Module 3: Routing

Syllabus

Routing Architecture – Cores, Peers and Algorithms, Routing among Autonomous system – BGP - The Scope Of A Routing Update Protocol, Determining A Practical Limit On Group Size, Autonomous System Concept, Exterior Gateway Protocols And Reachability, BGP Characteristics, BGP Functionality And Message Types, Routing Within An Autonomous System (RIP, RIPng, OSPF, IS-IS)- Introduction, Static Vs. Dynamic Interior Routes, Routing Information Protocol (RIP), Slow Convergence Problem, Solving The Slow Convergence Problem, The Disadvantage Of Using Hop Counts, Delay Metric (HELLO), Delay Metrics, Oscillation, And Route Flapping, The Open SPF Protocol (OSPF).

Original Internet Architecture And Cores

Much of our knowledge of routing and route propagation protocols has been derived from experience with the global Internet. When TCPDP was first developed, participating research sites were connected to the ARPANET, which served as the Internet backbone. During initial experiments, each site managed routing tables and installed routes to other destinations by hand. As the fledgling Internet began to grow, it became apparent that manual maintenance of routes was impractical; automated mechanisms were needed.

The routing table in a given router contains partial information about possible destinations. Routing that uses partial information allows sites autonomy in making local routing changes, but introduces the possibility of inconsistencies that may make some destinations unreachable from some sources.

Core Routers

Early Internet routers could be partitioned into two groups, a small set of core routers controlled by the Internet Network Operations Center (INOC), and a larger set of noncore routers controlled by individual groups. The core system was designed to provide reliable, consistent, authoritative routes for all possible destinations; it was the glue that held the Internet together and made universal interconnection possible. By fiat, each site assigned an Internet network address had to arrange to advertise that address to the core system. The core routers communicated among themselves, so they could guarantee that the information they shared was consistent. Because a central authority monitored and controlled the core routers, they were highly reliable.

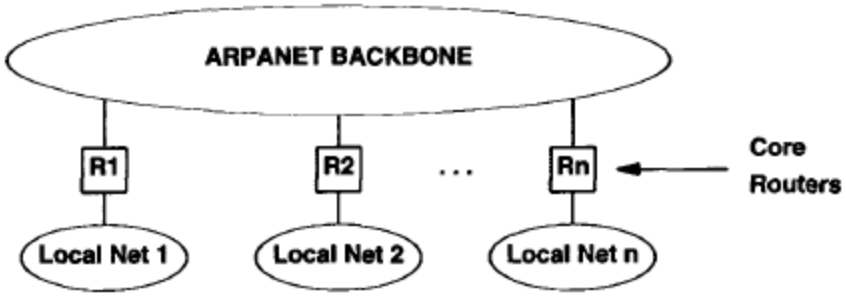


Figure 14.1 The early Internet core router system viewed as a set of routers that connect local area networks to the ARPANET. Hosts on the local networks pass all nonlocal traffic to the closest core router.

A core routing architecture assumes a centralized set of routers serves as the repository of information about all possible destinations in an internet. Core systems work best for internets that have a single, centrally managed backbone. Expanding the topology to multiple backbones makes routing complex; attempting to partition the core architecture so that all routers use default routes introduces potential routing loops.

Distance Vector (Bellman-Ford) Routing

The term distance-vectod refers to a class of algorithms routers use to propagate routing information. The idea behind distance-vector algorithms is quite simple. The router keeps a list of all known routes in a table. When it boots, a router initializes its routing table to contain an entry for each directly connected network. Each entry in the table identifies a destination network and gives the distance to that network, usually measured in hops (which will be defined more precisely later). For example, Figure 14.6 shows the initial contents of the table on a router that attaches to two networks.

Destination	Distance	Route
Net 1	0	direct
Net 2	0	direct

Figure 14.6 An initial distance-vector routing table with an entry for each directly connected network. Each entry contains the IP address of a network and an integer distance to that network.

Link-State (SPF) Routing

The main disadvantage of the distance-vector algorithm is that it does not scale well. Besides the problem of slow response to change mentioned earlier, the algorithm requires the exchange of large messages. Because each routing update contains an entry for every possible network, message size is proportional to the total number of networks in an internet. Furthermore, because a distance-vector protocol requires every router to participate, the volume of information exchanged can be enormous.

The primary alternative to distance-vector algorithms is a class of algorithms known as link state, link status, or Shortest Path First (SPF). The SPF algorithm requires each participating router to have complete topology information. The easiest way to think of the topology information is to imagine that every router has a map that shows all other routers and the networks to which they connect. In abstract terms, the routers correspond to nodes in a graph and networks that connect routers correspond to edges. There is an edge (link) between two nodes if and only if the corresponding routers can communicate directly.

