

Routing with OSPF

Introduction

The capabilities of an internet are largely determined by its routing protocol. An internet's scalability, its ability to quickly route around failures, and the consumption of network resources by the routing machinery are all issues directly related to the routing protocol. With the release of HP router software revision 5.70, OSPF (Open Shortest Path First) is available in addition to RIP.

RIP (Routing Information Protocol) is probably the most widely used IP routing protocol. Its popularity stems from having been included with Berkeley UNIX (*Routed*, the routing daemon) and from being standardized by the IETF (Internet Engineering Task Force). RIP is documented in RFC (Request for Comments) 1058. RIP is a distance-vector protocol. A distance-vector protocol frequently (at 30-second intervals for RIP) sends its routing table (a vector of distances) to neighbor (adjacent) routers. When a router receives its neighbor's routing update, it compares the update with its own routing table and changes its routing table if necessary.

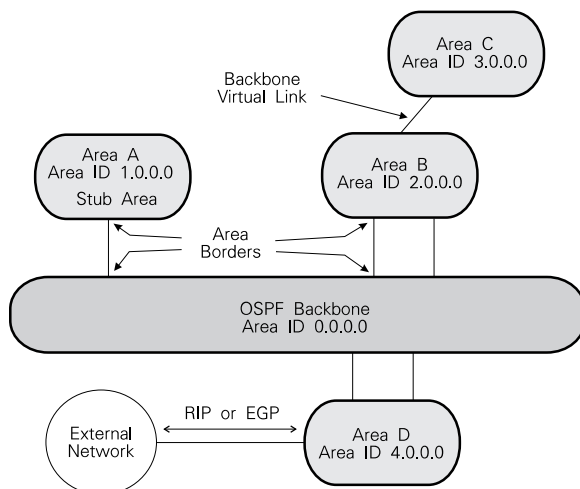
Distance-vector protocols are susceptible to two main problems. First, they can form routing loops, and second,

they can be slow to converge. Convergence is the time required for the routing tables in all of the connected routers to stabilize after an event such as a network link failure. Routing loops and slow convergence are more likely to occur as network size and complexity increase. A fundamental assumption for routing (in the design of RIP) was that the internet contained at most a few hundred networks. Thus, RIP is not well suited for use with today's large corporate and government internets, which have thousands of networks. The limitations inherent in distance-vector protocols such as RIP and the lack of a standard routing protocol suitable for use in large internets are in large part responsible for the development of OSPF.

OSPF (Open Shortest Path First) is a new IP routing protocol. HP routers implement Version 2 of OSPF, which is documented in RFC 1247. Unlike RIP, OSPF employs a link-state algorithm (also referred to as a shortest-path-first (SPF) algorithm). Link-state algorithms are those in which each routing node floods information about its attached links to all other routing nodes.

Figure 1 shows an OSPF AS (Autonomous System). AS is an IP term that refers to a collection of routers that all use the same IGP (interior gateway protocol). IGP is also an IP term. It refers to the routing protocol run within an AS. IGPs are a subset of routing protocols. They are referred to as IGPs to distinguish them from EGPs (exterior gateway protocols), another type of routing protocol. EGPs are used to route data between ASs. The Exterior Gateway Protocol (also known as EGP) is also the name of a specific EGP.

Figure 1



OSPF Autonomous System

OSPF has features that:

- Improve routing effectiveness and efficiency.
- Conserve IP address space and increase addressing flexibility.
- Enhance network security.
- Increase routing flexibility.

These features make OSPF attractive for use on both small and medium-sized internets as well as large internets. This application note provides an overview of OSPF and describes selected features in more detail using model networks as example.

Routing Improvements

OSPF has many features that are unarguably improvements over RIP. Those core features that improve the effectiveness and efficiency of routing include hierarchical routing, new routing metrics, the link-state protocol, and topological database.

