Internet Routing Architectures

Second Edition

The definitive BGP resource

ciscopress.com

Sam Halabi

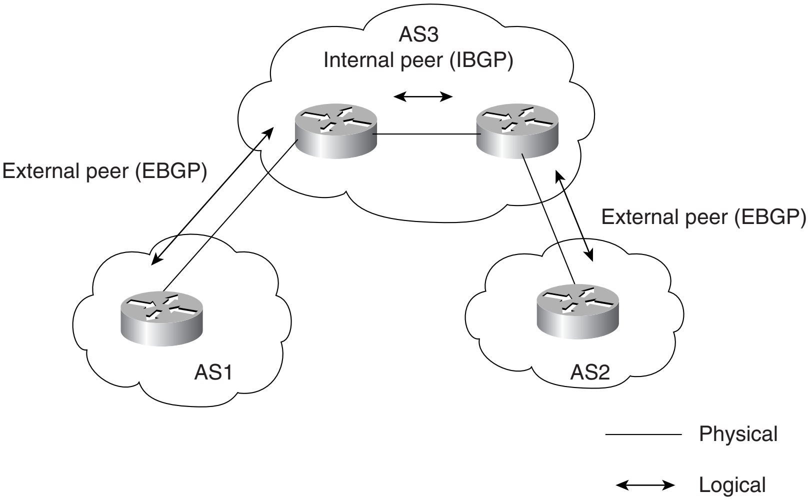
Internet Routing Architectures, Second Edition

Sam Halabi with Danny McPherson

Cisco Press

Cisco Press
800 East 96th Street
Indianapolis, IN 46240 USA

Figure 6-1 *Internal and External BGP Implementations*



Upon neighbor session establishment and during the OPEN message exchange negotiation, peer routers compare AS numbers and determine whether they are peers in the same AS or in different ASs. The difference between EBGP and IBGP manifests itself in how each peer would process the routing updates coming from the other peer and in the way different BGP attributes are carried on external connections versus internal connections.

The neighbor negotiation process is mainly the same for internal and external neighbors as far as building the TCP connection at the transport level. It is essential to have IP connectivity between the two neighbors for the transport session to be established. IP connectivity must be achieved via a protocol different from BGP; otherwise, the session will be in a race condition, an example of which follows:

> Neighbors can reach one another via some Interior Gateway Protocol (IGP), the BGP session is established, and BGP messages are exchanged. The IGP connection goes away for some reason, but the BGP TCP session is still up because neighbors can still reach each other via BGP. Eventually, the session will go down because the BGP session cannot depend on BGP itself for neighbor connectivity—the underlying substrate provides NEXT_HOP reachability. Another example is if a route more specific than that used to established the peer connection is learned via BGP.

Most commonly for IBGP peering sessions, an Interior Gateway Protocol (IGP) or static route can be configured to achieve IP connectivity. In essence, a ping packet, containing a source IP address (the IP address of one BGP peer) and a destination IP address (the IP address of the second peer), must succeed for a transport session to initiate. Generally, for external BGP sessions, a route through a directly connected interface establishes IP reachability.

