

# Interior Gateway Protocols

Eric Hall

Because routing protocols determine the path of IP packets, they also dictate whether packet delivery is timely or even successful. As a result, these services also control how well the higher-layer protocols, like TCP and SMTP, perform or if they are instead encumbered by lost packets, slow delivery, duplicate datagrams or any of the other problems that can result from routing troubles.

Simply put, the successful use of IP and all its higher-layer elements depends on efficient IP routing: If the basic routing service does not work well, everything else will suffer. And if everything else suffers, your online customers might not return to your site or your business might not get online information as fast as it should.

Unfortunately, the world of IP routing can be overwhelming, requiring detailed knowledge of multiple intricate protocols. Although many good books, courses and training programs are available for in-depth study (there is a cottage industry in support of IP routing), we will describe the fundamental principles for two of the most common routing protocols found on corporate networks: RIP (Routing Information Protocol) and OSPF (Open Shortest Path First).

## RIP (Routing Information Protocol)

RIP was originally distributed along with Berkeley's Unix, and like many of the other BSD services, it has become a critical element of IP networks everywhere, even though it was not developed as an Internet standard. Two distinct versions of RIP are now documented as IETF protocols: RFC 1058 describes the original RIP (version 1), while RIP v2 is defined by RFC 1722 (also published as Internet Standard 56). Although the protocols share many similarities, there are some important differences between them.

RIP uses a "distance-vector" algorithm that associates a specific "distance" (the number of hops) with a specific "vector" (a destination network or host). RIP devices learn about destinations and their distance from neighboring RIP routers, then choose the path to a destination based on the route with the lowest number of hops. Once a route for a destination has been chosen, it is stored in a local database, and any other routes for that destination are discarded. Periodically, each router advertises all the paths it has discovered. Eventually, all the devices on a network will discover the best routes to all the available destinations.

RIP defines hops as the number of routers between a sender and the destination network or system. If a router is attached to a network, the distance to that network is zero hops. Similarly, if a router can reach a neighboring network only by sending the datagrams to a neighboring router, the distance to that vector will be one hop. Whenever a router advertises a route, it increments the known distance metric by one hop. As these broadcasts arrive at the neighboring routers, they are compared with the entries already in those routers' databases. If one of the advertised routes to a destination is shorter than an existing entry, the advertised route will be incorporated into the local routing table, with the advertising router being listed as the next hop for the destination.

Some of these issues are clarified in "Enterprise Routing Model" (below), which shows five different network segments. In that example, Router A would advertise single-hop routes for the Ethernet segments and Internet links to which it was directly attached. Meanwhile, Router B would advertise single-hop routes for its local Ethernet attachment, a single-hop route for the WAN subnet and multihop routes for the networks in the remote areas (including the network segment that Server Z is attached to). All these routes will be advertised with broadcasts every 30 seconds, and both routers will republish the routes that they learn from each other.

In the RIP distance-vector model, this means that Router B would see the distance to Server Z's segment as being one hop through Router A, while Router A would see the WAN network as one hop through Router B. However, both routers would see their shared Ethernet as zero hops from