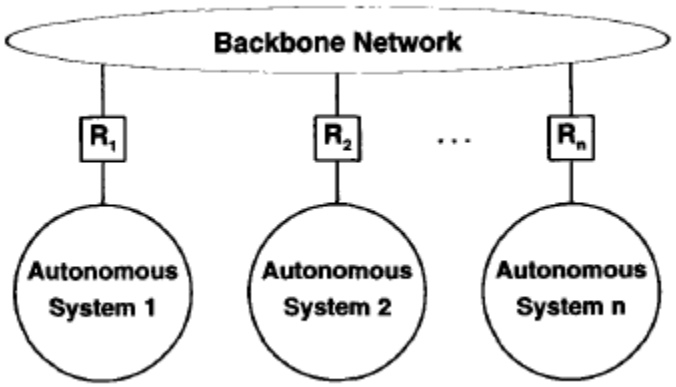
Determining A Practical Limit On Group Size

The above statement leaves many questions open. For example, what size internet is considered "large"? If only a limited set of routers can participate in an exchange of routing information, what happens to routers that are excluded? Do they function correctly? Can a router that is not participating ever forward a datagram to a router that is participating? Can a participating router forward a datagram to a non-participating router?

The answer to the question of size involves understanding the algorithm being used and the capacity of the network that connects the routers as well as the details of the routing protocol. There are two issues: delay and overhead.

Autonomous System Concept

a group of networks and routers controlled by a single administrative authority is called an autonomous system (AS). Routers within an autonomous system are free to choose their own mechanisms for discovering, propagating, validating, and checking the consistency of routes. Note that, under this definition, the original Internet core routers formed an autonomous system. Each change in routing protocols within the core autonomous system was made without affecting the routers in other autonomous systems.



Architecture of an internet with autonomous systems at backbone sites. Each autonomous system consists of multiple networks and routers under a single administrative authority.

Exterior Gateway Protocol

Computer scientists use the term Exterior Gateway Protocol (EGP) to denote any protocol used to pass routing information between two autonomous systems. Currently a single exterior protocol is used in most TCP/IP internets. Known as the Border Gateway Protocol (BGP), it has evolved through four (quite different) versions. Each version is numbered, which gives rise to the formal name of the current version: BGP-4.

When a pair of autonomous systems agree to exchange routing information, each must designate a router that will speak BGP on its behalf; the two routers are said to become BGP peers of one another. Because a router speaking BGP must communicate with a peer in another autonomous system, it makes sense to select a machine that is near the "edge" of the autonomous system. Hence, BGP terminology calls the machine a border gateway or border router. Figure 15.4 illustrates the idea.

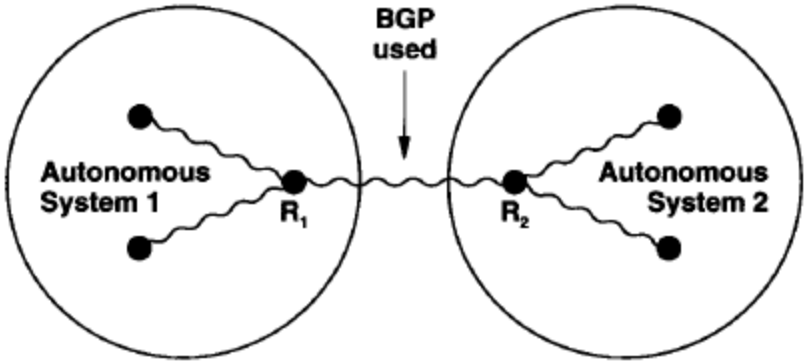


Figure 15.4 Conceptual illustration of two routers, R_1 and R_2, using BGP to advertise networks in their autonomous systems after collecting the information from other routers internally. An organization using BGP usually chooses a router that is close to the outer "edge" of the autonomous system.

In the figure, router R, gathers information about networks in autonomous system I and reports that information to router R2 using BGP, while router R2 reports information from autonomous system 2.

BGP Characteristics

BGP is unusual in several ways. Most important, BGP is neither a pure distance vector protocol nor a pure link state protocol. It can be characterized by the following:

1) *Inter-Autonomous System Communication*. Because BGP is designed as an exterior gateway protocol, its primary role is to allow one autonomous system to communicate with another.

2) *Coordination Among Multiple BGP Speakers*. If an autonomous system has multiple routers each communicating with a peer in an outside autonomous system, BGP can be used to coordinate among routers in the set to guarantee that they all propagate consistent information.