Hierarchical Routing--Areas

OSPF supports a routing hierarchy. Just as a hierarchical file structure is a better way to organize files than a flat file structure, so too a hierarchical network structure is a better way to organize networks than a flat network structure. The OSPF hierarchical structure helps to reduce the size of the topological database maintained by each router. (The topological database is discussed in more detail below.) It also helps to minimize routing control traffic. An AS may consist of one or more "areas". Small networks may consist of a single area; large networks may contain many areas. Each area comprises a group of networks (or subnets). Each area in the AS is attached to the OSPF Backbone. An OSPF AS is logically a star: areas extend in a radial fashion from the Backbone. Note that a single-area AS would not have a Backbone area.

Data is routed within an area when the destination system is in the same area as the source. This means that when two systems inside area A (figure 1) communicate, the data is routed within area A. When the destination system is in a different area than that of the source, data is routed from the source area to the Backbone area to the destination area. Therefore when a system in area A communicates with a system in area B, the traffic is routed from within area A to the Backbone. (The OSPF Backbone is an area in itself.) The router that connects area A to the Backbone is referred to as an "area border router". The data is then routed through the Backbone to area B, where it is finally routed to the destination system.

Area and Router IDs

Both areas and routers have IDs (identifiers) which are used by OSPF to build its topological database. Both IDs are given in dotted decimal notation. This is the same notation used for IP addresses. The range of the identifiers is thus 0.0.0.0 to 255.255.255.255. Area identifier 0.0.0.0 is reserved for the Backbone area. Area and router IDs are not IP addresses, however, and thus such concepts as IP address classes, subnetting, broadcast addresses, etc., do not apply.

The OSPF standard does not provide guidance on the selection of area identifiers except the Backbone area (0.0.0.0). Several schemes have been proposed. A unified way to select both area IDs and router IDs is to assign IDs using a scheme compatible with that of the corporation or organization, such as AREA . REGION . OFFICE . ROUTER.

In figure 1, area A has an area ID of 1.0.0.0, signifying a particular geographic area--a country, for example. Router IDs can then be selected based on their location within area 1.0.0.0. Thus the first router in office 1 of region 1 of area 1 receives the router ID 1.1.1.1. In any event, it is desirable to select identifiers that have some significance.

Area Sizing

A frequently asked question is: "How large should an OSPF area be?" Or: "What is the optimal number of routers in an OSPF area?" An area could contain as few as one router or as many as hundreds of routers. Carving a network up into smaller-sized areas will help to minimize the amount of OSPF protocol traffic. The costs associated with reducing the size of areas include additional complexity in the OSPF Backbone and the network as a whole. The answer to the question of area size probably has more to do with organizational structure than with the OSPF protocol. Generally, areas are suggested by the structure of the organization(s) responsible for managing the network. Within many large corporations, responsibility for network management is distributed. Corporate telecommunications departments manage the network in and around the corporate offices or sites, while network management in other regions is handled by other groups in those regions. These organizational boundaries are also natural OSPF area boundaries.

Stub Areas

Stub areas are areas into which external routes are not propagated. The term external route has a particular significance in OSPF. Routing information provided to OSPF from any protocol other than OSPF itself is considered external. Thus routes provided by RIP or EGP as well as static routes are considered external. An OSPF router that interfaces to an external router is called an "AS Boundary Router". Consider an AS connected to the open internet through EGP. There are potentially thousands of routes for which reachability information could be obtained. To prevent this information from being propagated into an area, an area can be configured as a stub area. In figure 1, area A is a stub area. Therefore, external routing information received in area D will not be propagated into area A.

Virtual Links

In figure 1, area C is not directly attached to the Backbone. Instead, it is attached to area B through a "Backbone virtual link". Virtual links allow areas to be configured where it would otherwise be inconvenient to do so due to the distance or cost to attach to the Backbone.

