BGP Overview

www.huawei.com

HUAWEI

Foreword

In accordance with the scope of operation, dynamic routing protocols can be divided into IGP and EGP types, IGP operating within the same AS, and mainly used to detect and calculate routes, and exchange routing information. An EGP runs between AS` to provide a loop-free routing information exchange. BGP is a typical EGP.

Objectives

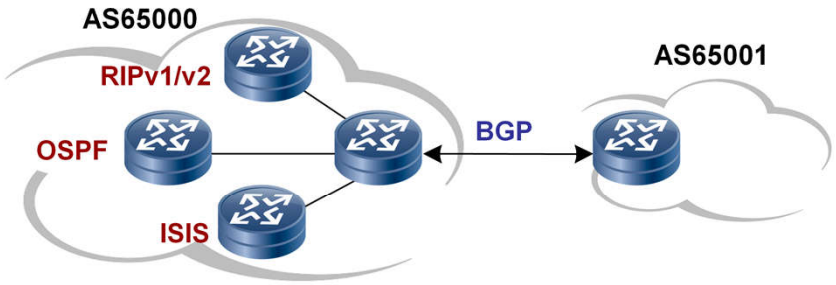
Upon completion of this section, you will be able to :

- Understand BGP`s operational scope
- Understand BGP functions
- Understand BGP features

Content

- **What is BGP?**
- Basic principal of BGP
- BGP massage type
- BGP database

Autonomous System



- **Autonomous System (AS):** refers to a set of routers managed by the same technical management organization and adopts the unified routing strategy
- Routing protocols within an AS ——IGP
- Routing protocols between the AS` ——EGP

  HUAWEI

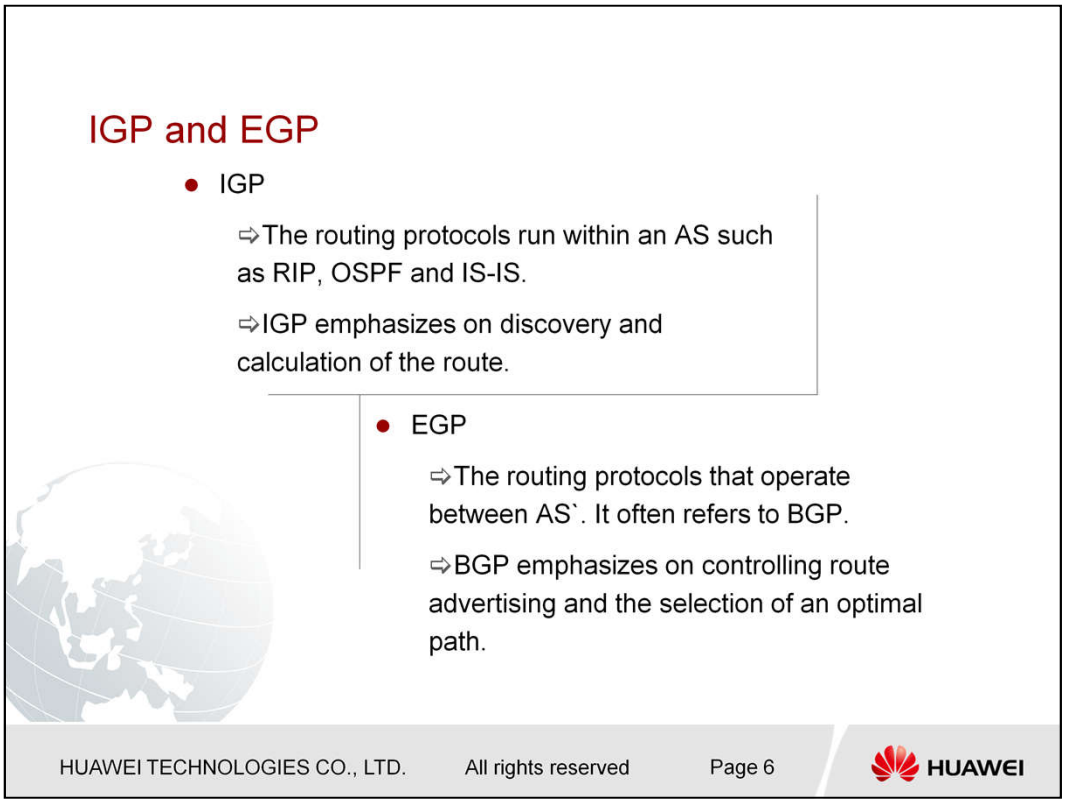
Autonomous system refers to a set of routers managed by the same technical organization and adopts the unified routing strategy.

Each autonomous system has a unique AS number which is allocated by IANA.

We distinguish different autonomous systems by using different numbers. When the network administrator does not want his/her communication data to pass some autonomous systems, these AS numbers become very useful. For example, the network administrator wants to avoid some autonomous systems managed by his competitor or avoid some autonomous systems which are lack of security mechanism. In this case, network

administrator can specify the path for data transmission by using routing protocol, routing policy and AS number.

The AS numbers range from 1 to 65535. Among them, AS numbers 1 to 64511 are the registered Internet numbers while those from 64512 to 65535 are reserved for private use.



The differences between the IGP and EGP are:

1. IGP is the routing protocols that run within an autonomous system such as RIP, OSPF, and IS-IS. It emphasizes on discovery and calculation of the route.