3) *Propagation Of Reachability Information*. BGP allows an autonomous system to advertise destinations that are reachable either in or through it, and to learn such information from another autonomous system.

4) *Next-Hop Paradigm*. Like distance-vector routing protocols, BGP supplies next hop information for each destination.

5) *Policy Support*. Unlike most distance-vector protocols that advertise exactly the routes in the local routing table, BGP can implement policies that the local administrator chooses. In particular, a router running BGP can be configured to distinguish between the set of destinations

reachable by computers inside its autonomous system and the set of destinations advertised to other autonomous systems.

6) *Reliable Transport*. BGP is unusual among protocols that pass routing information because it assumes reliable transport. Thus, BGP uses TCP for all communication.

7) *Path Information*. In addition to specifying destinations that can be reached and a next hop for each, BGP advertisements include path information that allows a receiver to learn a series of autonomous systems along a path to the destination.

8) *Incremental Updates*. To conserve network bandwidth, BGP does not pass full information in each update message. Instead, full information is exchanged once, and then successive messages carry incremental changes called deltas.

9) *Support For Classless Addressing*. BGP supports CIDR addresses. That is, rather than expecting addresses to be self-identifying, the protocol provides a way to send a mask along with each address.

10) *Route Aggregation*. BGP conserves network bandwidth by allowing a sender to aggregate route information and send a single entry to represent multiple, related destinations.

11) *Authentication*. BGP allows a receiver to authenticate messages (i.e., verify the identity of a sender).

BGP Functionality And Message Types

BGP peers perform three basic functions. The first function consists of initial peer acquisition and authentication. The two peers establish a TCP connection and perform a message exchange that guarantees both sides have agreed to communicate.

The second function forms the primary focus of the protocol - each side sends positive or negative reachability information. That is, a sender can advertise that one or more destinations are reachable by giving a next hop for each, or the sender can declare that one or more previously advertised destinations are no longer reachable.

The third function provides ongoing verification that the peers and the network connections between them are functioning correctly.

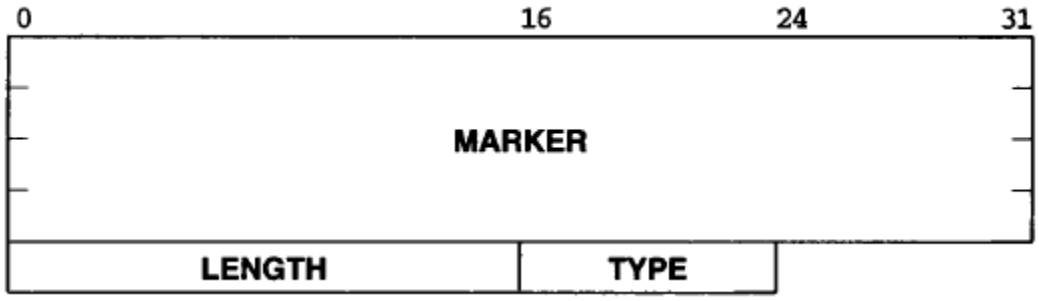
To handle the three functions described above, BGP defines four basic message types. Figure 15.5 contains a summary.

Type Code	Message Type	Description
1	OPEN	Initialize communication
2	UPDATE	Advertise or withdraw routes
3	NOTIFICATION	Response to an incorrect message
4	KEEPALIVE	Actively test peer connectivity

Figure 15.5 The four basic message types in BGP-4.

BGP Message Header

Each BGP message begins with a fmed header that identifies the message type.



The 16-octet MARKER field contains a value that both sides agree to use to mark the beginning of a message. The Zoctet LENGTH field specifies the total message length measured in octets. The minimum message size is 19 octets (for a message type that has no data following the header), and the maximum allowable length is 4096 octets. Finally, the 1-octet TYPE field contains one of the four values for the message type

BGP OPEN Message

As soon as two BGP peers establish a TCP connection, they each send an OPEN message to declare their autonomous system number and establish other operating parameters. In addition to the standard header, an OPEN message contains a value for a hold timer that is used to specify the maximum number of seconds which may elapse between the receipt of two successive messages. Figure 15.7 illustrates the format.