# Interior Gateway Protocols

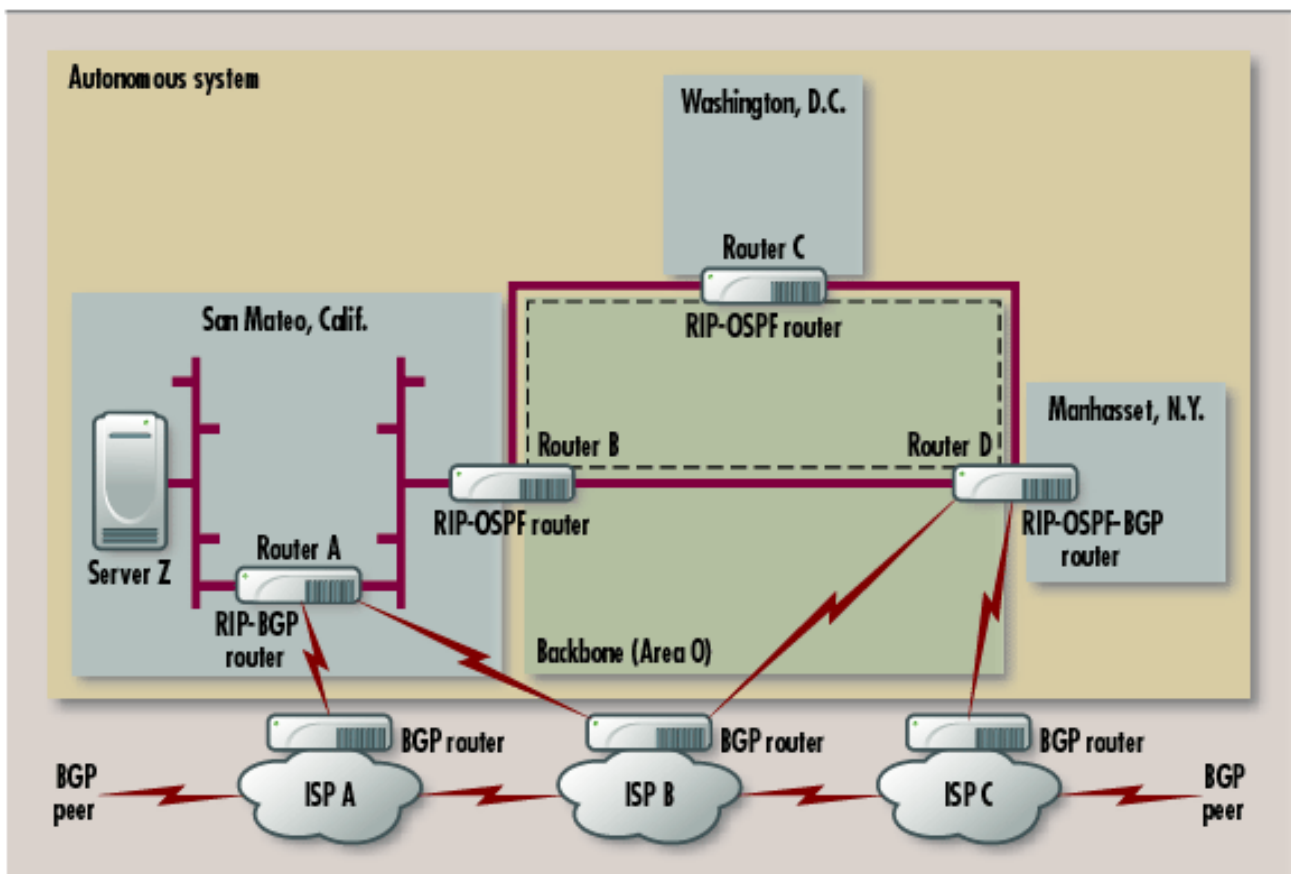
Eric Hall

themselves, and one hop from the other router. Because the advertised route for that segment is longer than the known route, the advertisements for the shared segment would be discarded by both routers.

## RIP Damage Control

RIP imposes a maximum hop count of 15, and any route that is advertised with a distance of 16 hops must be considered as unreachable through the advertised path. RIP uses an expiration timer of 180 seconds. If a route goes silent for three full minutes, the other routers that have learned about the route will set the hop count for the path to 16 (thereby flagging the route as invalid) and will advertise the route with the new distance value for another 120 seconds to ensure that downstream systems learn about the failure.

## ENTERPRISE ROUTING MODEL



However, an older route (with the legacy distance metric) may still be advertised by a distant router. Since that route would have a lower value than that of the infinite path, the remote route would propagate back through the network. For example, if a failure caused Router A to remove Server Z's network segment from its local routing table, then the next update from Router B (with a hop count of 2) could be incorporated into Router A's database as the shortest distance to that destination. In effect, Router A would see Router B as the shortest distance to the Server Z segment and would send traffic for that segment to Router B, though Router B would send it right back to Router A.

