4.5.2 Inter-Autonomous System Routing: BGP

The Border Gateway Protocol version 4, specified in RFC 1771 (see also RFC 1772; RFC 1773), is the de facto standard interdomain routing protocol in today’s Internet. It is commonly referred to as BGP4 or simply as BGP. As an inter-autonomous system routing protocol, it provides for routing among autonomous systems (ASs). Figure 4.5.2-new2 illustrates the intra-AS and inter-AS protocols used within AS2 – BGP is used to determine routes to destinations that are outside AS2, while OSPF is used to determine routes to destinations within AS2.

While BGP has the same general flavor as the distance vector protocol that we studied in Section 4.2, it is more appropriately characterized as a path-vector protocol. This is because neighboring BGP routers, known as BGP peers, exchange detailed path information (such as the list of ASs on a path to a given destination) rather than cost information. As in the generic distance vector protocol, BGP is a distributed protocol in that BGP routers only communicate with those BGP routers to which they are directly attached. Global information about routes to distant destinations propagates in an AS-by-AS manner via the exchange of BGP routing information between pairs of directly-connected BGP peers. We also note that BGP routes to destination networks (in the addressing sense), rather than to specific hosts or routers. Once a datagram reaches the destination network, that network’s intra-AS system routing will then be used to route the datagram to its final destination. In BGP, an autonomous system is identified by its globally unique autonomous system number (ASN) [RFC 1930]. (Technically, not every AS has an ASN. In particular, a so-called stub AS that only carries traffic for which it is a source or destination will not typically have an ASN; we ignore this technicality in our discussion below in order to see the forest from the trees). AS numbers, like IP addresses, are assigned by ICANN regional registries [ICANN 2002].

At the heart of BGP are route advertisements. A route advertisement is sent from one BGP peer to another BGP peer over a point-to-point connection. An advertisement consists of a CIDRized destination network address (e.g., 128.119.40/24) and a set of attributes associated with the path to that destination network. Two of the most important attributes are the PATH attribute (an explicit list of all ASs on the path to the specified destination network), and the identity of the NEXT-HOP router along that path to the destination network. For a full discussion route attributes, see [Griffin 2002; Stewart 1999; Halabi 2000].

BGP operation revolves around three activities, all involving route advertisements:

- **Receiving and filtering route advertisements from directly attached neighbor(s).** A BGP router will receive route advertisements from a BGP peer. We can think of a received BGP route advertisement as a “promise” A BGP peer advertising a route to a destination AS promises that if a neighboring AS forwards it a datagram destined to that destination AS then it will be able to forward that datagram further along on a path towards that destination. A BGP router can also filter (drop) received route advertisements. For example, a BGP router will ignore advertisements that contain its own AS number in the AS-PATH, since that path would result in a routing loop if used. Because the full AS path is specified, a network administrator can exercise significant control over where its datagrams are routed. For example, it is possible for an AS such as Hatfield.net to implement a policy such as "traffic from my AS should not cross the AS McCoy.net." (The Hatfields and the McCoys are two famous feuding families in the U.S.)

- **Route selection.** A BGP router may receive several route advertisements to the same destination AS and must choose which route it will use from among those advertised. The destination AS and next hop for the chosen path must then be installed in the router’s forwarding tables (e.g., as illustrated earlier in Figure 4-12). A BGP router may know many routes to a given destination, but will typically only install a single next-hop router for that destination in the forwarding table.
But how does a BGP router choose from among multiple paths? BGP makes a clear distinction between routing mechanism and routing policy. In particular, BGP does not specify how an AS should choose a path from among those advertised. This is a policy decision that is left up to the AS’s network administrator. A network administrator can specify so-called local preferences, e.g., indicating that routing via neighboring AS A is always to be preferred over routing via neighboring AS B, when there is a choice. In the absence of a local preference, the route selected is often the route with the shortest AS path (i.e., that crosses the smallest number of ASs on the path to the destination). A discussion of BGP best path selection algorithms can be found at [Cisco BGP 2002].

- Sending route advertisements to neighbors. Just as a BGP router will receive route advertisements from its neighbors, so too will it advertise routes to its neighbors. Once again, BGP provide a mechanism, but not a policy, for such advertisement. This allows the network administrator a significant degree of control over the traffic that will be routed into its network. Returning to our Hatfield/McCoy example, it is easy for the Hatfield’s to prevent the McCoys from directly sending traffic through the Hatfield’s network. For example, if the McCoys are immediate neighbors of the Hatfields, the Hatfields could simply not advertise any routes to the McCoys that contain the Hatfield network. But restricting traffic by controlling an AS’s route advertisement can only be partially effective. For example, if the Joneses are between the Hatfields and the McCoys, and the Hatfields advertise routes to the Joneses that pass through the Hatfields, then the Hatfields cannot prevent (using BGP mechanisms) the Joneses from advertising these routes to the McCoys.

