. The LocalAS feature works only with EBGP peerings One problem with the alternate ASN being prepended when receiving the routes is that other IBGP peers drop the network prefixes as part of a routing loop detection.
- To stop the alternate ASN from being prepended when *receiving routes*, the optional keyword **no-prepend** is used.
- To stop the alternate ASN from being prepended when *sending routes*, the optional keywords **no-prepend replace-as** is used.
- If both **no-prepend replace-as** keywords are used, all routers see the BGP advertisements as if they were running the original AS in the BGP process.

Common BGP Troubleshooting Dynamic BGP Peering

Common BGP Troubleshooting

BGP Peering Down Issues

Troubleshoot Blocked Process in IOS XR IOS XR is a distributed operating system, and every component (feature) runs as a separate process with its own set of threads that manages various tasks of the component.Unlike traditional IOS , In IOS XR, the BGP Process Manager (BPM) and BGP processes create the BGP protocol functionality.The BPM process also has the responsibility to calculate the router-id if one is not explicitly configured. It interacts with NETIO, TCP, and a few other processes internally to perform the necessary tasks in the system and finally installs the routes in the Routing Information Base (RIB)



Verify that the BGP and BPM processes are in Run state by using the command **show process** *process-name* [**detail** | **location** {**LC/RP** *location* | **all**}

Verify Blocked Processes
Execute **show process blocked** [**location** {**RP/LC**}] to verify whether there are any blocked processes, which could cause an impact on the BGP process. Primarily the *bgp*, *bpm*, *tcp*, and *netio* processes are the ones that are critical

Restarting a Process
If for some reason a process is in blocked state for a long period of time, restart the process using **process restart** [*job-id* | *process-name*].

Decode BGP Messages BGP generates a hex dump of the message. These hex dumps can then be analyzed to understand why the BGP router was unable to process the message. There are external websites that help decode BGP messages; for example, http://bgpaste.convergence.cx

Debugs for BGP
Running debugs should always be the last resort for troubleshooting any network problem. Debugs can sometimes cause an impact in the network if not used carefully. But sometimes they are the only options techniques cannot repair the problem when other troubleshooting
debug bgp ipv4 unicast ip-address

The peering down issue occurs because of one of the following circumstances

- During establishment of BGP sessions because of misconfiguration
- Triggered by network migration or event, or software or hardware upgrades
- Failure to maintain BGP keepalives due to transmission problems
- High CPU
- Blocked or stuck processes
- Firewall or ACL misconfiguration
- Software defects

A down BGP peer state is in either an Idle or Active state.

- **Idle** state No connected route to peer

Active state
- No route to peer address (IP connectivity not present)
- Configuration error, such as update-source missing or wrongly configured

Idle/Active state
- Transmission Control Protocol (TCP) establishes but BGP negotiation fails; for ex: misconfigured AS
- Router did not agree on the peering parameters

show running-config | section router bgp //Config
show bgp ipv4 unicast neighbor X.X.X.X | in TTL // *TTL Values*
ping X.X.X.X source Y.Y.Y.Y // Verifying Reachability
ping xxxx source loopback0

If the connectivity problem is in the forwarding path or the return path.

- Find the location and direction of packet loss **show ip traffic + include echo**
- Verify whether packets are being transmitted. If there is complete packet loss on the link, perform a ping connectivity test with the timeout set to 0 to confirm if the packet is actually leaving the router or if the other side is receiving the packets
  show interface Gi0/1 | in packets
  ping 10.1.12.1 timeout 0 repeat 10
- Use access control lists (ACL) to verify that packets are received ACLs prove to be really useful when troubleshooting packet loss or reachability issues. Configuring an ACL matching the source and the destination IP can help confirm whether the packet has reached the destination router

Many deployments have firewalls to protect the network from unwanted and malicious traffic. It is a better option tohave a firewall installed than to have a huge ACL configured on the routers and switches. Firewalls can be configured in two modes:
- Routed mode
- Transparent mode

In Routed mode, the firewall has routing capabilities and is considered to be a routed hop in the network. In Transparent mode, the firewall is not considered as a router hop to the connected device but merely acts like a "bump in the wire.