One technology that RIP uses to prevent these problems is called split horizon, which prohibits routers from advertising routes through the same interface from which the route was learned. In our

# Interior Gateway Protocols

Eric Hall

example, split horizon would prevent Router B from advertising Router A's local connections back to the Ethernet segment, and vice versa. With this technique, Router A may lose sight of the Server Z segment, but it would never learn of a better path through Router B.

Another technique often used to prevent localized routing loops is poison reverse, which uses a hop count of 16 to explicitly corrupt the reverse path. In that case, Router B would advertise Server Z's network segment with a distance of 16 hops when sending updates, ensuring that the route was not used by Router A. Note that neither of these mechanisms prevents routing loops from occurring when multihop loops are designed into the network but instead prevents only localized loops from occurring through normal operations.

## RIP Lacks Scalability

The biggest problem with RIP is that the 15-hop ceiling often isn't enough for complex campus networks. Many networks are simply too large and distributed to fit within this scope. Furthermore, the use of hops as a distance metric does not always reflect actual cost or bandwidth. Also, a less efficient route may be chosen simply because it has fewer hops or because it has an equal number of hops (all the Internet routes are two hops removed from Router B, and the slower path through Router D could be chosen simply because that advertisement arrived first). Finally, because each router broadcasts its routing table every 30 seconds, RIP can suck up a considerable amount of valuable WAN bandwidth through normal operations, particularly with large networks.

For all these reasons, RIP is not a good choice as a global routing protocol for networks that span multiple distributed sites.

Although RIP 2 improves on some core aspects of RIP 1, it still suffers from most of the same architectural problems as RIP 1. The major improvements in RIP 2 include support for variable-length subnet masks and route aggregation, both of which allow for better address allocation policies and smaller routing messages. In addition, RIP 2 uses multicasting to reduce the impact of frequent updates on nonparticipating local systems, and RIP 2 provides rudimentary authentication support, allowing for a modicum of security in mixed-user installations. Despite these benefits, RIP 2 is still hampered by the same architectural limitations that penalize RIP 1, making it equally inappropriate for complex, multisegment networks.

RIP is useful for small to midsize networks, particularly when those networks have RIP-enabled workstations and servers, and RIP 2 is even better. Because RIP is a relatively simple protocol, it is often implemented as a listen-only daemon, letting devices learn about their local network without needing to maintain static routes. However, it does have problems on large-scale networks.

## Enterprise Routing with OSPF

OSPF is based on the IS-IS (Intermediate System to Intermediate System) routing protocol but is optimized for IP networks in particular. Several IETF documents define the many iterative flavors of OSPF:

- RFC 1131 defines OSPF 1 (field obsolete).
- RFC 1583 is perhaps the most widely implemented rendition of OSPF 2.
- RFC 2328 defines the most current version of OSPF 2 (and is the current source for Internet Standard 54).

With OSPF, every router maintains an independent database of an administrative routing area, including information about the available networks, the routers on the networks and the per-interface

# Interior Gateway Protocols

Eric Hall

cost for each router's connections to each network. In this model, whenever a network, router or interface changes state, each of the routers within the area find out about it and incorporate the new information into the local database, then rebuild the routing maps accordingly. These calculations are performed according to the cost of the network paths for a specific destination, regardless of the number of hops that are required to get through the network. Taken as a whole, OSPF applies a cost-vector algorithm to a database of network objects and uses this information to determine optimum routes throughout an area.

This model opens the door for many compelling features (such as faster change synchronization), but it introduces significant memory and processor demands on the participating systems. For this reason, fewer OSPF-enabled devices are on the market than are RIP-enabled systems. For example, though many server-class operating systems provide OSPF daemons of some kind, not many network clients or low-end devices provide OSPF support, since even passive listeners have to implement the full database analysis engine for them to make use of the link-state data.