2. EGP is the routing protocols that run between the autonomous systems.

Nowadays, it is often referred to BGP. BGP emphasizes on control of route advertising and selection of optimal path.

BGP Characteristics

- BGP is an exterior routing protocol, used to transmit routing information between AS`
- BGP is an enhanced distance vector routing protocol
  - ⇨ Reliable updates
  - ⇨ Rich metrics
  - ⇨ No loop in design
- Many attributes for each route
- Support Classless Inter-Domain Routing (CIDR)
- Abundant route filtering and routing policies

  HUAWEI

BGP (Border Gateway Protocol) is a dynamic routing protocol that runs between the autonomous systems. Its basic function is to automatically exchange the loop-free network reachability information between the ASs. This network reachability information contains the list of autonomous system that the reachability information traverses. This reachability information is sufficient to construct the topology map of the AS from which the routing loops may be avoided and some routing policies at the AS level can be enforced. Protocols like OSPF and RIP are interior gateway protocols (IGPs) that run inside an autonomous system while BGP is exterior gateway protocol (EGP) that run between ISPs.

BGP was first introduced in 1988. The earliest versions of BGP are RFC1105 (BGP-1), RFC1163 (BGP-2) and RFC1267 (BGP-3). The current version of BGP is RFC4271/RFC1771 (BGP-4). BGP-4 has become the standard routing protocol for Internet.

The features of the BGP are as follow:

BGP provides the exchange of loop-free routing information between the autonomous systems (loop-free routing is guaranteed by using the AS-PATH attribute). BGP is policybased

routing protocol. It enforces the policy through abundant BGP

route attributes. It works on application layer and uses TCP as the transport layer protocol (BGP exchanges the route

between the neighbors on top of the reliable TCP connection). BGP is a distance vector routing protocol. This means that it will announce to its neighbors those networks that it can reach by

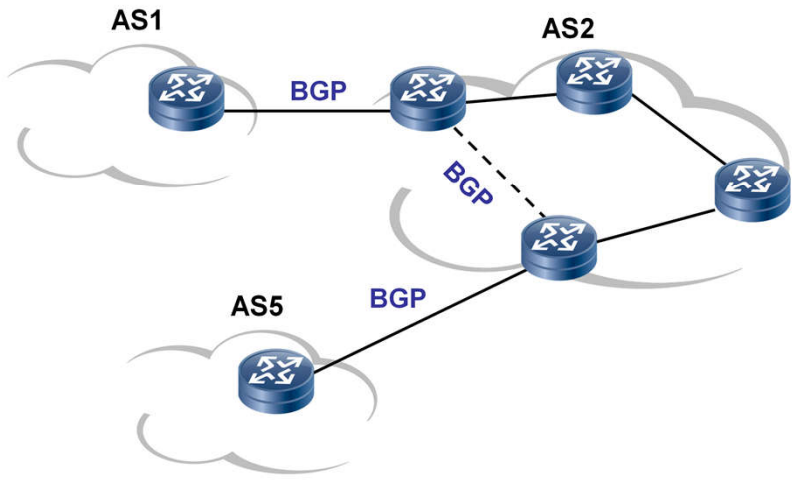
itself. However, the BGP route selection is not solely based on the distance (route selection is based on the bandwidth for most of the routing protocols). The BGP route selection is based on

the abundant route attributes. These attributes attached to the reachable IP subnets.

Therefore, we called BGP a distance vector routing protocol. It is easier to understand BGP as a distance vector routing protocol when we treat the whole AS as a single router. Apart from

that, BGP has some features of the link state protocol. For example, incremental updated, advertising route with IP subnet mask a, etc..

BGP Route Advertisement



To establish the TCP connection, two routers of each connection must know the IP address of each other. A router can learn the IP address of another router via direct connection, static route, or IGP.

The border router of the ISP will try to establish the TCP connection after successfully learning the IP address of the other end. If the connection is not established successfully, the routers

will try to re-establish the connection. This process will be repeated until the connection is established successfully.

When the TCP connection is successful, two routers will exchange some information to verify the capability of the other end or determine the next action to be performed. This is

necessary because any equipment that support IP protocol stack can support the establishment of TCP connection. However, not all the equipments that support IP protocol stack can support BGP. Therefore, the exchange of information is to guarantee the

capability of the router in supporting BGP. After confirming the capability of the routers, information from the BGP tables is exchanged.

The two routers that establish the BGP connection form the peer

relationship with each other. To guarantee the normal operation of the BGP process, the two ends of the peers will send the keepalive message periodically to ensure the validity of the connection.

If one end of the peer can’t receive any keepalive message from its peer within the hold time interval, the BGP process can be considered has been stopped in the neighbor. Therefore, the TCP connection will be closed and all the routes learnt from the neighbor will be removed.

BGP Reliable Updates

- Transport protocol: TCP, port 179
- Periodic updates are not required
- Route updates: incremental updates
- Sends keepalive messages periodically to maintain the TCP connection.

BGP uses TCP (port 179) as its transport protocol. TCP can ensure the reliable transmission of BGP.

Periodic updates is not required.

For route updating, BGP sends incremental routes only (for example new reachable route, changed route or withdrawn route). This greatly reduces the bandwidth occupied by BGP

route advertising. Therefore, BGP is suitable for advertising a large volume of routing information over the Internet.

During the initialization, BGP router sends all the routes to its peer and at the same time it saves the routes which have been sent in its local database. When the local BGP receives a

new route, it will compare this new route with the saved information which has been sent out.