In some deployments, network operators add NAT on the routed firewalls. In cases where NAT is configured on the router or on the firewall, the BGP peering should be configured with the translated IP address rather than the remote IP

- Verify TCP sessions A BGP session is a TCP session. Therefore, it is very important to verify if the TCP session is getting established to ensure successful BGP session establishment
  show sockets connection tcp
  show tcp brief all
- Simulate a BGP session. A good troubleshooting technique for BGP peers that are down is using Telnet on TCP port 179 toward the destination peer IP and implementing local peering IP as the source. This technique helps ensure that the TCP is not getting blocked or dropped between the two BGP peering devices
  telnet x.x.x.x source loopback 0

One way to minimize the configuration is by using BGP peer groups. If there are multiple neighbors that will share the same remote-as number or the same outbound policies, peer groups make it very easy to manage the configuration for those neighbors.This feature is not available for IPv6 addresses and dynamic BGP neighbor feature is not available on IOS XR and NX-OS.

The BGP dynamic neighbor concept is helpful in a hub-spoke topology where only the spoke router needs to have the peering configuration toward the hub. The spoke routers can be part of the same subnet. The hub router only needs to know the subnet. It can also be useful in topologies where RR is configured and there are huge numbers of RR clients. Similarly, a dynamic BGP peering concept can be used with confederations.

Dynamic BGP Peer Configuration

Step 1. Define the peer group by using Rtr(config-router)# **neighbor** *peer-group name* **peer-group**.
Step 2. Create a global limit of BGP dynamic subnet range neighbors. The value ranges from 1 to 5000. Rtr(config-router)# **bgp listen limit** *value*.
Step 3. Configure an IP Subnet Range and associate it with a peer group. Multiple subnets can be added to the same peer group. Rtr(config-router)# **bgp listen range** *subnet* **peer-group** *peer-group-name*
Step 4. Define the remote-as for the peer group. Optionally, define the list of ASnumbers that can be accepted to form neighborship with. The max limit of alternate-as numbers is 5. Rtr(config-router)# **neighbor** *peer-group-name* **remote-as** *asn* [**alternate-as** [*asn*] [*asn*] [*asn*] [*asn*]].
Step 5. Activate the peer group under ipv4 address-family by using Rtr(config-router af)# **neighbor** *peer-group-name* **activate**.

Note The alternate-as option is not available when configuring IBGP sessions.

```
R1# show running-config | section router bgp
router bgp 65535
 bgp log-neighbor-changes
 bgp listen range 10.1.0.0/16 peer-group DYNAMIC-BGP
 bgp listen limit 200
 neighbor DYNAMIC-BGP peer-group
 neighbor DYNAMIC-BGP remote-as 65530 alternate-as 65531 65532 65533
 !
 address-family ipv4
 ! Loopback Advertisement
  network 192.168.1.1 mask 255.255.255.255
 ! peer group activated
  neighbor DYNAMIC-BGP activate
 exit-address-family
```

Dynamic BGP Challenges
With dynamic BGP features, additional challenges are present, such as
- Misconfigured MD5 password
- Resource issues in a scaled environment
- TCP starvation

Misconfigured MD5 Password
This problem is very common and is generally caused by human error due to typo mistakes. You have to be careful when configuring passwords on the router configured for dynamically establishing a BGP neighbor relationship

Resource Issues in a Scaled Environment
The router Does not have any resources to serve any request coming to it. So proper planning must be done to determine how many neighbors can dynamically form BGP neighbor relationships on the router.

TCP Starvation
UDP occupies all the queues and makes TCP starve for bandwidth. Therefore, it is good to limit the number of BGP neighbors and be cautious during removal/addition of new IP subnet ranges

Demystifying BGP Notifications

Error Code	Subcode	Description
01	00	Message Header Error
01	01	Message Header Error—Connection Not Synchronized
01	02	Message Header Error—Bad Message Length
01	03	Message Header Error—Bad Message Type
02	00	OPEN Message Error
02	01	OPEN Message Error—Unsupported Version Number
02	02	OPEN Message Error—Bad Peer AS
02	03	OPEN Message Error—Bad BGP Identifier
02	04	OPEN Message Error—Unsupported Optional Parameter
02	05	OPEN Message Error—Deprecated
02	06	OPEN Message Error—Unacceptable Hold Time
03	00	Update Message Error
03	01	Update Message Error—Malformed Attribute List
03	02	Update Message Error—Unrecognized Well-known Attribute
03	03	Update Message Error—Missing Well-Known Attribute
03	04	Update Message Error—Attribute Flags Error
03	05	Update Message Error—Attribute Length Error
03	06	Update Message Error—Invalid Origin Attribute
03	07	(Deprecated)
03	08	Update Message Error—Invalid NEXT_HOP Attribute
03	09	Update Message Error—Optional Attribute Error
03	0A	Update Message Error—Invalid Network Field
03	0B	Update Message Error—Malformed AS_PATH
04	00	Hold Timer Expired
05	00	Finite State Machine Error
06	00	Cease
06	01	Cease—Maximum Number of Prefixes Reached
06	03	Cease—Peer Deconfigured
06	04	Cease—Administrative Reset
06	05	Cease—Connection Rejected
06	06	Cease—Other Configuration Change
06	07	Cease—Connection Collision Resolution
06	08	Cease—Out of Resources