Let’s illustrate some of the basic concepts of BGP route advertisement with a slightly more realistic example that the Hatfield’s and the McCoy’s. Figure 4.5-BGPnew shows six interconnected autonomous systems: A, B, C, W, X, Y. It is important to note that A, B, C, W, X, and Y are networks not routers. Let’s assume that autonomous systems W, X, and Y are stub networks and that A, B, and C are backbone provider networks. All traffic entering a stub network must be destined for that network, and all traffic leaving a stub network must have originated in that network. W and Y are clearly stub networks. X is a so-called multi-homed stub network, since it is connected to the rest of the network via two different providers (a scenario that is becoming increasingly common in practice). However, like W and Y, X itself must be the source/destination of all traffic leaving/entering X. But how will this stub network behavior be implemented and enforced? How will X be prevented from forwarding traffic between B and C? This can be easily accomplished by controlling the manner in which BGP routes are advertised. In particular, X will function as a stub network if it advertises (to its neighbors B and C) that it has no paths to any other destinations except itself. That is, even though X may know of a path, say XCY, that reaches network Y, it will not advertise this path to B. Since B is unaware that X has a path to Y, B would never forward traffic destined to Y (or C) via X. This simple example illustrates how selective route advertisement can be used to implement customer/provider routing relationships.
Let's next focus on a provider network, say AS B. Suppose that B has learned (from A) that A has a path AW to W. B can thus install the route BAW into its routing information base. Clearly B also wants to advertise the path BAW to its customer, X, so that X knows that it can route to W via B. But should B advertise the path BAW path to C? If it does so, then C could route traffic to W via CBAW. If A, B, and C are all backbone providers, than B might rightly feel that it should not have to shoulder the burden (and cost!) of carrying transit traffic between A and C. B might rightly feel that it is A and C’s job (and cost!) to make sure that C can route to/from A’s customers via a direct connection between A and C. There are currently no official "standards" that govern how backbone ISPs route amongst themselves. However, a rule of thumb followed by commercial ISPs is that any traffic flowing across an ISP’s backbone network must have either a source or a destination (or both) in a network that is a customer of that ISP; otherwise the traffic would be getting a “free ride” on the ISP’s network. Individual peering agreements (that would govern questions such as those raised above) are typically negotiated between pairs of ISPs and are often confidential; [Huston 199a] provides an interesting discussion of peering agreements.

Having outlined a few of the many policy issues associated with BGP, we can return to the more solid technical ground of the mechanisms within BGP. BGP peers communicate using the TCP protocol and port number 179. TCP thus provides for reliable and congestion-controlled message exchange between two BGP peers. In contrast, recall that we earlier saw that two RIP peers, e.g., the routed’s in Figure 4.31, communicate via unreliable UDP and that OSPF uses its own protocol to exchange OSPF messages. The BGP protocol defines the four types of messages: OPEN, UPDATE, NOTIFICATION, and KEEPALIVE.

• **OPEN.** When a BGP gateway wants to first establish contact with a BGP peer (for example, after the gateway itself or a connecting link has just been booted), an OPEN message is sent to the peer. The OPEN message allows a BGP gateway to identify and authenticate itself, and provide timer information. If the OPEN is acceptable to the peer, it will send back a KEEPALIVE message.

• **UPDATE.** A BGP gateway uses the UPDATE message to advertise a path to a given destination to the BGP peer. The UPDATE message can also be used to withdraw a path that had previously been advertised (that is, to tell a peer that a path that it had previously advertised is no longer valid). A BGP path is considered to be valid until it is explicitly withdrawn.

• **KEEPALIVE.** This BGP message is used to let a peer know that the sender is alive but that the sender doesn’t have other information to send. It also serves as an acknowledgment to a received OPEN message.

• **NOTIFICATION.** This BGP message is used to inform a peer that an error has been detected (for example, in a previously transmitted BGP message) or that the sender is about to close the BGP session.

Our discussion thus far has focussed exclusively on the use of BGP between routers in different ASs – a form of BGP more precisely known as External-BGP (E-BGP). There is another form of BGP, known as Internal-BGP (I-BGP), which is used to distribute routing information to routers within the AS regarding destination ASs outside of the AS. To see the need for an I-BGP, consider Figure 4.5.2 again. How should router R5 obtain routing information for distant ASs such as AS2 and AS3? We saw earlier that it is possible to specify a default route within an intra-AS routing protocol such as RIP, so that datagrams not destined to an network explicitly listed in the forwarding table will be forwarded along the default route. Another option is to use the I-BGP protocol to distribute routing information within the AS regarding remote ASs. I-BGP and E-BGP share the same message and path attribute formats, but they have several important differences. I-BGP routers within an AS are all logically connected to each other, i.e., all I-BGP routers with an AS are considered neighbors of each other. I-BGP routers are restricted also in how they can advertise paths that they learn from other I-BGP (as

As noted above, BGP is the de facto standard for inter-AS routing for the public Internet. BGP is used, for example, at the major network access points (NAPs) where major Internet carriers connect to each other and exchange traffic. To see the contents of today’s (less than two hours out-of-date) BGP routing table (large!) at one of the major NAPs in the U.S. see [IPMA 2002]. Statistics about the size and characteristics of BGP routing table are presented in [Huston 2001].

This completes our brief introduction of BGP. Although BGP is complex, it plays a central role in the Internet. We encourage readers to see the references [Griffin 2002; Stewart 1999; Labovitz 1997; Halabi 2000 Huitema 1998] to learn more about BGP.

Comment: This whole section (4.5.2) on BGP has been revised.