If the local BGP has not sent this new route before, it will send it out. Otherwise, the local BGP will compare this new route with the route already sent. If the new route is better, then local BGP will send out this new route and at the same time update the route already sent.

Else, if the new route is worse, it will not send the new route. What will the local BGP do if it finds out that one of the local

roules fails (for example the corresponding port fails)

and this route was once sent? In this case, the local BGP will send a route withdrawal message to the BGP peer. In conclusion, BGP does not necessarily broadcast all the routing information every time. It only sends the incremental routes after the initialization, which ensures the minimum communication between the BGP and its peer.

In addition, BGP sends and receives the keepalive message to verify the TCP connectivity.

BGP Message Types

- BGP uses four message types:
  - ⇨ Open: form the neighbor relationship with BGP peer
  - ⇨ Keepalive: sent periodically between the BGP peers to verify the TCP connectivity
  - ⇨ Update: advertise routing information between the BGP peers
  - ⇨ Notification: notify the peers when BGP speaker detects the error

  HUAWEI

Routers that run a BGP routing process are often referred to as BGP speakers. Four types of messages that are exchanged between the BGP speakers are Open, Keepalive, Update, and Notification. Among them, Open, Keepalive and Notification messages are used to establish and maintain the neighbor relationship.

Open Message: includes the BGP version, AS number of the sender and etc. After the establishment of TCP connection, two routers that try to establish the neighbor relationship will exchange the Open message and check whether the neighbor relationship can be established.

Keepalive Message: Keepalive message is exchanged periodically to maintain the neighbor relationship. It is used to verify the connectivity of the peer.

Update Message: Update message is used to exchange the routing information between peers.

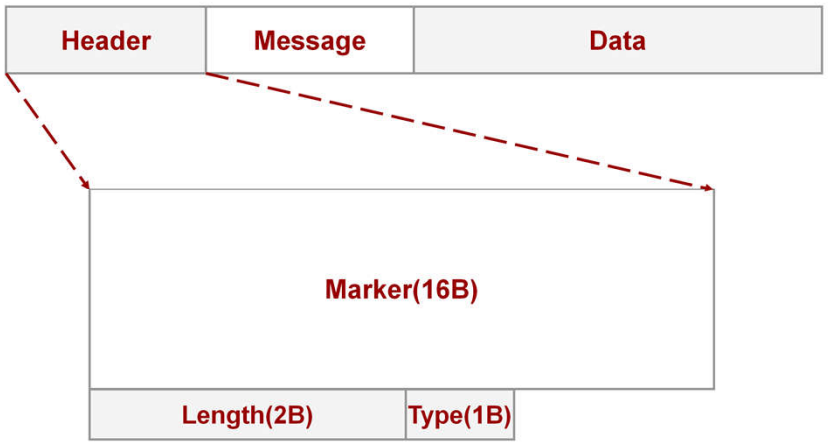
It consists of withdrawn route information, network layer reachability information and various paths attribute information. Among the four message types, update message is the most important message for BGP.

Notification Message: Notification message is the error checking

mechanism used in BGP.

BGP speaker will send the notification message when an error occurs. This will always cause the BGP connection to be closed.

BGP Message Header



  HUAWEI

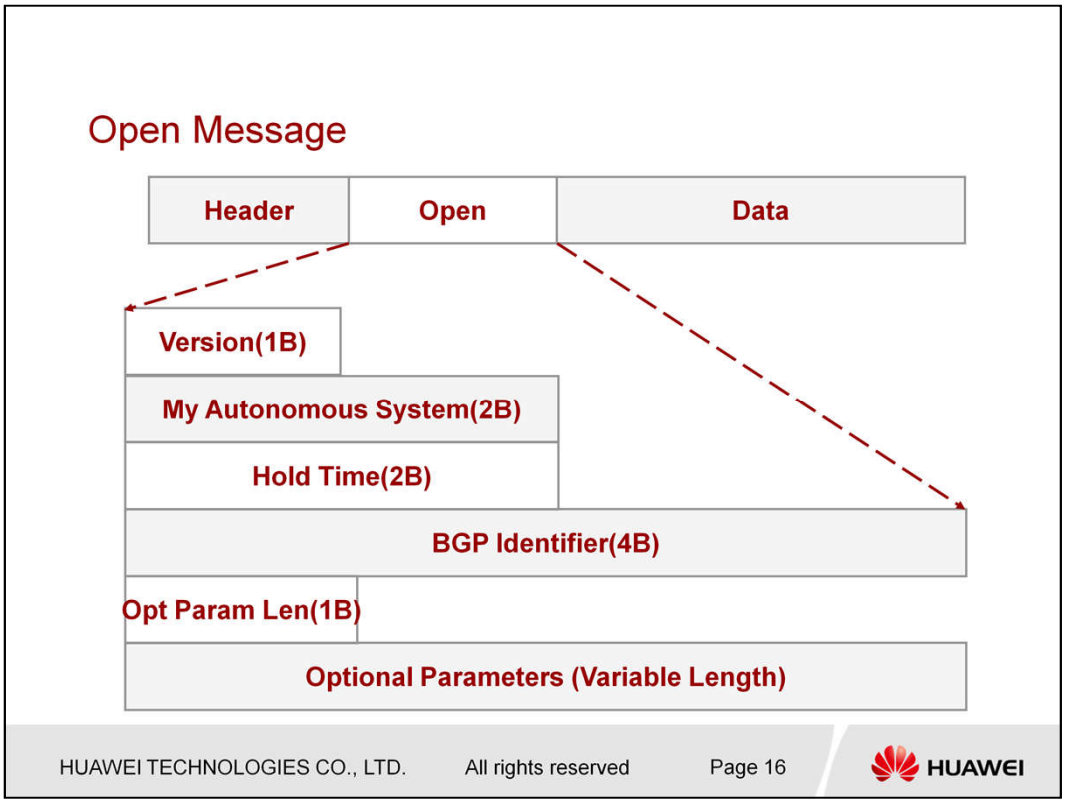
Marker：This 16-octet field contains a value that the receiver of the message can predict. In open message without authentication,then the Marker must be all ones. Otherwise, the value of the marker can be predicted by some a computation specified as part of the authentication mechanism used. The Marker can be used to detect loss of synchronization between a pair of BGP peers, and to authenticate incoming BGP messages. Length: This 2-octet unsigned integer indicates the total length of the message, including the header, in octets. Type：This 1-octet unsigned integer indicates the type code of the message. The following type codes are defined:

