3.5.4 Border Gateway Protocol Version 4 (BGP4)

- The Border Gateway Protocol version 4 (BGP4) is the only interdomain routing protocol used in the Internet today.
- BGP4 is based on the path-vector algorithm we described before, but it is tailored to provide information about the reachability of networks in the Internet.

Introduction

- BGP, and in particular BGP4, is a complex protocol. In this section, we introduce the basics of BGP and its relationship with intradomain routing protocols (RIP or OSPF).
- Figure 3.5.4.1 shows an example of an internet with four autonomous systems. AS2, AS3, and AS4 are stub autonomous systems; AS1 is a transient one. In our example, data exchange between AS2, AS3, and AS4 should pass through AS1.
- Each autonomous system in this figure uses one of the two common intradomain protocols, RIP or OSPF.

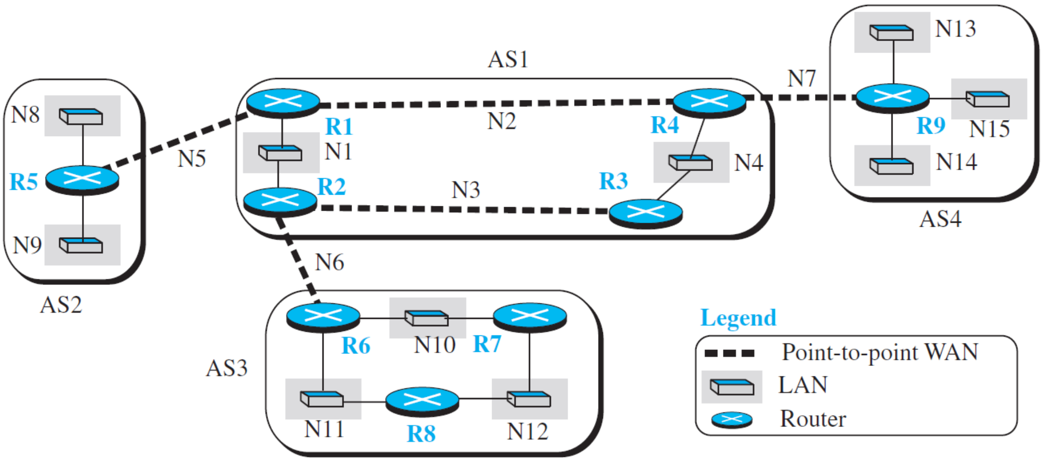


Figure 3.5.4.1 A sample internet with four ASs

- Each router in each AS knows how to reach a network that is in its own AS, but it does not know how to reach a network in another AS.
- To enable each router to route a packet to any network in the internet,
  - we first install a variation of BGP4, called external BGP (eBGP), on each border router (the one at the edge of each AS which is connected to a router at another AS).
  - We then install the second variation of BGP, called internal BGP (iBGP), on all routers.
- This means that the border routers will be running three routing protocols (intradomain, eBGP, and iBGP), but other routers are running two protocols (intradomain and iBGP).
- We discuss the effect of each BGP variation separately.

Operation of External BGP (eBGP)

- We can say that BGP is a kind of point-to-point protocol. When the software is installed on two routers, they try to create a TCP connection using the well-known port 179. In other words, a pair of client and server processes continuously communicate with each other to exchange messages. The two routers that run the BGP processes are called BGP peers or BGP speakers.

- We discuss different types of messages exchanged between two peers, but for the moment we are interested in only the update messages (discussed later) that announce reachability of networks in each AS.
- The eBGP variation of BGP allows two physically connected border routers in two different ASs to form pairs of eBGP speakers and exchange messages.
- The routers that are eligible in our example in Figure 20.24 (3.5.4.2) form three pairs:
  - R1-R5,
  - R2-R6,
  - and R4-R9.
- The connection between these pairs is established over three physical WANs (N5, N6, and N7).
- However, there is a need for a logical TCP connection to be created over the physical connection to make the exchange of information possible. Each logical connection in BGP parlance is referred to as a session. This means that we need three sessions in our example, as shown in Figure 20.25.