Example

```
R2#
*Sep 19 15:55:34.859: %BGP-3-NOTIFICATION: received from neighbor 10.1.13.2 active
  2/2 (peer in wrong AS) 4 bytes 0000FFFA
15:55:34.860: %BGP-5-NBR_RESET: Neighbor 10.1.13.2 active reset (BGP Notification
  received)
15:55:34.866: %BGP-5-ADJCHANGE: neighbor 10.1.13.2 active Down BGP Notification
received
15:55:34.867: %BGP_SESSION-5-ADJCHANGE: neighbor 10.1.13.2 IPv4 Unicast topology
base removed from session  BGP Notification received
```

Common BGP Troubleshooting

BGP Peer Flapping Issues

Route Advertisement

Troubleshooting BGP Route Advertisement

Local Route Advertisement Issues

An exact route must exist in the router's RIB (routing table) for the route to be installed into the BGP table so that it can be advertised to BGP neighbors. There are two solutions: modify the BGP configuration to match the local networks that already exist in the RIB or create a static route for the network in the BGP configuration

The static route uses the Null 0 interface as a safety mechanism to prevent routing loops. If Rtr has a more explicit route (longer match), it can forward the packet to that direction. If it does not have a more explicit route, the packet is dropped

Route Aggregation Issues

The aggregate route is not present because there are not any prefixes within the summary aggregate prefix range in the BGP table By adding the smaller network prefixes into the BGP table, the aggregate route can be created To keep the smaller prefixes from being advertised, they can be filtered with the router's outbound BGP policy or through the suppression locally by appending the keyword **summary-only** to the **aggregate-address** command

Route Redistribution Issues

Redistributing routes into BGP is a common method of populating the BGP table,Some of the OSPF and IS-IS routes were not redistributed into BGP for the following reasons:

- **OSPF:** When redistributing OSPF into BGP, the default behavior includes only routes that are internal to OSPF (O or O IA). The redistribution of external OSPF routes requires a conditional match in the redistribution statement and/or an optional redistribution route-map.
- **IS-IS:** IS-IS does not include directly connected subnets for any destination routing protocol. This behavior is overcome by redistributing the connected networks into BGP

redistribute ospf 1 match internal external 1 external 2

Note Although not directly related to the advertisement of routes into BGP, issues can arise when redistributing routes from BGP to an IGP protocol. By default, BGP does not redistribute internal routes (routes learned via an IBGP peer) into an IGP protocol (that is, OSPF) as a safety mechanism. The command **bgp redistribute-internal** allows IBGP routes to be redistributed into an IGP routing protocol

BGP Tables

Step 1. Verify next-hop reachability. Confirm that the next-hop address is resolvable in the global RIB. If the next-hop address is not resolvable in the RIB, the NLRI remains but does not process after Step 2. The next-hop address must be resolvable for the BGP best path process to occur in Step 3

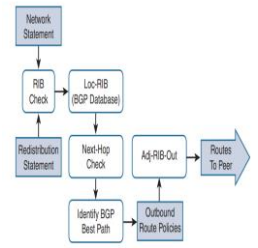
Step 2. Set BGP path attributes. The following BGP PAs are set dependent upon the location of the route in the local RIB: **Network /static route or routing protocol /redistribution**

Step 3. Identify the BGP best path. In BGP, route advertisements consist of the NLRI and the path attributes (PAs) A BGP router only advertises the best path to the neighboring routers BGP recalculates the best path for a prefix upon four possible events:

- BGP next-hop reachability change
- Failure of an interface connected to an External Border Gateway Protocol (EBGP) peer
- Redistribution change
- Reception of new paths for a route