Metrics

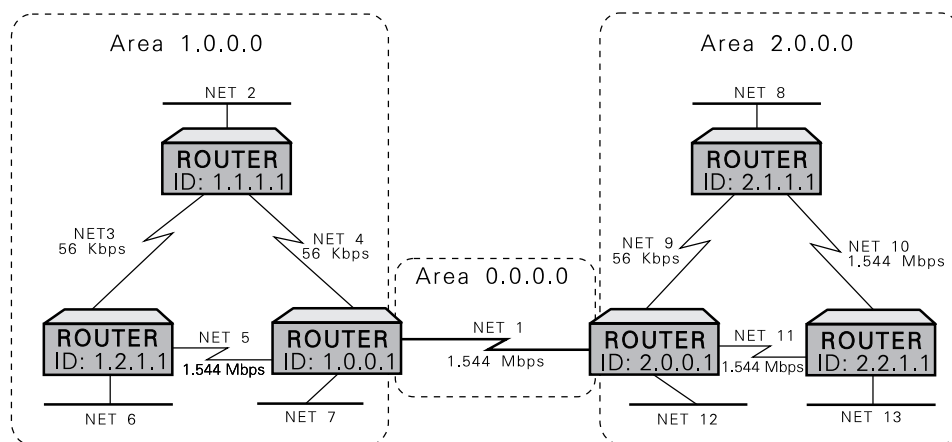
The OSPF routing algorithm calculates the shortest path to each destination network. A path is composed of a series of network links from a router to a particular destination network. Each link is assigned a metric. The metric for a path is simply the sum of all the link metrics. The shortest path to a network is thus the path with the lowest metric. As networks become more complex, the result of alternate routes and varying link speeds, metrics become more important. OSPF provides a 16-bit (0 to 65535) dimensionless metric for the assignment of link costs and allows 24 bits for inter-area paths.

The OSPF standard does not provide guidance for metric selection or assignment. The easiest way to assign metrics is on the basis of link speed. Table A shows one possible scheme for selecting metrics based on link speed.

Each link in the network is assigned a cost (metric). In figure 2 the cost from router 1.1.1.1 to net 8 is the sum of the costs for net 4, net 1, net 11, net 10, and net 8. Assuming costs are assigned in accordance with table A, the cost of the route from router 1.1.1.1 to net 8 is thus $110 + 40 + 40 + 40 + 10 = 240$.

Notice that the shortest path is not necessarily the one with the fewest number of hops, but rather the one with the lowest metric.

Figure 2



Multi-Area OSPF Internet

Link-State Protocol

When OSPF is started, usually during the router's boot procedure, it begins by synchronizing its database with those of its neighbor routers. Afterwards each router infrequently (at 30-minute intervals) floods LSAs (link-state advertisements) to all other routers in its area. Flooding is a way to send a message that will be relayed by all routers receiving the

Table A

Link Speed	Metric
100 Mbit/s	3
16 Mbit/s	7
10 Mbit/s	10
4 Mbit/s	15
2.048 Mbit/s	32
1.544 Mbit/s	40
768 Kbit/s	75
512 Kbit/s	85
256 Kbit/s	95
128 Kbit/s	100
64 Kbit/s	105
56 Kbit/s	110
38.4 Kbit/s	120
19.2 Kbit/s	150
9.6 Kbit/s	200

Metrics Based on Link Speed

message. Received LSAs are used to build and maintain a topological database from which each router builds its routing table. There are several types of LSAs. Consider the network in figure 2. Each router in area 1.0.0.0 sends a router links LSA to all other routers in area 1.0.0.0 at 30-minute intervals or whenever a link state changes. The router links LSA includes the Router_ID of the router that originated the message (called the advertising router) and a description of each of the links connected to it. Link descriptions vary by the type of connected network. However, in this example the information in the LSA about a synchronous link includes:

- IP address of the link.
- Subnet mask used on the link.
- The router ID of the remote end router.
- The type of the link (point-to-point).
- Metric or cost assigned to the link.

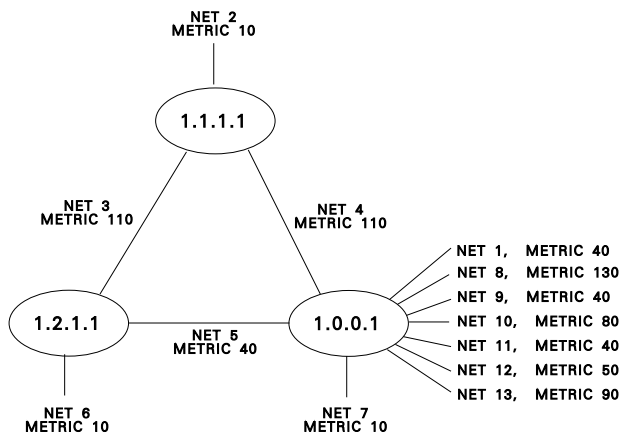
Like the routers in area 1.0.0.0, routers in area 2.0.0.0 flood router-links advertisements to the other routers in area 2.0.0.0.