1 –Open

2 –Update

3 –Notification

4 –Keepalive



After a transport protocol connection is established, the first message sent by each side is an OPEN message. If the OPEN message is acceptable, a KEEPALIVE message confirming the OPEN is sent back. Once the OPEN is confirmed, UPDATE, KEEPALIVE, and NOTIFICATION messages may be exchanged.

The following describe each of the Open message fields:

Version ： This 1octet unsigned integer number indicates the BGP version number of the originator.

My Autonomous System ： This 2-octet unsigned integer indicates the Autonomous System number of the sender.

Hold Time： This 2-octet unsigned integer indicates the number of seconds that the sender proposes for the value of the Hold Timer. Upon receipt of an OPEN message, a BGP speaker MUST calculate the value of the Hold Timer by using the smaller of its configured Hold Time and the Hold Time received in the OPEN message. It is the maximum number of seconds that may elapse between the receipt of successive KEEPALIVE and/or UPDATE messages. The value of hold timer increases from

0 to the hold time value. The hold timer will be reset to 0 when the Keepalive or Update message is receipt. The neighbor will be declared dead when the hold timer expired.

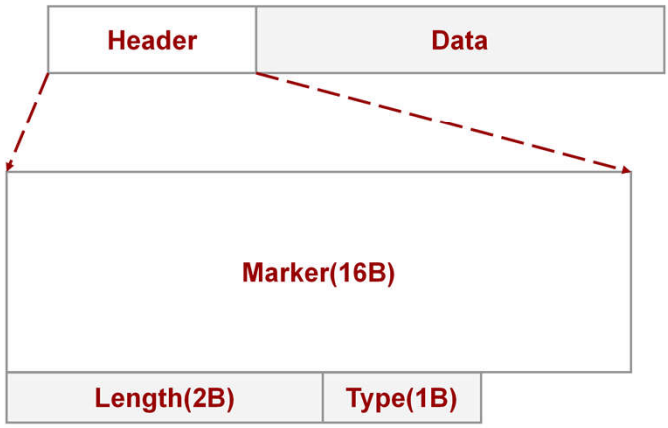
BGP Identifier ： indicates the router ID of the sender. This value is determined during the handshake operation between the BGP peers. The value of the BGP Identifier is the same for every local interface and every BGP peer.

Optional Parameters Len ： indicates the total length of the optional parameters in bytes. A length value of 0 indicates that no optional parameters are present.

Optional Parameters: indicates a list of optional parameters used in BGP neighbor session negotiation. This field is represented by one or several triplet <Parameters Types, Parameter Length, Parameter Value> with lengths of 1 byte, 1 byte and variable length, respectively.You can refer the optional parameters from RFC3392.