Step 4. Process outbound neighbor route policies The NLRI is processed through any specific outbound neighbor route policies

Step 5. Advertise the NLRI to BGP peers. The router advertises the NLRI to BGP peers. If the NLRI's next-hop BGP PA is 0.0.0.0, then the next-hop address is changed to the IP address of the BGP session



Receiving and Viewing Routes

Step 1. Perform a quick validity check. This is performed on the route to ensure that a routing loop is not occurring.like (ASN) in the AS-Path or its router-ID (RID)

Step 2. Store the route in Adj-RIB-In and process inbound route policies. The NLRI is stored in the Adj-RIB-In table in its original state. The inbound route policy is applied based on the neighbor the route was received.

Step 3. Update the Loc-RIB. The BGP Loc-RIB database is updated with the NLRI after inbound route-policy processing has occurred

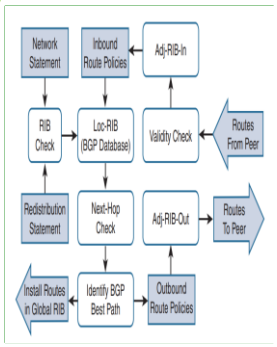
Step 4. Verify next-hop reachability. Confirm that the next-hop address is resolvable in the global RIB.

Step 5. Compute the BGP best path. Multiple NLRIs (paths) can exist for the same network prefix in the Loc-RIB table

Step 6. Install the BGP best path into the global RIB and advertise to peers. Install the prefix into the Global RIB using the next-hop IP address from the BGP Loc-RIB table command RIB failure is seen with the command show ip bgp rib-failure.

Step 7. Process outbound neighbor route policies. The NLRI is processed through any specific outbound neighbor route policies.

Step 8. Advertise the NLRI to BGP peers. Advertise the NLRI to BGP peers. If the NLRI's next-hop BGP PA is 0.0.0.0, then the next-hop address is changed to the IP address of the BGP session



the BGP state keeps flapping between Idle and Established states

Flapping BGP peers could be due to one of several reasons:

- **Bad BGP update**
- **Hold timer expired**
- **MTU mismatch**
- **High CPU**
- **Interface and platform drops**
- **Improper control-plane policing**

Bad BGP Update

- Bad link carrying the update; bad hardware
- Problem with BGP update packaging
- Malicious update by an attacker (hacker)

Hold Timer Expired

- **Interface issues**
- **Physical connectivity** packets are not transmitted correctly through the wire.
- **Physical interface** the interface was unable to process the packet because it was receiving traffic at an excessive rate
- **Input hold queue** Packets arrive to the router but are dropped in the input hold queue of the incoming Interface **Show interface xxx**
- **TCP receive queue and BGP InQ** BGP keepalives arrive at the TCP receiving queue but are not being processed and moved to the BGP InQ. When a non-zero value is seen for the BGP neighbor in the **show bgp** *afi safi* **summary** command, it indicates that the TCP messages are waiting in queue to be processed.

MTU Mismatch Issues

- **Mismatch MTU**
- Improper planning and network design
- Device not supporting Jumbo MTU or certain MTU values
- Unknown transport circuits such as EoMPLS (may not support Jumbo MTU end to end)
- Change due to application requirement
- Change due to end customer requirement

MSS value defaults to 536 bytes as defined in RFC 879

RFC 1191, PMTUD is introduced to reduce the chances of IP packets getting fragmented along the path and to help with faster convergence. Using PMTUD, the source identifies the lowest MTU along the path to destination and then decides what packet size to send

- The interface MTU on both the peering routers do not match.
- The Layer 2 path between the two peering routers do not have consistent MTU settings.
- PMTUD didn't calculate correct MSS for the TCP BGP session.
- BGP PMTUD could be failing because of blocked ICMP messages by a router or a firewall in path

High CPU Causing Control-Plane Flaps

- CPU process issues
- Interrupt (traffic processing)

show process cpu sorted | exclude 0.0

If the CPU is high due to interrupts, it could be due to one of the following problems:

- Excess process switched packets
- Packets with TTL value of 1
- Excess control plane packets

The following methods help mitigate the problems caused by packets hitting the CPU:

- Configuring an ACL to block the packets once identified
- Configuring rate limiters
- Using Control Plane Policing (CoPP)

Control Plane Policing

This scenario can result in one of the following issues:

- Loss of line protocol keepalives, update which can cause a line to go down and lead to route flaps and major network transitions.
- Near 100% CPU utilization can lock up the router and prevent it from completing high-priority processing,
- When the RP is near 100% utilization, the response time at the user command line interface (CLI) is very slow or the CLI is locked out.
- Resources including memory, buffers, and data structures can be consumed causing negative side effects. Drops of important packets.
- Router crashes.

