Network Layer Routing - V
Border Gateway Protocol -4

Yatindra Nath Singh

ynsingh@ieee.org

Dept. Of Electrical Engineering

IIT Kanpur-208016

22 August 2001

Border Gateway Protocol - 4

BGP-4 (RFC 1771 - http://www.ietf.org/)

- intended to be used for routing between Autonomous systems.
- Autonomous systems - network domains administered by single entity.
- previous two protocols were designed to work within AS (as IGP)
- BGP is exterior gateway protocol (EGP).

What is expected of EGP?

- routing between ASs - depends on relation between enitities administring AS's. (e.g., traffic originating in one country to outside world, should not go through neighbouring country.)
- A network administration may not allow transit traffic.
- Routing policies are important factors.

- BGP can only support policies with ”hop-by-hop” paradigm.
  - AS A send packets via AS B. A cannot define a routing policy which are different then for packets for same destination from B.
  - Based on policy in A, only next hop can be decided. Afterwards, policies of other routers will decide.
  - For policies which cannot be supported, *strict source routing* need to be used. (not supported by BGP)

Operation of BGP

- BGP protocol uses reliable transport connection for communication between BGP routers.
- In internet world, TCP is used.
- to establish the connections, TCP port 179 is used.
- After establishing the connection, messages exchanged to open and confirm parameters.
- initial data flow - whole routing table.
- increamental updates are sent afterwards.
- periodic refresh of whole routing table - not done in BGP.

- BGP routers should keep current version of routing tables of all of its peers for duration of connection.
- to keep the connection open, keepalive messages are periodically sent.
- in case of error or special conditions - notification messages are sent.
  - for errors, notification messages are sent, thereafter connections are closed.

BGP speakers need not be router.

- A host can exchange routing information with router using EGP or IGP.
- Then this can exchange information with BGP speakers in other ASs.

When an AS has multiple BGP speakers, and providing transit service.

- consistent view of routing withing AS, must.
- BGP speakers, should maintain connection among themselves to provide consistent view in AS of routing to external destinations.

BGP speakers from different ASs - external links

BGP speakers from same AS - internal links

A peer in other AS - external peer

A peer in same AS - internal peer

Routes

- pair of BGP speakers advertise routes to each other in update messages.
- each update message has IP address of network in Network Layer Reachability Information (NLRI) field of update message, and path to destination as attribute.
- Routes are stored in Routing Information Bases (RIBs).