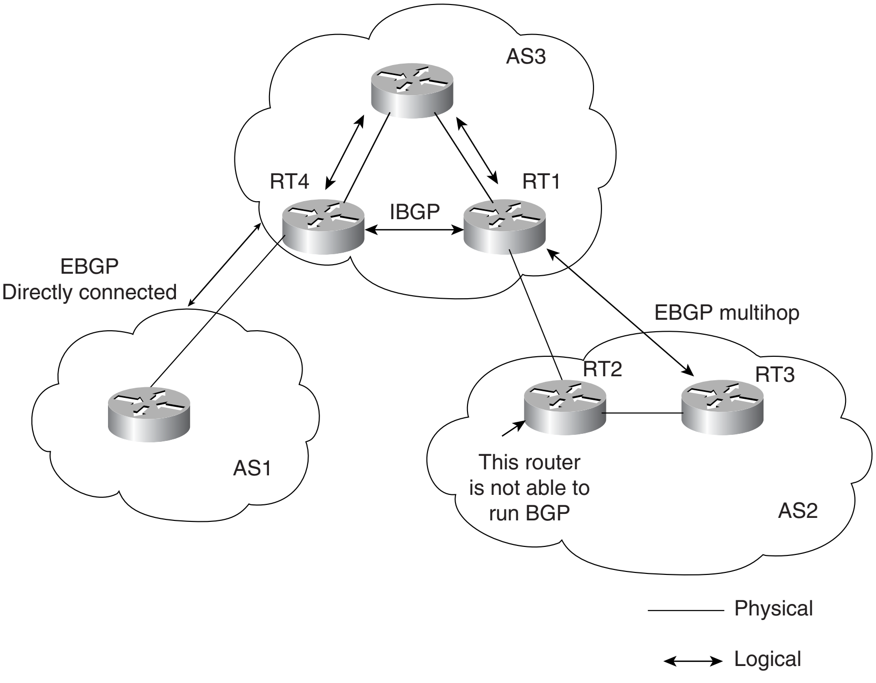
Physical Versus Logical Connections

External BGP neighbors have a restriction in that they must be physically connected, adjacent to one another. BGP drops any UPDATE message from its external BGP peer if the peer is not physically connected, unless otherwise specified. However, some situations arise in which external neighbors cannot be on the same physical segment. Such neighbors are logically connected (multiple IP hops away) but not physically connected. An example would be running BGP between external neighbors across non-BGP routers. In this situation, Cisco (and most other vendors) offers an extra knob to override this restriction. BGP would require some extra configuration to indicate that its external peer is not physically attached.

NOTE Indirectly connected external neighbors require extra configuration.

A BGP session formed between external BGP peers that are not physically connected is referred to as *multihop EBGP*. In Figure 6-2, RT2 can't run BGP, but RT1 and RT3 can. Thus, external neighbors RT1 and RT3 are logically connected and peer with one another via multihop EBGP. (Note, however, that RT2 must somehow learn the appropriate routing information to avoid potential forwarding loops or black-holing packets.)

Figure 6-2 *External BGP Multihop Environment*



On the other hand, neighbors within the same autonomous system (internal neighbors) have no restrictions whatsoever on whether the peer is physically connected or separated by multiple IP hops. As long as there is IP connectivity between the two neighbors, BGP requires no additional configuration. In Figure 6-2, RT1 and RT4 are logically, but not physically, connected. Because both are in the same AS, no additional configuration is required for them to run IBGP.

Obtaining an IP Address

The neighbor's IP address could be the address of any of the routers' interfaces, such as Ethernet, Token Ring, or serial. Keep in mind that the stability of the neighbor connection depends on the stability of the IP address you choose.

NOTE Session stability depends on the stability of neighbor IP addresses.

If the IP address belongs to an Ethernet card that has some hardware problems and is shutting down every few minutes, the neighbor connection and the stability of the routing system will suffer. Cisco provides the capability to configure a virtual interface, referred to as a *loopback interface*, that is supposed to be up at all times. Tying the BGP neighbor connection to a loopback interface will ensure that the BGP session is not reliant on any hardware interface that might be problematic.

Adding loopback interfaces is not necessary in every situation (it actually requires more configuration). If external BGP neighbors are directly connected and the IP addresses of the directly connected segment are used for the neighbor negotiation, a loopback address is of no added value. If the physical link between the two peers is problematic, the session will break with or without loopback.

TIP See the section "Building Peering Sessions" in Chapter 11, "Configuring Basic BGP Functions and Attributes," on page 301.