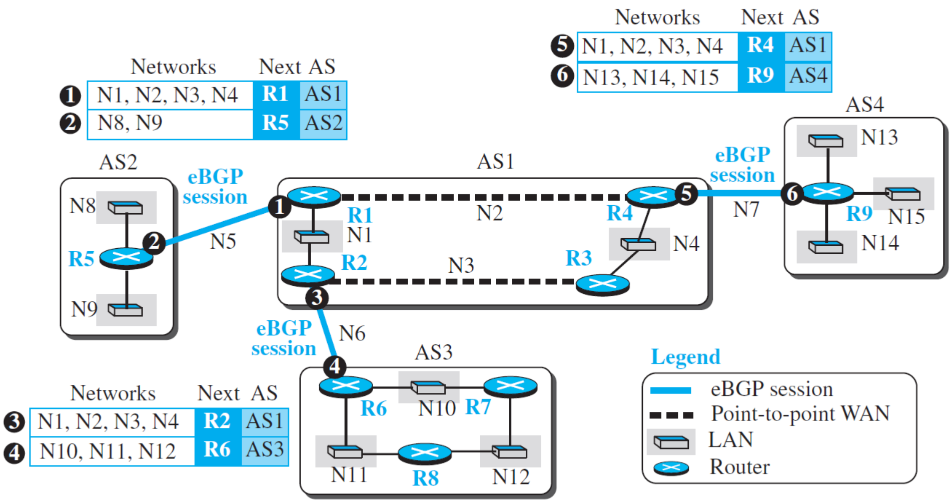


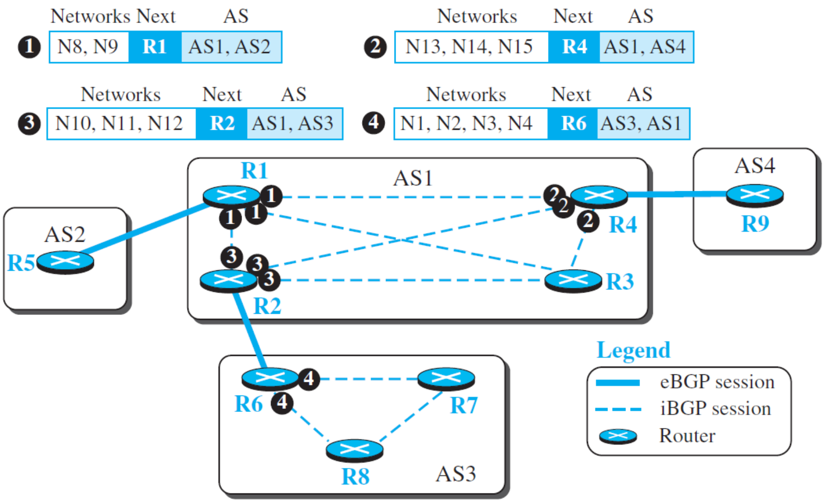
Figure 3.5.4.2 eBGP operation

- The figure also shows the simplified update messages sent by routers involved in the eBGP sessions.
- The circled number defines the sending router in each case.
  - For example,
    - Message number 1 is sent by router R1 and tells router R5 that N1, N2, N3, and N4 can be reached through router R1 (R1 gets this information from the corresponding intradomain forwarding table).Router R5 can now add these pieces of information at the end of its forwarding table. When R5 receives any packet destined for these four networks, it can use its forwarding table and find that the next router is R1.

- The reader may have noticed that the messages exchanged during three eBGP sessions help some routers to know, how to route packets to some networks in the internet, but the reachability information is not complete. There are two problems that need to be addressed:
    - 1. Some border routers do not know how to route a packet destined for nonneighbor ASs. For example, R5 does not know how to route packets destined for networks in AS3 and AS4. Routers R6 and R9 are in the same situation as R5: R6 does not know about networks in AS2 and AS4; R9 does not know about networks in AS2 and AS3.
    - 2. None of the nonborder routers know how to route a packet destined for any networks in other ASs. To address the above two problems, we need to allow all pairs of routers (border or nonborder) to run the second variation of the BGP protocol, iBGP.

