An Introduction to IP Routing

Geoff Huston
January 2001

The structure of the global Internet can be likened to a loose coalition of semi-autonomous constituent networks. Each of these networks operates with its own policies, prices, services and customers. Each network makes independent decisions about where and how to secure supply of various components that are needed to create the network service. With all this independent freedom of decision how is it that the result is one of an apparently cohesive network, rather than a disparate collection of isolated islands of local connectivity?

That's a pretty big question, and the answer is equally big, with technical, business and social dimensions. Lets limit our attention to the technical part of the answer, and take a look at what binds these thousands of component networks into a single Internet. The basic technical glue of the Internet is routing and addressing. In the address real, each network uses a unique set of addresses drawn from a single global address space. Each connected device has a unique address that it uses to label its network interface. Each IP packet generated by these devices has a source and destination address. The source address references the local interface address, and, logically, the destination address is the corresponding interface address of the intended recipient. As it is being passed within the network from router to router, the router can identify this intended recipient. But within a network identity is only half of the solution. Complementing identity, the network must be able to know location, or, where the packet is to be directed. The task of associating location with identity, or in other words maintaining routing information within a network is undertaken by routing protocols.

The intended result of the routing system is quite impressive: at every decision point within the entire Internet the local router has adequate information to switch any IP packet to the 'correct' output port. In a routing sense 'correct' not only means 'closer to the destination' but also means 'consistent with the best possible path from the sender to the recipient'. Such an outcome requires the routing protocol to maintain both local and global state information, as the router must be able to identify a set of output ports that will carry a packet closer to its destination, but also select a port from this set which represents the best possible path to the destination. Again this is not all that a routing protocol must achieve. We have to add to this picture the observation that routers and links are not perfectly reliable. Whenever a network component, such as a router or a link fails, the routing protocol must attempt to repair the break by establishing a new set of paths that avoids the failed component. When a component is restored, or new routers and links are added to the network, again the routing protocol must reevaluate the topology of the network and possibly set up a new collection of switching paths through the network. And, of course, this information must be flooded to all routers in the network as soon as possible after the event.

In a small network this can be a forbidding problem. In a large network such as the Internet, with millions of end devices and hundreds of thousands of links and routers, it's a even more forbidding problem. The technique used by the Internet to achieve this functionality is one of dividing the problem into more manageable tasks. In the routing domain this division of the problem corresponds to the structure of the Internet itself: each separate network runs its own local internal routing protocol (or *Interior Gateway Protocol*, or IGP) and the collection of networks is joined into one large routing domain through the use of an inter-network routing protocol (or *Exterior Gateway Protocol*, or EGP).

There are a number of interior routing protocols, including RIPv2, EIGRP, OSPF and IS-IS. They all perform a similar function, that of maintaining an accurate view of the current topology

of the local network. For all addresses that are reachable within the network the routing protocol computes the best path to the address from all points in the network. In the event of failure of a network component, or a change in operational state of a component, the protocol is intended to react quickly to create an updated view of the network. There are two basic approaches to implementing this protocol function.

Both RIPv2 and EIGRP use a so-called distance vector algorithm, where each router computes a local view of address reachability and passes this information to its neighbors. The neighbors incorporate this view into their address reachability tables and pass this updated information to its neighbors, and so on. While simple to set up, these algorithms do have some weakness. They work by sending full address tables to all neighbors at regular intervals. In a network with a large address table this can become a significant overhead. When the network experiences heavy congestion these large routing updates may fail, causing the routing protocol to become unstable. The protocol takes some time to converge to a stable state following a link failure, as the iterative process of updates takes some time to ripples across the network. During this time the routing protocol may form loops as an intermediate network state.

OSPF and IS-IS are instances to a link state flooding protocol. Such protocols use a technique of uniquely identifying each link within the network, and when a link changes state this information is rapidly passed across all routers in the network. Each router maintains the same table of link states and is able to compute a complete picture of the current topology of the network. From this topology view the local router is able to generate a local forwarding table corresponding to the best paths to reach each address destination. Link state protocols converge rapidly, and without forming routing loops in the process. However there are some issues with stability of very large networks. To address this both OPSF and IS-IS use the a two level hierarchy of routing, termed *areas*. Within an area, the protocol undertakes a complete flooding of link state. Between areas the protocol passes address reachability information rather than link states. The intent is to localize the fine-grained control of topology state to a defined area. Careful design of areas assists in scaling the routing system to quite large and complex networks. Within an ISP network it is common to see OSPF or IS-IS being used as the local routing protocol.

For many networks the routing design begins and ends with OSPF or IS-IS. The network carries full information about all addresses used within the network and computes paths to each destination. For packets moving entirely within the local network this is all that's necessary. But what about packets destined to addresses within some remote network? Does the IGP need to maintain full routing information for all known Internet addresses? As there some 100,000 such addresses, that's a large task! Mercifully, the answer is 'no'. At the point where a network connects to its upstream ISP you can originate a special address, the *default route*. The default route is passed within the IGP in the same fashion as any other internal address. Its purpose is to direct all packets destined to non-local networks towards the upstream ISP.

So, in many cases, OSPF plus default, or IS-IS plus default is all that's need to set up a network to be part of the routing of the Internet. However that's not all of the routing picture. When a network is connected to a number of other networks and not just one, then the routing system needs to perform more work. The routing system must learn which destinations are reachable from each of the neighboring networks, and direct traffic accordingly. A single default is now no longer adequate. This task of exchanging reachability information between networks is undertaken by an exterior routing protocol.

As with IGPs, there are a number of exterior routing protocols. The most common in the Internet today is BGP4. When two networks exchange information using BGP4 they do not tell each other the precise path to a particular destination. Instead, they simply inform the neighboring network that if they receive a packet addressed to a particular destination, then they will be able to deliver it. This does not necessarily mean that the address is part of the local network. The network may have been learned from another neighboring network via BGP, and the network is

willing to allow transit traffic between the two neighboring networks. When an ISP network connects to multiple ISPs it is often the case that the same address is reachable via two or more neighboring networks. Left to its own devices BGP will select paths that traverse the fewest possible number of providers to reach each destination. But BGP does not have to run in such a fully automated mode. BGP4 has an additional function not found in IGPs – that of policy specification and enforcement. One upstream provider may be cheaper than another, or one neighbor may be a peer while the other may be a customer. BGP allows a network to express preferences in which neighbor to prefer when choosing a path for external addresses. A common policy is that of preferring customer routes over peer routers and peer routers over upstream routes. An ISP may prefer to use routes from one upstream provider over another, and so on. Path selection is not only possible for outgoing traffic. A network may attempt to bias incoming traffic to use particular networks over others, and do so for particular addresses. It may sound somewhat clumsy, but it is on these foundations that traffic engineering and load balancing is constructed in today's Internet.

This is a quick pass across the building blocks of the Internet's routing system. The big question is how will it deal with the demands of tomorrow's Internet? We will look at this question in a future article.