Keepalive Message

- Keepalive message only carries the header

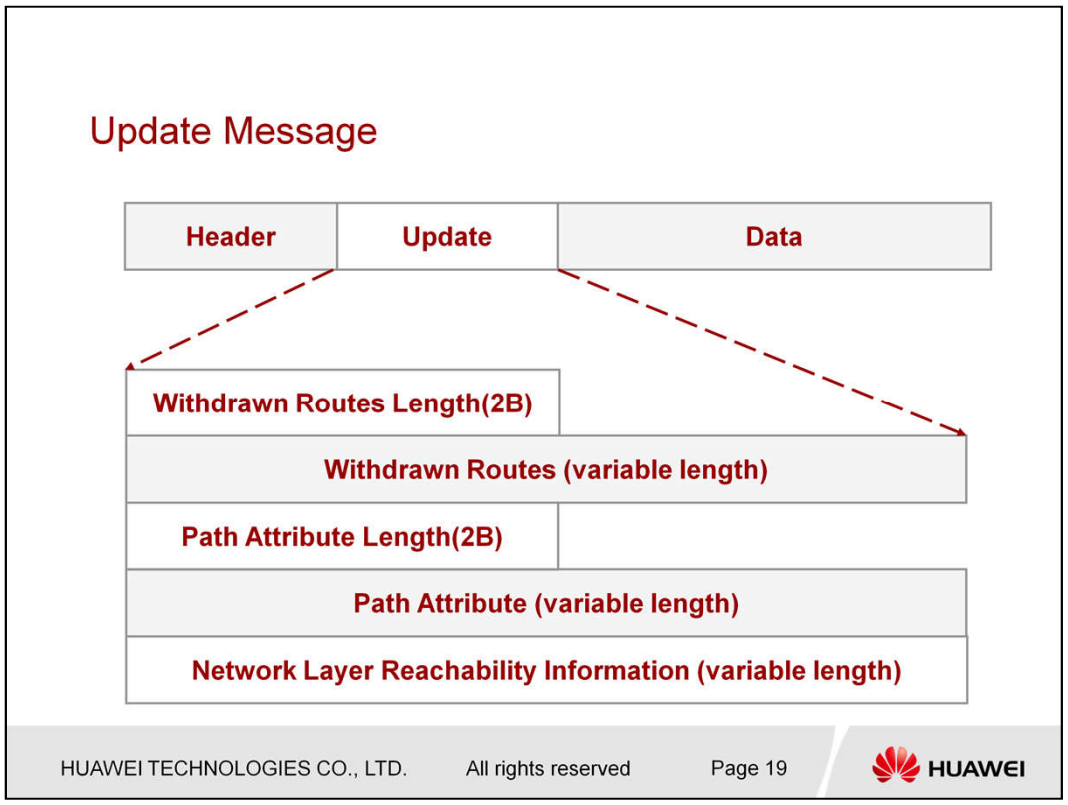


  HUAWEI

KEEPALIVE messages are sent periodically between BGP neighbors to ensure that the connection is valid. Keepalive message consists of only message header and has a length

of 19 octets. The KEEPALIVE messages are sent at a rate that ensures that the hold time will not expire When the BGP connection has been formed between a router and its neighbor, the Keepalive message will be sent periodically to the peer at every keepalive interval. This is to ensure the availability of the connection.

The recommended KEEPALIVE rate is one-third of the Hold Timer value. By default, the Keepalive interval is 60s while the hold time interval is 180s. The value of hold timer increases from 0 to the hold time value. The hold timer will be reset to 0 when the Keepalive or Update message is receipt. The neighbor will be declared dead when the hold timer expired.



Update messages are used to transfer routing information between the BGP peers. Update message consists of the following field:

Withdrawn Routes Length: (2 bytes unsigned integer) indicates the length of the withdrawn route. A withdrawn Routes Length of 0 indicates that no routes are to be withdrawn and that no Withdrawn Routes field is included in the message.

Withdrawn Routes: (Variable Length) indicates a list of routes to be withdrawn. Each route in the list is described with a (Length, Prefix) tuple in which the Length is the length of the prefix and the Prefix is the IP address prefix of the withdrawn route. For example, <19, 198.18.160.0> indicates network 198.18.160.0 255.255.224.0.

Path Attribute Length: (2 bytes unsigned integer) indicates the total length of the Path Attribute field in octets. A value of zero indicates that the path attributes field is empty.

Path Attributes: (variable length) lists the attributes associated with the NLRI. Each path attribute is a variable-length triple of (Attribute Type, Attribute Length, Attribute Value).

Network Layer Reachability Information: (variable length) consists of a list of (Length, Prefix) tuples in which the format is

the same as withdrawn route field. The Length

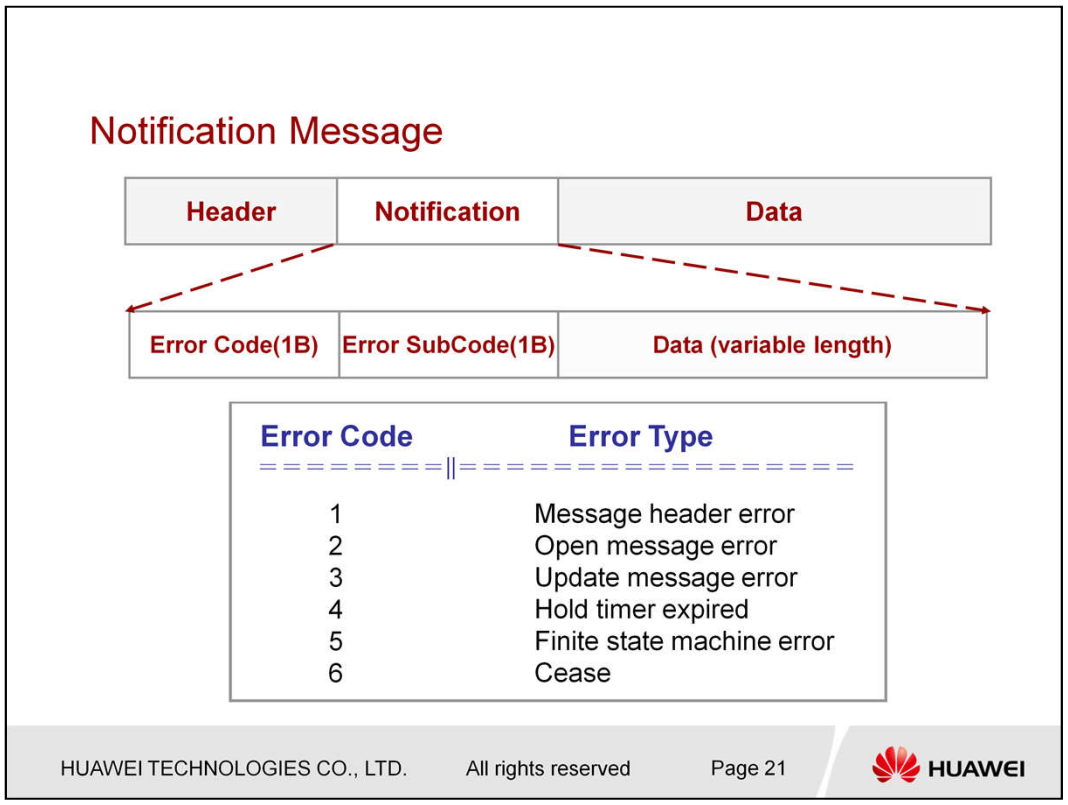
indicates the length in bits of the following prefix, and the Prefix is the IP address prefix of the NLRI.

