CHAPTER 27

Border Gateway Protocol

Background

Exterior gateway protocols are designed to route between routing domains. In the terminology of the Internet (a large, international network connecting research institutions, government agencies, universities, and private businesses), a routing domain is called an *autonomous system (*AS*)*. The first exterior gateway protocol to achieve widespread acceptance in the Internet was the *Exterior Gateway Protocol* (EGP). For more information about EGP, see Chapter 26, "Exterior Gateway Protocol." Although EGP is a useful technology, it has several weaknesses, including the fact that it is more of a reachability protocol than a routing protocol.

The Border Gateway Protocol (BGP) represents an attempt to address the most serious of EGP's problems. BGP is an inter-AS routing protocol created for use in the Internet. Unlike EGP, BGP was designed to detect routing loops. BGP may be thought of as a next-generation EGP. Indeed, BGP and other inter-AS routing protocols are (slowly) replacing EGP in the Internet. BGP Version 3 is specified in *Request For Comments* (RFC) 1163.

To address the needs of the growing Internet, BGP continues to evolve. A future version of BGP will give it the ability to aggregate or summarize groups of similar routes into one route.

Technology Basics

Although BGP was designed as an inter-AS protocol, it can be used both within and between ASs. Two BGP neighbors communicating between ASs must reside on the same physical network. BGP routers within the same AS communicate with one another to ensure that they have a consistent view of the AS and to determine which BGP router within that AS will serve as the connection point to or from certain external ASs.

Some ASs are merely pass-through channels for network traffic. That is, some ASs carry network traffic that did not originate within the AS and is not destined for the AS. BGP must interact with whatever intra-AS routing protocols exist within these pass-through ASs.

BGP update messages consist of network number/AS path pairs. The AS path contains the string of ASs through which the specified network can be reached. These update messages are sent over the Transmission Control Protocol transport mechanism to ensure reliable delivery.

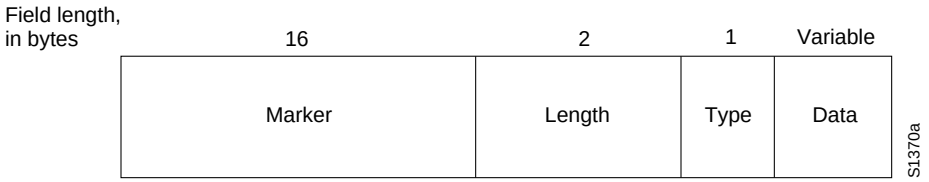
The initial data exchange between two routers is the entire BGP routing table. Incremental updates are sent out as the routing tables change. Unlike some other routing protocols, BGP does not require a periodic refresh of the entire routing table. Instead, routers running BGP retain the latest version of each peer routing table. Although BGP maintains a routing table with all feasible paths to a particular network, it advertises only the primary (optimal) path in its update messages.

The BGP metric is an arbitrary unit number specifying the degree of preference of a particular path. These metrics are typically assigned by the network administrator through configuration files. Degree of preference may be based on any number of criteria, including AS count (paths with a smaller AS count are generally better), type of link (is the link stable? fast? reliable?), and other factors.

Packet Format

The BGP packet format is shown in Figure 27-1.

Figure 27-1 BGP Packet Format



BGP packets have a common 19-byte header consisting of the following three fields:

- *Marker*—Contains a value that the receiver of the message can predict. This field is used for authentication.
- *Length*—Contains the total length of the message, in bytes.
- *Type*—Specifies the message type.

Messages

Four message types are specified in RFC 1163:

- Open
- Update
- Notification
- Keepalive

Open

After a transport protocol connection is established, the first message sent by each side is an open message. If the open message is acceptable to the recipient, a keepalive message confirming the open message is sent back. Upon successful confirmation of the open message, updates, keepalives, and notifications may be exchanged.

In addition to the common BGP packet header, open messages define several fields. The *version* field provides a BGP version number, and allows the recipient to check that it is running the same version as the sender. The *autonomous system* field provides the AS number of the sender. The *hold-time* field indicates the maximum number of seconds that may elapse without receipt of a message before the transmitter is assumed to be dead. The *authentication code* field indicates the authentication type being used (if any). The *authentication data* field contains actual authentication data (if any).

Update

BGP update messages provide routing updates to other BGP systems. Information in these messages is used to construct a graph describing the relationships of the various ASs. In addition to the common BGP header, update messages have several additional fields. These fields provide routing information by listing path attributes corresponding to each network.

BGP currently defines five attributes:

- *Origin*—Can take on one of three values: *IGP*, *EGP*, or *incomplete*. The IGP attribute means that the network is part of the AS. The EGP attribute means that the information was originally learned from the EGP. BGP implementations would be inclined to prefer IGP routes over EGP routes because EGP fails in the presence of routing loops. The incomplete attribute is used to indicate that the network is known via some other means.
- *AS path*—Provides the actual list of ASs on the path to the destination.
- *Next hop*—Provides the IP address of the router that should be used as the next hop to the networks listed in the update message.
- *Unreachable*—If present, indicates that a route is no longer reachable.
- *Inter-AS metric*—Provides a way for a BGP router to advertise its cost to destinations within its own AS. This information can be used by routers external to the advertiser's AS to select an optimal route into the AS to a particular destination.

Notification

Notification messages are sent when an error condition has been detected, and one router wishes to tell another why it is closing the connection between them. Aside from the common BGP header, notification messages have an *error code* field, an *error subcode* field, and *error data*. The error code field indicates the type of error, and can be one of the following:

- *Message header error*—Indicates a problem with the message header such as an unacceptable message length, an unacceptable marker field value, or an unacceptable message type.
- *Open message error*—Indicates a problem with an open message such as an unsupported version number, an unacceptable AS number or IP address, or an unsupported authentication code.
- *Update message error*—Indicates a problem with the update message. Examples include a malformed attribute list, an attribute list error, and an invalid next-hop attribute.
- *Hold time expired*—Indicates a hold-time expiration, after which a BGP node will be declared dead.

Keepalive

Keepalive messages do not contain any additional fields beyond those in the common BGP header. These messages are sent often enough to keep the hold-time timer from expiring.