Operation of Internal BGP(iBGP)

- The iBGP protocol is similar to the eBGP protocol in that it uses the service of TCP on the well-known port 179, but it creates a session between any possible pair of routers inside an autonomous system.
- However, some points should be made clear. First, if an AS has only one router, there cannot be an iBGP session. For example, we cannot create an iBGP session inside AS2 or AS4 in our internet. Second, if there are n routers in an autonomous system, there should be [n × (n − 1) / 2] iBGP sessions in that autonomous system (a fully connected mesh) to prevent loops in the system.
- In other words, each router needs to advertise its own reachability to the peer in the session instead of flooding what it receives from another peer in another session.
- Figure 3.5.4.3 shows the combination of eBGP and iBGP sessions in our internet.



❶

Networks	Next	AS
N8, N9	R1	AS1, AS2

❷

Networks	Next	AS
N13, N14, N15	R4	AS1, AS4

❸

Networks	Next	AS
N10, N11, N12	R2	AS1, AS3

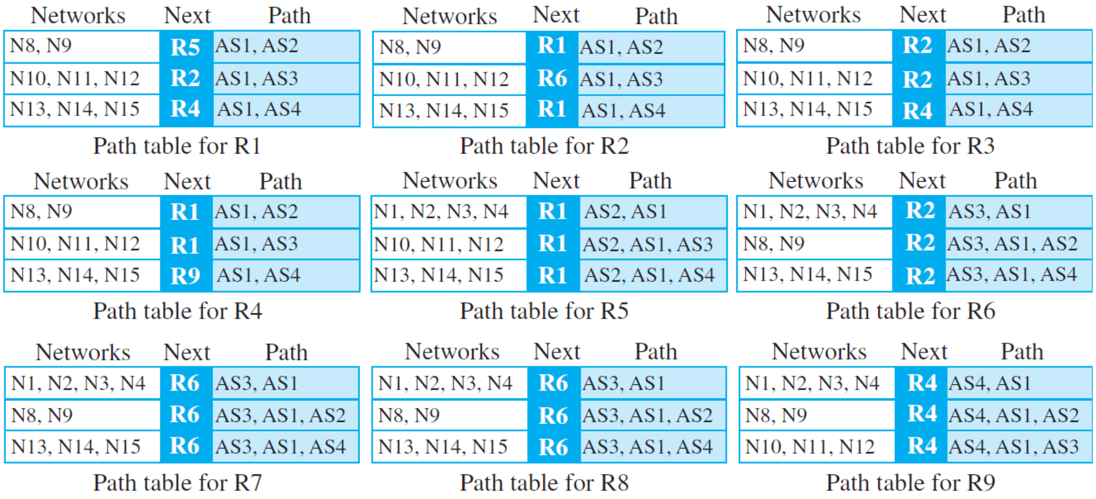
❹

Networks	Next	AS
N1, N2, N3, N4	R6	AS3, AS1

Figure 3.5.4.3 Combinations of eBGP and iBGP sessions in our internet

- Note that we have not shown the physical networks inside ASs because a session is made on an overlay network (TCP connection), possibly spanning more than one physical network as determined by the route dictated by intradomain routing protocol.
- Also note that in this stage only four messages are exchanged.

- The first message (numbered 1) is sent by R1 announcing that networks N8 and N9 are reachable through the path AS1-AS2, but the next router is R1. This message is sent, through separate sessions, to R2, R3, and R4.
- Routers R2, R4, and R6 do the same thing but send different messages to different destinations.
- The interesting point is that, at this stage, R3, R7, and R8 create sessions with their peers, but they actually have no message to send. The updating process does not stop here.
- For example, after R1 receives the update message from R2, it combines the reachability information about AS3 with the reachability information it already knows about AS1 and sends a new update message to R5. Now R5 knows how to reach networks in AS1 and AS3. The process continues when R1 receives the update message from R4.
- The point is that we need to make certain that at a point in time there are no changes in the previous updates and that all information is propagated through all ASs.At this time, each router combines the information received from eBGP and iBGP and creates what we may call a path table after applying the criteria for finding the best path, including routing policies that we discuss later.
- To demonstrate, we show the path tables in Figure 3.5.4.4 for the routers in Figure 3.5.4.1.