The smallest Update information is 23 bytes (19 bytes of header + 2 bytes of withdrawn route length + path attribute length). This type of update information is called End-of-RIB and it is used in BGP GR.

One UPDATE message can advertise one route only at one time, but it can also carry multiple path attributes.

One UPDATE message can also advertise multiple routes at one time, but the path attributes must be the same.

One UPDATE message can list multiple withdrawn routes at one time.



A notification message is used when error occurs or the peer connection is stopped. This message carries various error codes (e.g. timer expiry), error subcode and error information.

Errorcode: A 1 byte field indicaties the type of error. Every errors is identified by the unique error code. Every error code can contain one or more error sub codes. If no appropriate

Error Sub code is defined, a zero value is used for the Error Subcode field.

Errsubcode：

Message Header Error Sub Codes:

1 – Connection not synchronized.

2 – Incorrect message length

3 – Incorrect message type

Open Message Error Sub Codes:

1 – Unsupported Version Number.

2 – Incorrect Peer AS.

3 – Incorrect BGP Identifier.

4 – Unsupported Optional Parameter.

5 – RFC1771 defines it as Authentication Failure. It is deprecated in RFC4271.Please refer to RFC1771/RFC4271

6 – Unacceptable Hold Time.

Update Message Error Sub Codes:

1 – Malformed Attribute List.

2 – Unrecognized Well-known Attribute.

3 – Missing Well-known Attribute.

4 – Attribute Flags Error.

5 – Attribute Length Error.

6 – Invalid ORIGIN Attribute

7 – RFC1771 defines it as AS Routing Loop. It is deprecated in RFC4271.Please refer to RFC1771/RFC4271.

8 – Invalid NEXT_HOP Attribute.

9 – Optional Attribute Error

10 – Invalid Network Field.

11 – Malformed AS_PATH.

Data: This variable-length field is used to diagnose the reason for the NOTIFICATION. The contents of the Data field depend upon the Error Code and Error Subcode. Note that the length of the Data field can be determined from the message Length field by the formula:

Message Length = 21 + Data Length. The minimum length of the NOTIFICATION message is 21 octets (including message header).

Application of Messages in BGP

- After a transport protocol connection is established, the first message sent by each side is an OPEN message.
- After the connection is established, the UPDATE message is sent to notify the peer of the routing information if a route needs to be sent or route change occurs
- After stabilization, it is necessary to send the KEEPALIVE message periodically to ensure the validity of the BGP connection
- NOTIFICATION message is sent to notify the BGP peer when an error is detected during the running of local BGP.

  HUAWEI

BGP uses TCP port 179 to establish the connection with its peer. Similar to the establishment of TCP connection, BGP uses a series of session and handshakes to establish the BGP connection. TCP uses the handshake negotiation to advertise the

parameters like port. The handshake negotiation parameters of BGP include BGP version, hold timer of BGP connection, local router ID, authentication information and so on. These parameters are included in the Open message.

After the BGP connection is formed successfully, the Update message is sent to advertise the routing information to the peer if there is a route to be sent. Update message carries the attribute of the route when it is used to distribute the routing information to the peer. This attribute information can help the peer to select the best route. Update message can also be used to inform the changes to the BGP peer when the route of the local BGP changes.

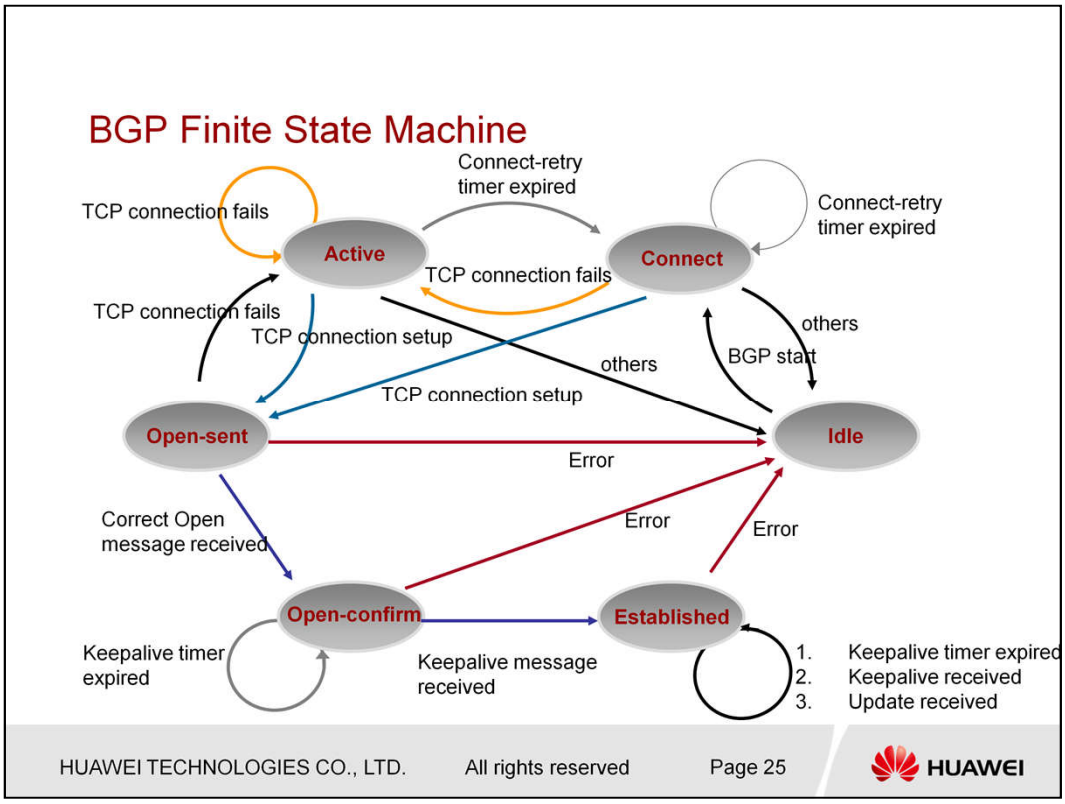
After exchanging the routing information for a period of time between the local BGP and the peer BGP, the status become stable when no new route to be advertised. At this moment,