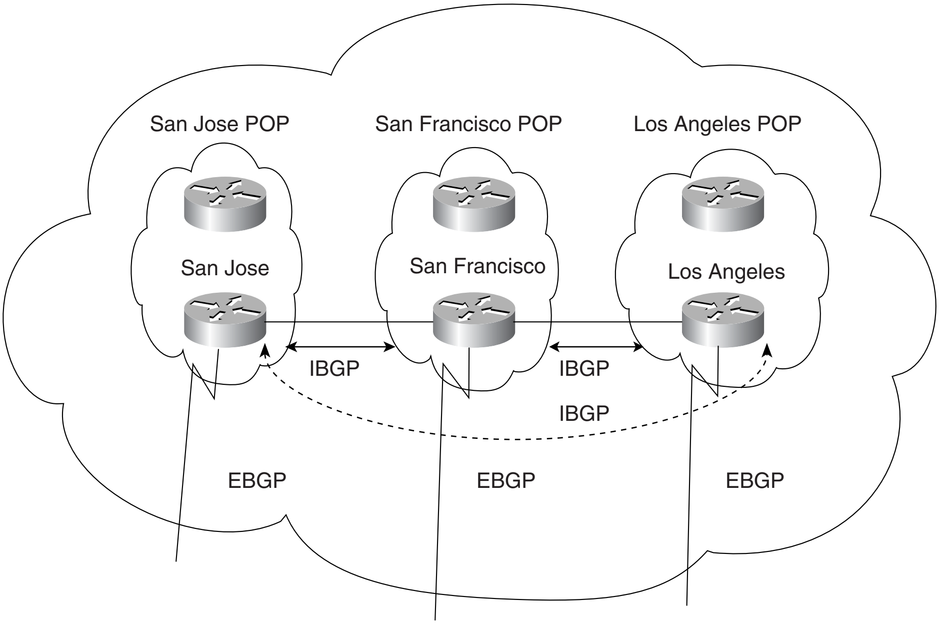
Authenticating the BGP Session

As you saw in Chapter 4, "Interdomain Routing Basics," the BGP message header allows for authentication. Authentication is a precaution against hackers who might present themselves as one of your BGP peers and feed invalid routing information into your AS. Authentication between two BGP peers provides the capability to validate the session between you and your neighbor by using shared secret keys. A neighbor that attempts to establish a session without using the proper key will be ignored. The current authentication features available in BGP-4 use the message-digest version 5 (MD5) algorithm. A detailed discussion of the MD5 authentication algorithm is beyond the scope of this book, but as previously discussed, it can provide added security to the underlying TCP transport connection.

BGP Continuity Inside an AS

Aside from the special case of route reflection, in order to avoid routing information loops inside an AS, BGP does not readvertise to internal BGP peers routes that are learned from other IBGP peers. Thus, it is important to maintain a full IBGP mesh within the AS. In other words, every BGP router in the AS has to establish a BGP session with all other BGP routers inside the AS. Figure 6-3 illustrates one of the common mistakes administrators make when setting BGP routing inside the AS.

Figure 6-3 *Common BGP Continuity Mistake*



In the situation illustrated in Figure 6-3, an ISP has points of presence (POPs) in San Jose, San Francisco, and Los Angeles. Each POP has multiple non-BGP routers and a BGP border router running EBGP with other ASs. The administrator configures an IBGP connection between the San Jose border router and the San Francisco border router. He configures another IBGP connection between the San Francisco border router and the Los

Angeles border router. In this configuration, EBGP routes learned via San Jose will be given to San Francisco, EBGP routes learned via San Francisco are given to San Jose and Los Angeles, and EBGP routes learned via Los Angeles are given to San Francisco. Routing in this picture is not complete: EBGP routes learned via San Jose will not be given to Los Angeles, and EBGP routes learned via Los Angeles will not be given to San Jose. This is because the San Francisco router will not pass on IBGP routes between San Jose and Los Angeles. What is needed is an additional IBGP connection between San Jose and Los Angeles (shown via the dotted line). You will see in Chapter 9, "Controlling Large-Scale Autonomous Systems," how this situation could be handled by using the concept of route reflectors, an option that scales much better in cases where the AS has a large number of IBGP routers.