Networks	Next	Path
N8, N9	R5	AS1, AS2
N10, N11, N12	R2	AS1, AS3
N13, N14, N15	R4	AS1, AS4

Path table for R1

Networks	Next	Path
N8, N9	R1	AS1, AS2
N10, N11, N12	R6	AS1, AS3
N13, N14, N15	R1	AS1, AS4

Path table for R2

Networks	Next	Path
N8, N9	R2	AS1, AS2
N10, N11, N12	R2	AS1, AS3
N13, N14, N15	R4	AS1, AS4

Path table for R3

Networks	Next	Path
N8, N9	R1	AS1, AS2
N10, N11, N12	R1	AS1, AS3
N13, N14, N15	R9	AS1, AS4

Path table for R4

Networks	Next	Path
N1, N2, N3, N4	R1	AS2, AS1
N10, N11, N12	R1	AS2, AS1, AS3
N13, N14, N15	R1	AS2, AS1, AS4

Path table for R5

Networks	Next	Path
N1, N2, N3, N4	R2	AS3, AS1
N8, N9	R2	AS3, AS1, AS2
N13, N14, N15	R2	AS3, AS1, AS4

Path table for R6

Networks	Next	Path
N1, N2, N3, N4	R6	AS3, AS1
N8, N9	R6	AS3, AS1, AS2
N13, N14, N15	R6	AS3, AS1, AS4

Path table for R7

Networks	Next	Path
N1, N2, N3, N4	R6	AS3, AS1
N8, N9	R6	AS3, AS1, AS2
N13, N14, N15	R6	AS3, AS1, AS4

Path table for R8

Networks	Next	Path
N1, N2, N3, N4	R4	AS4, AS1
N8, N9	R4	AS4, AS1, AS2
N10, N11, N12	R4	AS4, AS1, AS3

Path table for R9

Figure 3.5.4.4 Finalized BGP Path tables

- For example, router R1 now knows that any packet destined for networks N8 or N9 should go through AS1 and AS2 and the next router to deliver the packet to is router R5. Similarly, router R4 knows that any packet destined for networks N10, N11, or N12 should go through AS1 and AS3 and the next router to deliver this packet to is router R1, and so on.

Injection of Information into Intradomain Routing

- The role of an interdomain routing protocol such as BGP is to help the routers inside the AS to augment their routing information. In other words, the path tables collected and organized by BPG are not used, per se, for routing packets; they are injected into intradomain forwarding tables (RIP or OSPF) for routing packets. This can be done in several ways depending on the type of AS.
- In the case of a stub AS, the only area border router adds a default entry at the end of its forwarding table and defines the next router to be the speaker router at the end of the

eBGP connection. In Figure 3.5.4.1, R5 in AS2 defines R1 as the default router for all networks other than N8 and N9. The situation is the same for router R9 in AS4 with the default router to be R4. In AS3, R6 set its default router to be R2, but R7 and R8 set their default router to be R6. These settings are in accordance with the path tables we describe in Figure 3.5.4.4 for these routers. In other words, the path tables are injected into intradomain forwarding tables by adding only one default entry.

- In the case of a transient AS, the situation is more complicated. R1 in AS1 needs to inject the whole contents of the path table for R1 in Figure 3.5.4.4 into its intradomain forwarding table. The situation is the same for R2, R3, and R4. One issue to be resolved is the cost value. We know that RIP and OSPF use different metrics. One solution, which is very common, is to set the cost to the foreign networks at the same cost value as to reach the first AS in the path. For example, the cost for R5 to reach all networks in other ASs is the cost to reach N5. The cost for R1 to reach networks N10 to N12 is the cost to reach N6, and so on. The cost is taken from the intradomain forwarding tables (RIP or OSPF).
- Figure 3.5.4.5 shows the interdomain forwarding tables. For simplicity, we assume that all ASs are using RIP as the intradomain routing protocol. The shaded areas are the augmentation injected by the BGP protocol; the default destinations are indicated as zero.

Des.	Next	Cost
N1	—	1
N4	R4	2
N8	R5	1
N9	R5	1
N10	R2	2
N11	R2	2
N12	R2	2
N13	R4	2
N14	R4	2
N15	R4	2

Table for R1

Des.	Next	Cost
N1	—	1
N4	R3	2
N8	R1	2
N9	R1	2
N10	R6	1
N11	R6	1
N12	R6	1
N13	R3	3
N14	R3	3
N15	R3	3

Table for R2

Des.	Next	Cost
N1	R2	2
N4	—	1
N8	R2	3
N9	R2	3
N10	R2	2
N11	R2	2
N12	R2	2
N13	R4	2
N14	R4	2
N15	R4	2