The Control Plane Policing (CoPP) feature increases the device security by protecting its CPU (Route Processor) from unwanted and excess traffic or Denial of Service (DoS) attacks

SAJID Abdallah Naid

Common BGP Troubleshooting BGP Peer Flapping Issues

Regular Expressions (Regex)

Modifier	Description
_ (Underscore)	Matches a space
^ (Caret)	Indicates the start of the string
$ (Dollar Sign)	Indicates the end of the string
[] (Brackets)	Matches a single character or nesting within a range
- (Hyphen)	Indicates a range of numbers in brackets
[^] (Caret in Brackets)	Excludes the characters listed in brackets
() (Parentheses)	Used for nesting of search patterns
| (Pipe)	Provides *or* functionality to the query
. (Period)	Matches a single character, including a space
* (Asterisk)	Matches zero or more characters or patterns
+ (Plus Sign)	One or more instances of the character or pattern
? (Question Mark)	Matches one or no instances of the character or pattern

Note The.^$*+()[]? characters are special control characters that cannot be used without using the backslash (\) escape character. For example, to match on the * in the output you would use the * syntax.

Looking Glass and Route Servers
Hands-on experience is helpful when learning technologies such as regex
http://www.bgp4.net or
http://www.traceroute.org

Troubleshooting Missing BGP Routes

Reasons that route advertisement fails between BGP peers are as follows:
- Next-Hop Check Failure
- Bad Network Design
- Validity Check Failure
- BGP Communities
- Mandatory EBGP Route Policy for IOS XR
- Route filtering

- **BGP Loc-RIB:** Just because a route is missing from the Global RIB show bgp afi safi [prefix/prefixlength]
- **BGP Adj-RIB-in:** The BGP Loc-RIB table contains only valid routes that passed the router's inbound route policies
- **BGP Adj-RIB-out:** Viewing the BGP Adj-RIB-out table on the advertising router verifies that the route was advertised and provides a list of the BGP PAs that were included with the route
- **Viewing BGP Neighbor Sessions:** The information contained in the BGP neighbor session varies from platform to platform, but still provides a lot of useful information, such as the number of prefixes advertised….. So on
show bgp afi safi neighbor ip-address.
- **NX-OS Event History:** NX-OS contains a form of logging that runs in the background and is not as intensive as running a debug.
show bgp event-history detail.
- **Debug Commands:** Debug commands provide the most amount of information about BGP.

Troubleshooting Convergence Issues

A BGP speaker faces convergence issues primarily because of a large BGP table size and an increase in the number of BGP peers. The different dimensional factors while investigating BGP convergence issues that need to be considered include the following:
- Number of peers
- Number of address-families
- Number of prefixes/paths per address-family
- Link speed of individual interface, individual peer
- Different update group settings and topology
- Complexity of attribute creation and parsing for each address-family

Faster Detection of Failures One of the biggest factors leading to slower convergence is the mechanism to detect failures. BFD is used in conjunction with BGP to help detect failures

Jumbo MTU for Faster Convergence 9176 byte update messages can be sent to the neighbors instead of the default 536 byte update messages. This increases the efficiency because fewer update messages need to be sent to the peer

Slow Convergence due to Periodic BGP Scan To overcome this issue, the BGP scan time is reduced by using the command **bgp scantime *time-in-seconds***, where the timer can be set to any value between 5 seconds and 60 seconds. But this is not an effective solution A better way to overcome this issue is by using the BGP next-hop tracking (NHT) feature

Slow Convergence due to Default Route in RIB Default route makes the configuration simpler by allowing all traffic, but it is very important to understand where the default route needs to be advertised in the network and what impact it can potentially have. Although at times a default route is required, if configured inappropriately, it can lead to convergence issues and traffic black hole

Selective Next-Hop Tracking BGP NHT overcomes the problem faced because of periodic BGP scan by introducing the event-driven quick scan paradigm, but it still does not resolve the inconsistencies caused by default route or summarized route present in the RIB. To overcome these problems, a new enhancement was introduced in BGP NHT called the BGP SelectiveNext-Hop Tracking or BGP Selective Next-Hop Route Filtering the command bgp nexthop route-map route-map-name