The central architectural concept in OSPF is the administrative routing area, which provides a scope control to the network database. All OSPF messages are bound to a specific area; the routers that participate in a shared area will exchange detailed information about that area with each other but will exchange only summary information with routers in remote areas. If an organization needs to use multiple areas, a special backbone area is required to exchange information between the multiple areas. Edge areas must exchange their summary information through the backbone area, meaning OSPF imposes a mandatory two-tier hierarchy on interarea route exchanges (this applies only to the route-control messages, not to all network traffic).

Areas have 32-bit identifiers (which are normally written as IPv4 network addresses), while the backbone area is always numbered 0. Routers can participate in multiple areas simultaneously, but each area will have its own link-state database in the router and require its own memory and processor cycles. In OSPF lingo, a router that participates in multiple areas simultaneously is an ABR (Area Border Router), while a router that exchanges data with foreign routing protocols is known as an ASBR (Autonomous System Border Router).

Remember that routers must make forwarding decisions only one hop at a time. In this regard, OSPF routers do not need to have detailed information about every network in the world (a default route is good enough to move the packets closer to an off-site destination). As such, OSPF uses the detailed information to build routing maps for the local area, but once a router has learned routes for other areas, that information will be used to forward packets to the remote areas directly.

## OSPF Cost Controls

OSPF uses cost as a metric when building routing maps. Cost can be associated with any kind of input mechanism, but most implementations associate cost with available bandwidth. This is achieved by dividing a baseline value (such as 100 million) by the available bandwidth on a specific interface. For example, dividing the baseline value of 100 million by 10 million (for a 10 Mbps Ethernet segment) produces a cost value of 10 for that interface, which is 10 times more expensive than a 100-Mbps interface (which has a cost of 1). All other factors being equal, a router will prefer the lower-cost path, which would be the 100-Mbps interface with a cost of 1 in this case.

When OSPF routers build their routing maps, the total cost for all the outbound interfaces in between the router and the known destinations is incorporated in the cost-vector algorithm. Because some networks provide asynchronous transfer rates, two endpoint systems on a link may have different costs associated with a given link, so only the outbound cost toward a destination is included in these calculations.

# Interior Gateway Protocols

Eric Hall

These concepts are illustrated in "Enterprise Routing Model", which shows three routers attached to a frame relay network labeled as Area 0 (a standalone backbone area). In this example, the three routers have T1 connections to one another, with an associated cost of 64 for each of the network interfaces. However, the routers also have 128-Kbps ISDN backup links to each other (the dotted lines), which have costs of 781.

In this design, Router B would normally send all data for Router C via the frame relay network. But if that link failed, Router B would send the data to Router D for forwarding to Router C, since the cumulative cost for that path would be 128, which is cheaper than using the dial-up backup line, which has a cost of 781. If the frame relay network collapsed entirely, Router B would have no choice but to begin using the backup line. Even if it couldn't connect with Router C directly, however, Router B may be able to send data to Router C by way of an ISDN connection to Router D (at a cumulative cost of 1,562).

When sending data to the Internet, Router C would likely choose to route the data through Router D, since it has the lowest cost (Router B has an extra Ethernet segment, with a minimal cost of 1). However, Router C may end up using Router B if the San Mateo RIP routing has been given a fixed-cost metric, which represents that the San Mateo network is available through a high-speed Ethernet link rather than providing explicit costs for all the networks at San Mateo separately. In that case, Router C would see a cost of 64 plus 64 for the two hops through Manhasset, while seeing a cost of only 64 plus 1 for the two known hops through San Mateo (rather than the actual cost of 64 plus 1 plus 64, which should be associated with the three distinct network segments). As should be obvious from this example, the use of cost as a metric works only when the values have been assigned appropriately and consistently.

## OSPF Database Maintenance

The process of building and maintaining the link-state database is the most complex part of understanding OSPF and unfortunately requires a working understanding of the OSPF protocol structure and mechanisms. For this reason, a completely detailed discussion on this subject is beyond the reach of this primer. However, some of the fundamental principles of the protocol are simple. For more information on this topic, we encourage you to get one of the books available on OSPF design or to read RFC 2328.