Table for R3

Des.	Next	Cost
N1	R1	2
N4	—	1
N8	R1	2
N9	R1	2
N10	R3	3
N11	R3	3
N12	R3	3
N13	R9	1
N14	R9	1
N15	R9	1

Table for R4

Des.	Next	Cost
N8	—	1
N9	—	1
0	R1	1

Table for R5

Des.	Next	Cost
N10	—	1
N11	—	1
N12	R7	2
0	R2	1

Table for R6

Des.	Next	Cost
N10	—	1
N11	R6	2
N12	—	1
0	R6	2

Table for R7

Des.	Next	Cost
N10	R6	2
N11	—	1
N12	—	1
0	R6	2

Table for R8

Des.	Next	Cost
N13	—	1
N14	—	1
N15	—	1
0	R4	1

Table for R9

Figure 3.5.4.5 Forwarding tables after injection from BGP

Address Aggregation

- The reader may have realized that intradomain forwarding tables obtained with the help of the BGP4 protocols may become huge in the case of the global Internet because many destination networks may be included in a forwarding table.
- Fortunately, BGP4 uses the prefixes as destination identifiers and allows the aggregation of these prefixes, as we discussed in the Unit-1. For example, prefixes 14.18.20.0/26, 14.18.20.64/26, 14.18.20.128/26, and 14.18.20.192/26, can be combined into 14.18.20.0/24 if all four subnets can be reached through one path. Even if one or two of the aggregated prefixes need a separate path, the longest prefix principle we discussed earlier allows us to do so.

Path Attributes

- In both intradomain routing protocols (RIP or OSPF), a destination is normally associated with two pieces of information: next hop and cost. The first one shows the address of the next router to deliver the packet; the second defines the cost to the final destination.
- Interdomain routing is more involved and naturally needs more information about how to reach the final destination. In BGP these pieces are called path attributes.
- BGP allows a destination to be associated with up to seven path attributes. Path attributes are divided into two broad categories: well-known and optional. A well-known attribute must be recognized by all routers; an optional attribute need not be.
    - A well-known attribute can be mandatory, which means that it must be present in any BGP update message, or discretionary, which means it does not have to be.
    - An optional attribute can be either transitive, which means it can pass to the next AS, or intransitive, which means it cannot.
- All attributes are inserted after the corresponding destination prefix in an update message (discussed later).
- The format for an attribute is shown in Figure 3.5.4.6.

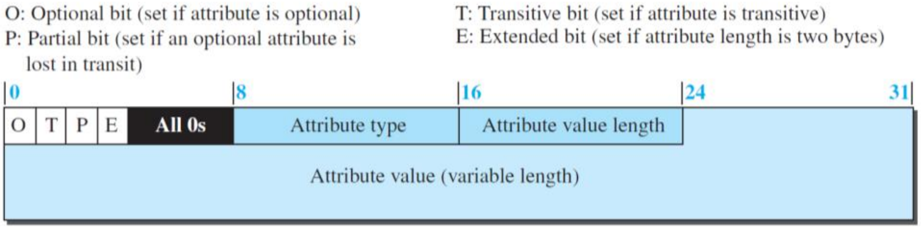

O: Optional bit (set if attribute is optional)
P: Partial bit (set if an optional attribute is lost in transit)
T: Transitive bit (set if attribute is transitive)
E: Extended bit (set if attribute length is two bytes)
0
8
16
24
31
O
T
P
E
All 0s
Attribute type
Attribute value length
Attribute value (variable length)

Figure 3.5.4.6 Format of Path Attribute

- The first byte in each attribute defines the four attribute flags (as shown in the figure). The next byte defines the type of attributes assigned by ICANN (only seven types have been assigned, as explained next). The attribute value length defines the length of the attribute value field (not the length of the whole attributes section). The following gives a brief description of each attribute.
- ORIGIN (type 1).
    - This is a well-known mandatory attribute, which defines the source of the routing information.
    - This attribute can be defined by one of the three values: 1, 2, and 3.
        - Value 1 means that the information about the path has been taken from an intradomain protocol (RIP or OSPF).
        - Value 2 means that the information comes from BGP.
        - Value 3 means that it comes from an unknown source.
- AS-PATH (type 2).
    - This is a well-known mandatory attribute, which defines the list of autonomous systems through which the destination can be reached. We have used this attribute in our examples.
    - The AS-PATH attribute, as we discussed in path-vector routing in the last section, helps prevent a loop. Whenever an update message arrives at a router that lists the current AS as the path, the router drops that path.