Routers maintain detailed topological information about area(s) of which they are members, and they maintain summary information about networks in other areas. Area border routers (1.0.0.1 and 2.0.0.1 in figure 2) exchange summary information about their own areas with other area border routers. Network summary information received by an area border router is then transmitted to routers in its attached area(s). The type of LSA used to send network summaries is called a summary-links advertisement. Like router-links advertisements, summary-links advertisements are also sent at 30-minute intervals and are triggered when a link state changes. A summary-links advertisement includes the following for each link advertised:

- IP network number or IP address of the link.
- Router ID (set to that of the local area border router).
- Subnet mask used on the link
- Metric or cost from the area border router to the destination network.

Routers exchange LSAs to build and maintain their topological databases.

Figure 3



Area 1.0.0.0 Topological Database

Topological Database

Routers in each area have identical routing databases. This topological database describes which routers are connected to which networks. Attributes of the networks and routers such as subnet masks, metrics, etc., are also maintained in the database. Each router constructs its routing table from the database. The routing table contains the shortest path to every network the router can reach. The benefit of maintaining a topological database is that when a change

occurs to the network, new loop-free routes can be computed quickly using minimal network resources. When a link outage occurs, for example, each router that detects the change floods LSAs that describe the change to all other routers in the network. Each router then modifies its database and recomputes the shortest path to each remaining network.

RIP, by comparison, maintains just the best route to any given network. When a change to the network occurs, routers must exchange their entire routing table with each neighboring router to relearn the best route to every network.

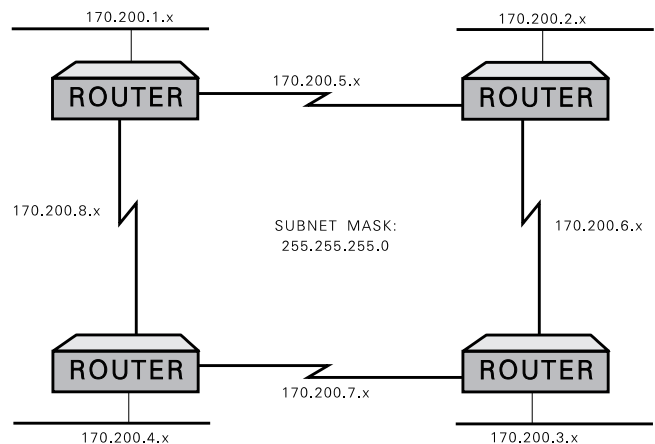
Figure 3 shows a summary of the topological database of area 1.0.0.0 (from figure 2). Each router that is a member of area 1.0.0.0 has a copy of this database.

One aspect of the use of areas is that it simplifies the database. Notice that routers in area 0.0.0.0 and 2.0.0.0 do not appear in the topological database, although the networks in those areas do.

Conserving IP Address Space

On internets using RIP, the subnet mask used throughout the internet must be identical on all subnets. The information about individual networks included in a RIP update includes the network (or subnet) number and metric (hop count). RIP updates do not include the subnet mask associated with a network. This often results in the over-allocation of IP address space--especially on synchronous point-to-point links.