Keepalive message is sent periodically to verify the validity of the BGP connection. When the hold time for a particular peer is

expired and the local BGP still doesn’t receive any

BGP message from its peer, this BGP connection will be regarded as invalid. As a result, the BGP connection is closed and the local BGP will withdraw all the BGP routes learnt from that BGP peer.

A Notification message is sent to notify the BGP peer when an error is detected during the running of the BGP. For example, the local BGP does not support the BGP version of the peer; the local BGP receives the Update message with illegal structure from the peer and so on. Besides, the local BGP that exits the BGP connection will also send a Notification message.



Idle： This is the first state of the BGP connection in which BGP is waiting for a start event.

Example of start event are establishing a BGP session through router configuration or resetting an already existing session. After the Start event, BGP initializes its resources, resets a ConnectRetry timer, initiates a TCP connection, and starts listening for a connection that may be initiated by a remote peer. BGP then transitions to a Connect state.

In case of errors, BGP falls back to the Idle state.

Connect: In this state, BGP establish the firs TCP connection. If the TCP connection is successful, the state transitions to OpenSent (this is where the OPEN message is sent). If the connection is fail, the state transitions to Active. If the Connect Retry timer expires, the state remains in the Connect stage, the timer is reset, and a TCP connection is initiated again.

Active: In this state, BGP always attempt to establish the TCP connection. If the Connectretry timer expires, it will return to the connect state. Otherwise, it will enter the OpenSent state. If the TCP connection fails, it will remain in the Active state and keep initiating the TCP connection. In addition, BGP continues to listen for a connection that might be initiated from another peer.

The state might go back to Idle in case of other events, such as a Stop event initiated by the system or the operator.

OpenSent: In this state, BGP connection has been established. The Open message has been sent, and BGP is waiting to hear an Open message from its neighbor. The OPEN message is checked for correctness. In case of errors, such as a bad version number or an unacceptable AS, the system sends an error NOTIFICATION message and goes back to Idle. If there are no errors, BGP starts sending KEEPALIVE messages and resets the KEEPALIVE timer. Meanwhile, it will enter the OpenConfirm state.

OpenConfirm： In this state, the BGP process waits for a Keepalive or Notification message. If a Keepalive is received, the state transitions to Established. If a Notification is received, or a TCP disconnect is received, the state transitions to Idle.If the Hold timer expires, an error is detected, or a Stop event occurs, a Notification is sent to the neighbor and the BGP connection is closed, changing the state to Idle.

Established：This is the final stage in the neighbor negotiation. At this stage, BGP starts exchanging UPDATE packets with its peers. The Hold Timer restarts at the receipt of an UPDATE or KEEPALIVE message. If the system receives any NOTIFICATION message (if an error has occurred), the state falls back to Idle. The UPDATE messages are checked for errors, such as missing attributes, duplicate attributes, and so on. If errors are found, a NOTIFICATION message is sent to the peer, and the state falls back to Idle. If the Hold Timer expires, or a disconnect notification is received from the transport protocol, or a Stop event, the system falls back to the Idle state.

By using"display bgp peer" command, we always observe these 2 states: Active and Established.

The neighbor state Active indicates that TCP connection fail to establish. This could be due to the inability of a neighbor to reach the IP address of its peer or mistakes in the configuration. As a result, the neighbors not able to exchange the routing information.

BGP Routing Information Base

- IP Routing Table (IP-RIB)
  - ⇨ Entire routing information base, includes all the IP routing information.
- BGP Routing Table (Loc-RIB)
  - ⇨ BGP routing information base, it includes the routes that will be used by the local BGP speaker.
- Neighbor List
  - ⇨ The list of BGP peers
- Adj-RIB-In
  - ⇨ Store the routes that are received from other BGP speakers. Adj-RIBs-In contain unprocessed routing information that has been advertised to the local BGP speaker by its peers
- Adj-RIB-Out
  - ⇨ Stores the routes that will be advertised to other BGP speakers.

  HUAWEI

•IP Routing Table (IP-RIB)

Entire routing information base, includes all the IP routing information.