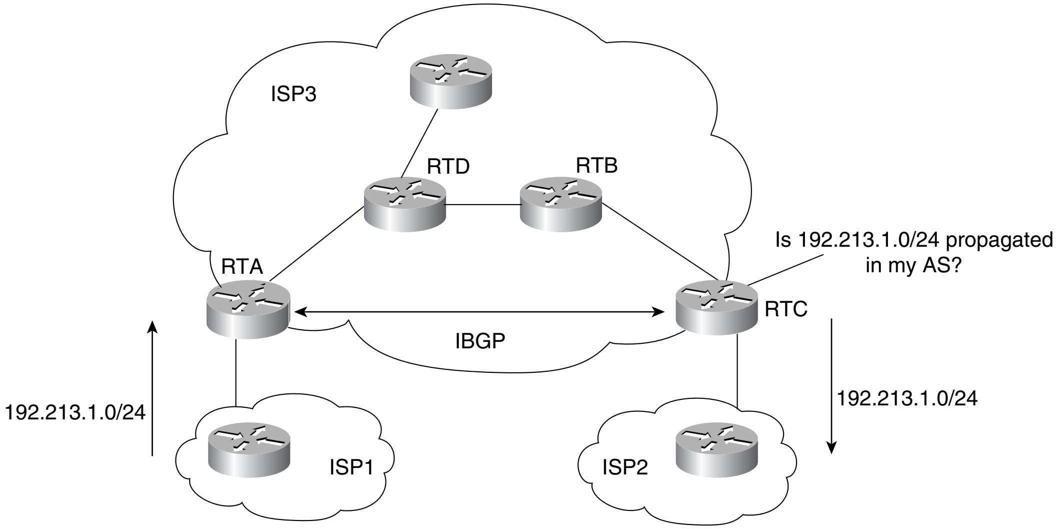
Synchronization Within an AS

By definition, the default behavior of BGP requires that it must be synchronized with the IGP before BGP may advertise transit routes to external ASs. It is important that your AS be consistent about the routes it advertises to avoid unnecessarily black-holing traffic. For example, if an IBGP speaker were to advertise a route to an external peer before all routers within your AS had learned about the route through the IGP, your AS could receive traffic to destinations for which some of the routers might not yet have the information to reach.

Whenever a router receives an update about a destination from an IBGP peer, the router tries to verify internal reachability for that destination before advertising it to other EBGP peers. The router does this by checking the destination prefix first to see if a route to the next-hop router exists and second to see if a destination prefix in the IGP exists. This router check indicates whether non-BGP routers can deliver traffic to that destination. Assuming that the IGP recognizes that destination, the router announces it to other EBGP peers. Otherwise, the router treats the destination prefix as not being synchronized with the IGP and does not advertise it.

Consider the situation illustrated in Figure 6-4. ISP1 and ISP2 use ISP3 as a transit AS. ISP3 has multiple routers in its AS and is running BGP only on the border routers. (Even though RTB and RTD are carrying transit traffic, ISP3 has not configured BGP on these routers.) ISP3 is running an Interior Gateway Protocol inside the AS for internal connectivity.

Figure 6-4 *BGP Route Synchronization*



Assume that ISP1 is advertising route 192.213.1.0/24 to ISP3. Because RTA and RTC are running IBGP, RTA propagates the route to RTC. Note that other routers besides RTA and RTC are not running BGP and have no knowledge so far of the existence of route 192.213.1.0/24.

In the situation illustrated in Figure 6-4, if RTC advertises the route to ISP2, traffic toward the destination 192.213.1.0/24 will start flowing toward RTC. RTC will perform a lookup in its IP routing table and will direct the traffic toward RTB. RTB, having no visibility to the BGP routes, will drop the traffic because it has no knowledge of the destination. The traffic is dropped because BGP and the IGP are not synchronized.

The BGP rule states that a BGP router should not advertise to external neighbors destinations learned from IBGP neighbors unless those destinations are also known via an IGP. This is known as synchronization. If a router knows about these destinations via an IGP, it assumes that the route has already been propagated inside the AS, and internal reachability is ensured.

The consequence of injecting BGP routes inside an IGP is costly. Redistributing routes from BGP into the IGP will result in major overhead on the internal routers, primarily from an IGP scalability perspective, because (as discussed earlier) IGPs are not designed to handle that many routes. Besides, carrying all external routes inside an AS is not necessary. Routing can easily be accomplished by having internal non-BGP routers default to one of the BGP routers. Of course, this will result in suboptimal routing because there is no guarantee that the shortest path for each route will be used, but this cost is minimal compared to maintaining thousands of routes inside the AS. Of course, managing default