Figure 4



Subnetting in a RIP Internet

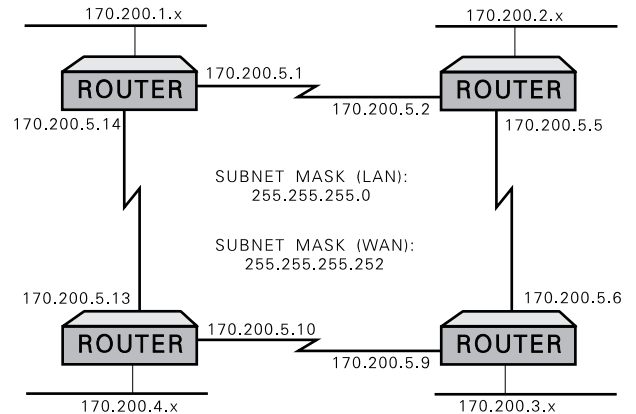
The internet in figure 4 uses the IP class B address 170.200.x.x. This internet is subnetted using the subnet mask 255.255.255.0. There are 254 possible addresses on each subnet (addresses 0 and 255 are reserved). On each of the wide-area network subnets, however, there are only two members, namely the router on each end of the point-to-point link. Of the 254 possible (or allocated) addresses, only two will be used. Thus 252 addresses are unused or wasted. Of the approximately 2000 IP addresses allocated in the internet in figure 4, roughly half are unusable because they are allocated to point-to-point networks. Variable-length subnet masks, a feature of OSPF, can be used to minimize the allocation of IP addresses, such as in the case of the point-to-point links in figure 4.

Variable-Length Subnet Masks

The mechanics of conserving IP address space using variable-length subnet masks are straightforward. Again consider the internet in figure 4. With OSPF the subnet mask can be specified per network or subnetwork. The subnet masks for the LAN subnets can be specified as before (255.255.255.0). This allows 254 addressable nodes per subnet. The point-to-point links, however, require only two IP addresses--one for each router. To restrict the number of addresses on the point-to-point links, a subnet mask is needed that allows only two addresses. It is tempting to use the subnet mask 255.255.255.254. The two addresses that this mask provides, however, are the "broadcast" address and the "any host" address for the subnet. Therefore, one additional bit is required in the subnet mask, so that reserved addresses will not be used. Thus, the subnet mask 255.255.255.252 is the minimum subnet mask that will provide only two addresses per subnet. Now that a suitable subnet mask for use on point-to-point links has been determined, subnet numbers and addresses must be selected.

The following method of selecting subnet numbers and point-to-point link addresses illustrates one method of allocating addresses that seemingly maintains the address structure used in allocating LAN subnets. The first point-to-point subnet defined in figure 4 was subnet 170.200.5.x, the second was 170.200.6.x, and so on. To maintain consistency with that addressing scheme, the point-to-point

Figure 5



Internet with Variable-Length Subnet Masks

subnets to be defined will use 170.200.5.x as the base for selecting link addresses. The last octet (the "x" octet) will be used to define individual point-to-point subnets and addresses. Table C shows the allocation of subnets and point-to-point addresses. Extending the subnet mask to include the upper 6 bits of the last octet for point-to-point links provides enough address space for 64 point-to-point links.

Using the addresses in table C, the internet from figure 4 is shown in figure 5. Full point-to-point link addresses are shown in figure 5 rather than subnets only, since the subnet mask is split on an octet boundary.

Reserved Addresses

When using variable-length subnet masks, special attention is required to avoid assigning addresses that are reserved. Avoiding the use of the "all hosts broadcast" address (all ones assigned to host address field of an IP address) and the "any host" address (all zeros assigned to the host-address field) was discussed above.

The other consideration that warrants special attention is avoiding the assignment of reserved subnet addresses.

Reserved subnet addresses are those that are all ones (subnet broadcast) and those that are all zeros (any subnet). Variable-length subnet masks introduce extra complexity to this issue, since there may now be several different subnet definitions in a single network.

Table C

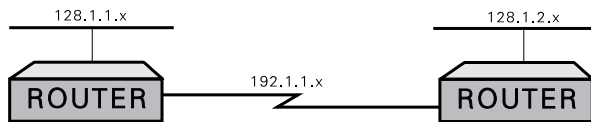
16 Bits	8 Bits	Upper 6 Bits	Lower 2 Bits	32 Bits / 32 Bits
Network Number	Subnet Number	Subnet Extension Component	Link Address Component	Full Point-to-Point Link Addresses
170.200	5	000000	01 /10	170.200.5.1 / 170.200.5.2
170.200	5	000001	01 /10	170.200.5.5 / 170.200.5.6
170.200	5	000010	01 /10	170.200.5.9 / 170.200.5.10
170.200	5	000011	01 /10	170.200.5.13 / 170.200.5.14

Point-to-Point Subnet and Address Definition

For example, consider the internet in figure 5. IP subnets in the range 170.200.0.4 through 170.200.0.248 appear to be valid IP subnets when used with the WAN subnet mask 255.255.255.252. However, these addresses are within the range of the reserved LAN subnet addresses, 170.200.0.0 through 170.200.0.255 (subnet mask 255.255.255.0), and thus must not be assigned. Similarly, subnets 170.200.255.4 through 170.200.255.248 appear to be valid IP subnets when used with the 255.255.255.252 subnet mask. This address range is also within the reserved LAN subnet address range and must not be assigned.