•BGP Routing Table (Loc-RIB)

BGP routing information base, it includes the routes that will be used by the local BGP speaker.

•Neighbor List

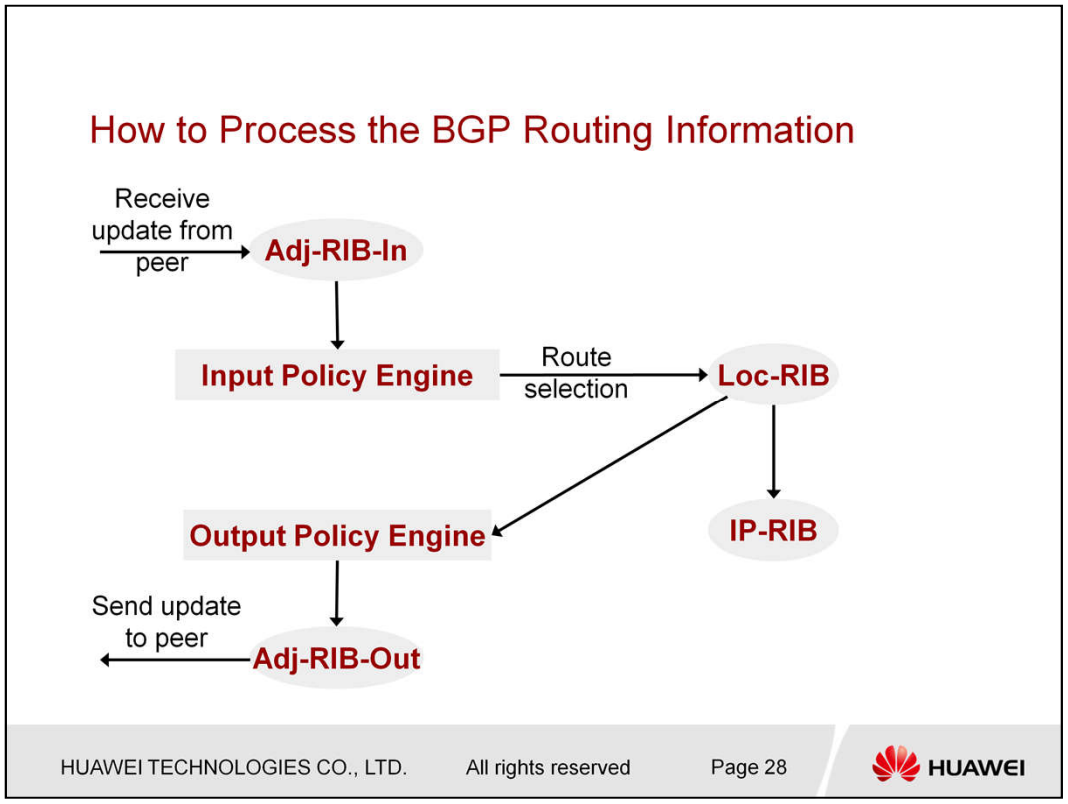
The list of BGP peer

•Adj-RIB-In

Store the routes that are received from other BGP speakers. Adj-RIBs-In contain unprocessed routing information that has been advertised to the local BGP speaker by its peers.

Adj-RIB-Out

Store the routes that will be advertised to other BGP speakers.



Upon receiving the update data packet from the peer, the BGP speaker will store this update data packet into the BGP Routing Information Base (RIB) for route selection. The Adj-RIB-In which is associated with each individual peer of the BGP speaker is specified. Then, the update data packets are being manipulated or filtered by the Input Policy Engine associated with the peer. The router will then execute the route selection criteria and a best path is selected for every IP prefix. The Loc-RIB contains only the preferred routes that have been selected as the best path to each available destination. This best route is then sent to the local IP-RIB, and it is under the installation consideration.

When multiple routes to the same IP subnet exist, the best path and all the equal cost paths are sent to the IP-RIB for consideration. In addition to the best route received from the BGP peer, LOC-RIB also contains route that the local router originates (if configured to do so) about the network inside its autonomous systems. This is how an AS advertises its internal networks to the outside world. Before the contents of the LOC-RIB are advertised to other BGP peers, it must be processed by the output policy engine. Only the route that has been filtered by the output policy engine can be installed in the Adj-RIBOut.

Summary

- How are neighbors discovered in BGP?
- What is the underlying protocol and port number used by BGP?
- What are the four BGP message types, and how is each one used?

Q. How to discover the neighbor in BGP?

A. BGP does not use any neighbor discovery mechanism. Therefore, we have to specify all the neighbors manually.

2. Name the underlying protocol and the port number that BGP used.

A. BGP uses TCP port 179.

3. What are the four BGP message types, and how is each one used?.

A: OPEN: The Open message includes BGP version number, AS number of the originator and so on. After the TCP session is established, both neighbors send Open messages to each other and determine whether the neighbor relationship can be formed.

KEEPALIVE: Keepalive message is exchanged periodically to maintain the neighbor relationship. It is used to verify the connectivity of the peer.

NOTIFICATION: Notification message is the error checking mechanism used in BGP. BGP speaker will send the notification message when an error occurs. This will always cause the BGP connection to close.

UPDATE: Among the 4 message types, update message is the most important message in the BGP system. Update message is used to exchange the routing information between the peers. It consists of all the information used by BGP to form the loop-free network structure.

It comprises of Network Layer Reachability Information (NLRI), path attributes and withdrawn route fields.

Thank You

www.huawei.com