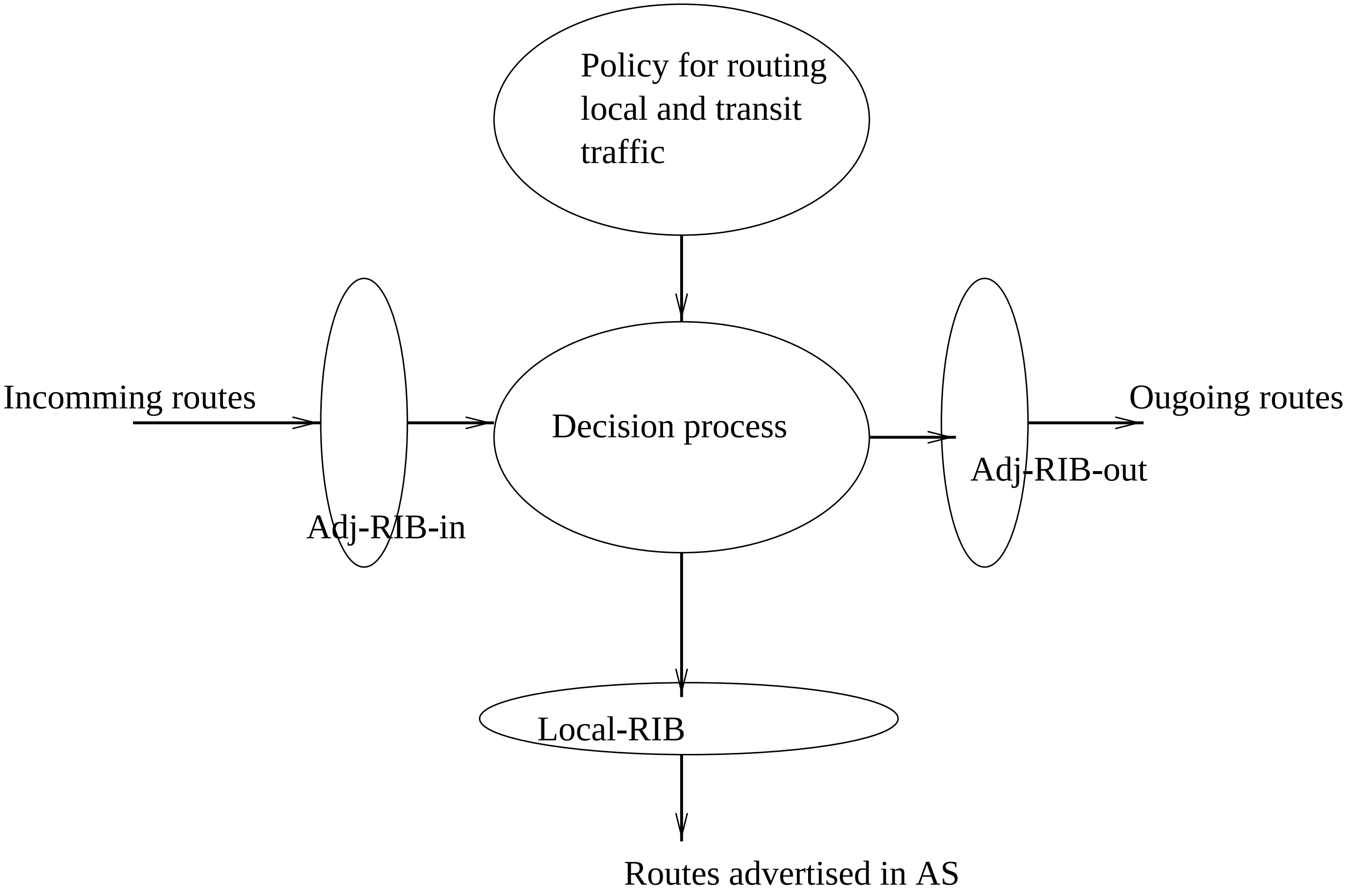
Three kind of RIBs.

- Adj-RIB-in : routes received from adjacent BGP speakers. These are available as input to decision process.

- Loc-RIB : After running decision process (applying local policies) on routes in Adj-RIB-in, these routes are obtained. Used for locally originated traffic.
- Adj-RIB-out : routes which are advertised to other adjacent BGP speakers. Policies for transit traffic and locally destined traffic are used to generate the routes from Adj-RIB-in for inclusion in Adj-RIB-out.
- BGP speaker can add or modify the path attributes before advertising a route to other BGP speakers.

BGP speakers can inform peers of non-availability of an existing routes

- IP prefix denoting the destination of previously advertised route - denoted in WITHDRAWN ROUTES field in update message.
- Replacement route with same NLRI is advertised.
- BGP speaker - BGP speaker connection can be closed. Removes all the routes which speakers hav advertised to each other.



The protocol is essentially like distance vector routing.

But there is not meteric. Instead of meteric path to the destination is exchanged.

Message format

- messages are processed only when they are completely received.
- maximum message size - 4096 octets.
- smallest message size - 19 octes, message header without data.

Message header

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+                                                               +
|                                                               |
+                                                               +
|                           Marker                              |
+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Length               |      Type     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Marker

- 16 octet field.
- contains a value which reciever of message can predict.
- for message type OPEN and it not carrying optional parameter for authentication. marker is all 'one's.
- The value of marker is predictable due calculation specified as part of authentication information.
- It is used to find loss of synchronization between to peers, to authenticate incomming BGP messages.

Length

- 2 octets, unsigned integer.
- indicates the length of message (max 4096 and min 19).
- no padding or extra data after message is allowed, length is smallest value of length required for message.

Type

- one octet indicates type of message
  - 1 : OPEN
  - 2 : UPDATE
  - 3 : NOTIFICATION
  - 4 : KEEPALIVE

OPEN message format

- After opening TCP connection, first message is OPEN message.
- OPEN message is responded by KEEPALIVE message as confirmation.
- After this, KEEPALIVE, UPDATE and NOTIFICATION messages are exchanged.