  - The AS-PATH can also be used in route selection.
- NEXT-HOP (type 3).
  - This is a well-known mandatory attribute, which defines the next router to which the data packet should be forwarded.
  - We have also used this attribute in our examples. As we have seen, this attribute helps to inject path information collected through the operations of eBGP and iBGP into the intradomain routing protocols such as RIP or OSPF.
- MULT-EXIT-DISC (type 4).
  - The multiple-exit discriminator is an optional intransitive attribute, which discriminates among multiple exit paths to a destination.
  - The value of this attribute is normally defined by the metric in the corresponding intradomain protocol (an attribute value of 4-byte unsigned integer).
  - For example, if a router has multiple paths to the destination with different values related to these attributes, the one with the lowest value is selected.
  - Note that this attribute is intransitive, which means that it is not propagated from one AS to another.
- LOCAL-PREF (type 5).
  - The local preference attribute is a well-known discretionary attribute. It is normally set by the administrator, based on the organization policy.
  - The routes the administrator prefers are given a higher local preference value (an attribute value of 4-byte unsigned integer).
  - For example, in an internet with five ASs,
    - the administrator of AS1 can set the local preference value of 400 to the path AS1 → AS2 → AS5, the value of 300 to AS1 → AS3 → AS5, and the value of 50 to AS1 → AS4 → AS5. This means that the administrator prefers the first path to the second one and prefers the second one to the third one. This may be a case where AS2 is the most secured and AS4 is the least secured AS for the administration of AS1. The last route should be selected if the other two are not available.
- ATOMIC-AGGREGATE (type 6).
  - This is a well-known discretionary attribute, which defines the destination prefix as not aggregate; it only defines a single destination network.
  - This attribute has no value field, which means the value of the length field is zero.
- AGGREGATOR (type 7).
  - This is an optional transitive attribute, which emphasizes that the destination prefix is an aggregate.
  - The attribute value gives the number of the last AS that did the aggregation followed by the IP address of the router that did so.

Route Selection

- So far in this section, we have been silent about how a route is selected by a BGP router mostly because our simple example has one route to a destination.

- In the case where multiple routes are received to a destination, BGP needs to select one among them.The route selection process in BGP is not as easy as the ones in the intradomain routing protocol that is based on the shortest-path tree.
- A route in BGP has some attributes attached to it and it may come from an eBGP session or an iBGP session.
- Figure 3.5.4.7 shows the flow diagram as used by common implementations.
- The router extracts the routes which meet the criteria in each step. If only one route is extracted, it is selected and the process stops; otherwise, the process continues with the next step.
- Note that the first choice is related to the LOCAL-PREF attribute, which reflects the policy imposed by the administration on the route.

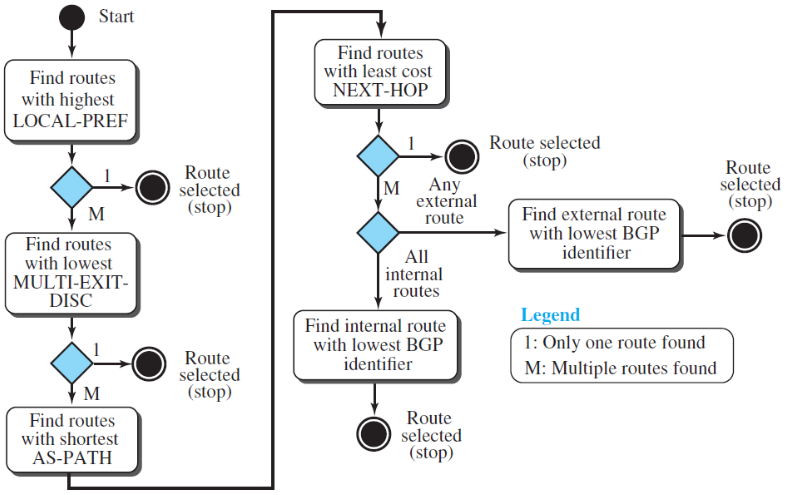

Start
Find routes with highest LOCAL-PREF
1
Route selected (stop)
M
Find routes with lowest MULTI-EXIT-DISC
1
Route selected (stop)
M
Find routes with shortest AS-PATH
Find routes with least cost NEXT-HOP
1
Route selected (stop)
M
Any external route
Find external route with lowest BGP identifier
Route selected (stop)
All internal routes
Find internal route with lowest BGP identifier
Route selected (stop)
Legend
1: Only one route found
M: Multiple routes found