Each OSPF node within a routing area maintains its own link-state database for all the networks, routers and interfaces associated with that area. During steady-state operations, the routers simply exchange OSPF Hello messages, which are small datagrams that advertise only that a particular router is still up and running. During synchronization operations, however, a variety of complex LSA (Link-State Advertisement) messages will be exchanged, depending on the event that occurred, the state of the database and other factors.

If an interface changes state, only a small amount of database activity is required to fully integrate the information into the area databases on all the routers within that area. If a new OSPF router is brought online, however, that router will have to discover all the routers, networks and interfaces within its area, and this process can consume a significant amount of network resources.

On broadcast and multiaccess networks, OSPF supports the use of a designated router, which lets new routers obtain complete copies of the database with minimal network impact. On point-to-point networks, however, each router has to obtain link-state data from each of the other routers independently.

# Interior Gateway Protocols

Eric Hall

This database-synchronization model represents what is perhaps the greatest challenge with running OSPF in large, complex networks, since a significant amount of time can be spent maintaining database synchronization in the face of network stability problems.

However, OSPF has many other features that make it compelling for large and complex corporate networks. But despite these advantages, OSPF is overkill for small or self-contained networks, and the use of RIP can often pay greater dividends.

## Routing Protocols at a Glance

Many routing protocols can be used to automate the process of path discovery, with each of these protocols providing different benefits and having different costs. Here are some of the more common of these protocols.

- **Border Gateway Protocol (BGP):** Version 4 is the exterior routing protocol of choice on the Internet and is used as an interior routing protocol in some environments. BGP routers exchange routes to organizational networks (as identified by autonomous system numbers), with these routes being tweaked according to the monetary cost of the connected link, the bandwidth available through the advertising organization's link and other details. In practical terms, this means that BGP is a contract vector-routing protocol, where any number of routes may be available for a destination, but the chosen path will most often be determined by the terms of the contracts that govern the available paths.
- **ICMP Router Discovery Protocol (IRDP):** ICMP is best known for providing the kind of diagnostic and alert messages that are used by programs such as ping and traceroute. ICMP also provides a series of router-information messages that can be used to discover and advertise the presence of a default router on a network. In this model, routers can advertise themselves as offering default routes with a specific preference weight (operator assigned), and workstations can pick the best router based on the preference value. This means that managers do not have to configure workstations with specific routes, nor do they have to assign subnet-specific routes to DHCP assignment pools. Instead, they can simply enable IRDP on the workstations and walk away. However, IRDP is useful only for choosing default routes and cannot be used to assign routes with any higher granularity, so this restricts its functionality.
- **Intermediate System to Intermediate System (IS-IS):** OSPF is a derivative of IS-IS, which was the first fully blown link-state routing protocol. IS-IS was intended for use with the OSI protocols but was extendible to other protocol families. Because of this flexibility and its earlier presence, it is still found on some older large-scale networks, but its younger cousins, OSPF and NLSP, are much more common.
- **NetWare Link-Services Protocol (NLSP):** NLSP is the IPX-centric version of IS-IS. NLSP provides a link-state replacement for the IPX-specific version of RIP and the NetWare SAP (Service Advertisement Protocol), such that IPX routes and services have to be transmitted over WAN links only whenever those resources change rather than being advertised every 60 seconds.
- **Open Shortest Path First (OSPF):** OSPF was originally designed as an IP-specific link-state protocol similar to IS-IS, though it has diverged from that design dramatically over the years.
- **Routing Information Protocol (RIP):** In theory, RIP supports multiple protocols, including IP and IPX, though in practice these are covered by two separate protocols that have slightly different mechanics. For example, the IP implementation of RIP broadcasts all known routes every 30 seconds, while the IPX version broadcasts routes every 60 seconds.