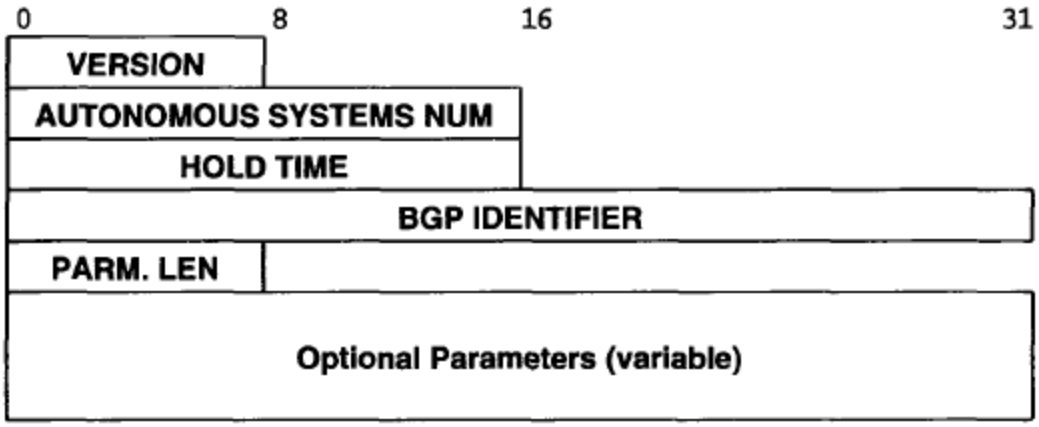


Figure 15.7 The format of the BGP OPEN message that is sent at startup. These octets follow the standard message header.

Most fields are straightforward. The VERSION field identifies the protocol version used (this format is for version 4). Recall that each autonomous system is assigned a unique number. Field AUTONOMOUS SYSTEMS NUM gives the autonomous system number of the sender's system. The HOLD TIME field specifies a maximum time that the receiver should wait for a message from the sender. The receiver is required to implement a timer using this value. The timer is reset each time a message arrives; if the timer expires, the receiver assumes the sender is no longer available (and stops forwarding datagrams along routes learned from the sender).

Field BGP IDENTIFIER contains a 32-bit integer value that uniquely identifies the sender. If a machine has multiple peers (e.g., perhaps in multiple autonomous systems), the machine must use the same identifier in all communication. The protocol specifiesthat the identifier is an IP address. Thus, a router must choose one of its IP addresses to use with all BGP peers.

BGP UPDATE Message

Once BGP peers have created a TCP connection, sent OPEN messages, and acknowledged them, the peers use UPDATE messages to advertise new destinations that are reachable or to withdraw previous advertisements when a destination has become unreachable. Figure 15.8 illustrates the format of UPDATE messages.

As the figure shows, each UPDATE message is divided into two parts: the first lists previously advertised destinations that are being withdrawn, and the second specifies new destinations being

advertised. As usual, fields labeled variable do not have a fixed size; if the information is not needed for a particular UPDATE, the field can be omitted from the message. Field WITHDRAWN LEN is a 2-octet field that specifies the size of the Withdrawn Destinations field that follows. If no destinations are being withdrawn, WlTHDRAWN LEN contains zero. Similarly, the PATH LEN field specifies the size of the Path Attributes that are associated with new destinations being advertised. If there are no new destinations, the PATH LEN field contains zero.

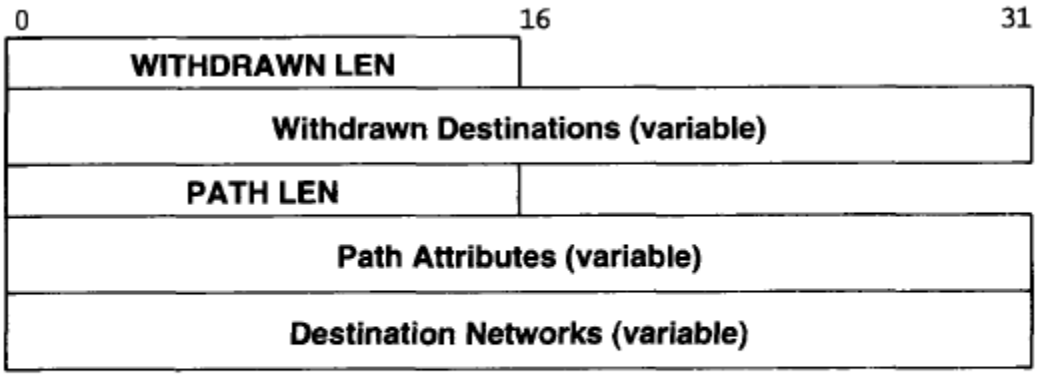


Figure 15.8 BGP UPDATE message format in which variable size areas of the message may be omitted. These octets follow the standard message header.

BGP Path Attributes

The additional information is contained in the Path Attributes field of an update message. A sender can use the path attributes to specify: a next hop for the advertised destinations, a list of autonomous systems along the path to the destinations, and whether the path information was learned from another autonomous system or derived from within the sender's autonomous system