Slow Convergence due to Advertisement Interval BGP neighbor advertisement interval or MRAI causes delays in update generation if set to a higher value configured manually. It is a good practice to have the same MRAI timer at both ends of the neighbor and also across different platforms

Computing and Installing New Path BGP always selects only one best path (assuming BGP multipath is not configured). In case of failure of the best path, BGP has to go through the path selection process again to compute the alternative best path. This takes time and thus impacts convergence time. Also, features such as BGP NHT help improve the convergence time by providing fast reaction to IGP events, but that is still not significant because it depends on the total number of prefixes to be processed for best-path selection. With the BGP multipath feature, equal cost paths can be used for both redundancy and faster failover

Next-Hop Check Failures

The next-hop IP address could be not available in the RIB.
- advertises the peering link into BGP
- Establish an IGP routing protocol within AS and advertise the peering link but make the peering link interface passive
- configure the next-hop-self feature in the address-family for the BGP peering

there are many solution it's depending in design and other factors .

Bad Network Design

Networks that use BGP are more sensitive to design flaws than networks that use only IGP routing protocols. An improperly design BGP network can result in an inconsistent routing policy, missing routes, or worse.

Validity Check Failure

BGP performs a validity check upon receipt of prefixes. Specifically, BGP is looking for indicators of a loop, such as
- Identifying the router's ASN in the AS-Path : **AS-Prepending,Route Aggregation** After configuring the as-set keyword on Rtr, Rtr includes the PAs from the smaller aggregate routes
- Identifying the router's RID in as the Route-Originator ID
- Identifying the router's RID as the Cluster ID

BGP Communities

BGP communities provide additional capability for tagging routes and are considered either *well-known* or *private* BGP communities. Private BGP communities are used for conditional matching for a router's route policy
There are three well-known communities that affect only outbound route advertisement: No-Advertise, No-Export, and Local-As
show bgp *afi safi* [**community** {**local-AS** | **no-advertise** | **no-export**}]
BGP Communities: No-Advertise
The No_Advertise community (0xFFFFFF02 or 4,294,967,042) specifies that routes with this community should not be advertised to any BGP peer.
BGP Communities: No-Export
The No_Export community (0xFFFFFF01 or 4,294,967,041) specifies that when a route is received with this community, the route is not advertised to any EBGP peer. If the router receiving the No-Export route is a confederation member, the route can be advertised to other sub-ASs in the confederation.
BGP Communities: Local-AS (No Export SubConfed)
The No_Export_SubConfed community (0xFFFFFF03 or 4,294,967,043) known as the Local-AS community specifies that a route with this community is not advertised outside of the local AS. If the router receiving a route with the Local-AS community is a confederation member, the route can be advertised only within the sub-AS (Member-AS) and is not advertised between Member-ASs.

BGP Slow Peer

BGP Slow Peer Symptoms
There are two common symptoms when the BGP slow peer condition is seen:
- High CPU due to BGP Router process
- Prefixes not getting replicated and traffic black hole

BGP Slow Peer Detection
BGP slow peer condition can be easily detected with the help of show commands. The
following steps help identify a BGP slow peer:
Step 1. Verify OutQ in **show bgp ipv4 unicast summary** output.
Step 2. Verify SndWnd field in the **show bgp ipv4 unicast neighbor** *ip-address*
Step 3. Verify CSize along with Current Version and Next Version fields in **show bgp ipv4 unicast replication** output.
Step 4. Verify CPU utilization due to BGP Router process.

Filtering of Prefixes by Route Policy

The last component for finding missing BGP routes is through the examination of the BGP routing policies. As stated before, BGP route policies are applied before routes are inserted into the Loc-RIB table and as prefixes leave the Loc-RIB before they are advertised to a BGP peer.
- **Prefix-list:** A list of prefix matching specifications that permit or deny network prefixes in a top-down fashion similar to an ACL.
- AS-Path ACL/Filtering: A list of regex commands that allows for the permit or deny of a network prefix based on the current AS-Path values.
- Route-maps: Route-maps provide a method of conditional matching on a variety of prefix attributes and taking a variety of actions. Actions could be a simple permit or deny or could include the modification of BGP path attributes.

SAJID Abdallah Naif