In addition to BGP header, OPEN has the following.

```
 0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+
|    Version    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     My Autonomous System      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Hold Time           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         BGP Identifier                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Opt Parm Len  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|                       Optional Parameters                     |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- current version number - 4.
- My autonomous system : 2 octets indicating autonomous system of sender.
- Hold time : Proposed value of hold time. Hold time is maximum time between KEEPALIVE and/or UPDATE messages. The receiver send the minimum of its configured value and received value of Hold Time. Should be either zero or minimum 3 secs.
- BGP identifier : This is basically set to IP address assigned to sender.
- Opt Param Len : optional parameter length gives total number of octets used for optional parameters in the message. 0 means no optional parameters.

- Optional Parameters : have Parm. type (1 octet), Parm. length (1 octet), and Parameter value (variable length - at most 255 octets).

```
 0                   1
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...
|  Parm. Type   | Parm. Length  |  Parameter Value (variable)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...
```

Optional parameter types

- 1 : authentication information

minimum length of OPEN message is 29 octets including message header.

UPDATE message format (after header)

```
+-----------------------------------------------------+
|   Unfeasible Routes Length (2 octets)               |
+-----------------------------------------------------+
|  Withdrawn Routes (variable)                        |
+-----------------------------------------------------+
|   Total Path Attribute Length (2 octets)            |
+-----------------------------------------------------+
|    Path Attributes (variable)                       |
+-----------------------------------------------------+
|   Network Layer Reachability Information (variable) |
+-----------------------------------------------------+
```

Unfeasible Routes Length

- represents unsigned integer indicating total length of withdrawn routes field in octets.
- value of 0 indicates that no routes are being withdrawn from service; WITHDRAWN ROUTES field is not present.

Withdrawn routes

- Variable length field
- contains list of IP address prefixes for the routes being withdrawn from service.

- Each IP address prefix encoded as 2-tuple.

```
+---------------------------+
|   Length (1 octet)        |
+---------------------------+
|   Prefix (variable)       |
+---------------------------+
```

  - length : length of bits of IP address prefix. A length of 0 means prefix matching to all IP address (no prefix field).
  - Prefix : IP address prefix followed by enough 0 bits to make end of field at boundry of octet.

Total path attribute length (2 octets)

- indicates total length of path attributes fields in octets.
- value of 0 means there are not path attributes, hence not network layer reachability information. (NLRI).

Path attributes

- variable length
- 3-tuple attribute type, attribute length, attribute value.
  - Attribute type is 2 octet field

```
 0                   1
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Attr. Flags  |Attr. Type Code|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- Bit 0 of Attr. Flags - attribute is optional (1) or not (0).
- Bit 1 - transitive bit - optional attribute is transitive (1) or not (0). For Well known attributes - transitive bit (1).
- Bit 2 - Partial bit - value in optional transitive attribute partial (1) or not (0). Well known and optional non-transitive attributes bit (0).
- Bit 3 - Extended length bit - (0) attribute length is one octet (3 octet of path attribute). (1) - attribute length is two octet (3 and 4 octet of path attribute).
- lower order 4 bits - unused, ignored.

- Attribute type code : five types of attributes are defined.
  - ORIGIN (type code 1) : origin of path.
    - data octet 0 - route learned through IGP (NLRI is interior to AS).
    - data octet 1 - route learned through EGP
    - data octet 2 - route learned through some other means.
  - AS_PATH (type code 2) : 3-tuple of path segment type, path segment length, path segment value.
  - NEXT_HOP (type code 3).
  - MULTI_EXIT_DISC (type code 4).
  - LOCAL_PREF (type code 5).
  - ATOMIC_AGGREGATE (type code 6).
  - AGGREGATOR (type code 7).

- remaining is attribute value.

Network Layer Reachability Information.

- variable length field
- contains many IP prefixes.
- length of NLRI not encoded - can be computed by $length\ of\ NLRI = update\ message\ length - 23 - unfeasable\ route\ length - total\ path\ attribute\ length$
- encoded as 2-tuple

```
+---------------------------+
|   Length (1 octet)        |
+---------------------------+
|   Prefix (variable)       |
+---------------------------+
```

Conclusion

- attempt made to understand the routing, RIP-1, RIP-2, RIPng, OSPF2 and BGP4.
- More details can be seen in RFCs.
- Packet format details - not necessary for ISP operations as such.
- but helps in understanding *what is happenning in the network.*