Figure 3.5.4.7 Flow diagram for route selection

BGP Messages

- BGP uses four types of messages for communication between the BGP speakers across the ASs and inside an AS: open, update, keepalive, and notification (see Figure 3.5.4.8).

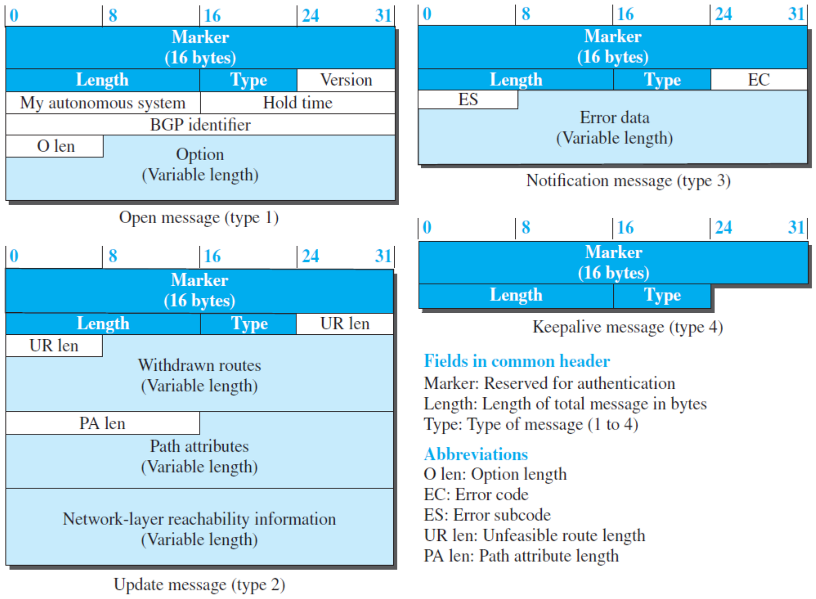

0
8
16
24
31
Marker (16 bytes)
Length
Type
Version
My autonomous system
Hold time
BGP identifier
O len
Option (Variable length)
Open message (type 1)
Marker (16 bytes)
Length
Type
EC
ES
Error data (Variable length)
Notification message (type 3)
Marker (16 bytes)
Length
Type
UR len
UR len
Withdrawn routes (Variable length)
PA len
Path attributes (Variable length)
Network-layer reachability information (Variable length)
Update message (type 2)
Marker (16 bytes)
Length
Type
Keepalive message (type 4)
Fields in common header
Marker: Reserved for authentication
Length: Length of total message in bytes
Type: Type of message (1 to 4)
Abbreviations
O len: Option length
EC: Error code
ES: Error subcode
UR len: Unfeasible route length
PA len: Path attribute length

Figure 3.5.4.8 BGP Messages

- All BGP packets share the same common header.
- Open Message.
  - To create a neighborhood relationship, a router running BGP opens a TCP connection with a neighbor and sends an open message.
- Update Message.
  - The update message is the heart of the BGP protocol.
  - It is used by a router to withdraw destinations that have been advertised previously, to announce a route to a new destination, or both.
  - Note that BGP can withdraw several destinations that were advertised before, but it can only advertise one new destination (or multiple destinations with the same path attributes) in a single update message.
- Keepalive Message.
  - The BGP peers that are running exchange keepalive messages regularly (before their hold time expires) to tell each other that they are alive.
- Notification.
  - A notification message is sent by a router whenever an error condition is detected or a router wants to close the session.

Performance

- BGP performance can be compared with RIP.
- BGP speakers exchange a lot of messages to create forwarding tables, but BGP is free from loops and count-to-infinity.
- The same weakness we mention for RIP about propagation of failure and corruption also exists in BGP.