The rule is to avoid assigning addresses within the reserved address ranges given by the subnet mask with the fewest number of bits in the subnet ID field. In figure 5 the subnet mask with the fewest number of bits in the subnet ID is 255.255.255.0. Thus the reserved address ranges are 170.200.0.0 through 170.200.0.255 and 170.200.255.0 through 170.200.255.255.

Figure 6



A Partitioned Network

Addressing Flexibility

OSPF enhances IP addressing flexibility by allowing IP networks to be partitioned. Network partitioning is not permissible with RIP. A network becomes partitioned when one or more subnets of a network become separated from the other subnets of the same network by a second network. In figure 6, the class B network 128.1.x.x is partitioned by network 192.1.1.x.

To understand why network partitioning might be useful, refer again to figure 6. Suppose network 128.1.x.x is a large network with very few remaining subnets.

Subnets of the class C network 192.1.1.x can be used for synchronous links so that none of the remaining class B subnets have to be used to extend the network.

Another situation in which support for partitioned networks is helpful is when networks are combined as a result of a merger or acquisition. Networks often overlap in this case, often resulting in capacity imbalances. Being able to

partition a network allows the network's administrators much more flexibility in assembling the combined networks.

Network Security

Routing Authentication

To enhance security, routing updates may optionally be authenticated using a simple password. When routing authentication is enabled, all OSPF protocol packets are password protected. Passwords are from 1 to 8 characters and are configurable on a link basis. The determination to use routing authentication is made on an area basis. When routing authentication is used in an area, passwords must be configured on all links in the area.

Information Hiding

As discussed above, routers in each area have identical topological databases. Each router knows the topology of the areas to which it is attached. Only summary information is exchanged between areas. Thus, the topology of an area is hidden from routers outside the area.

Type-of-Service Routing

Type-of-service routing is not yet available on HP routers. Currently, HP routers implement TOS 0 routing only. The following discussion is intended only to explain the type-of-service routing concept.

Type-of-service routing is based on the type of service defined in the Internet Protocol Specification, RFC 791. Three abstract quality-of-service parameters are given--delay, throughput, and reliability. These quality-of-service parameters (when used) are set in the IP protocol header by the system originating the IP datagram. Type-of-service values range from 0 to 7. Table D shows the four most common types of service. Note, the other types of service

Table D

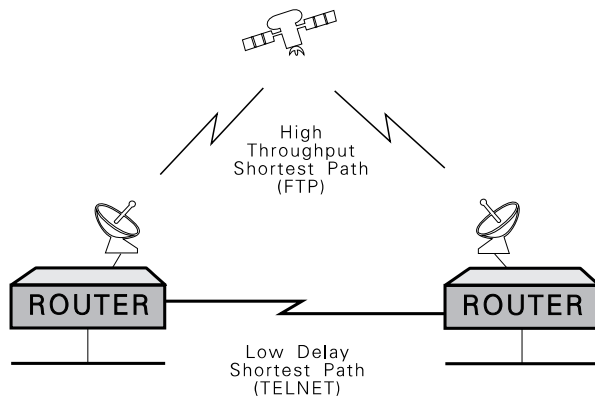
Type of Service (TOS)	Delay	Throughput	Reliability	Service Description
0	0	0	0	Default
1	0	0	1	High Reliability
2	0	1	0	High Throughput
4	1	0	0	Low Delay

IP Types of Service

are based on combinations of the quality of service parameters given in table D.

When multiple types of service are supported by HP routers, routing decisions can be based on the TOS requested in the header of a datagram. Thus a separate set of routes can be calculated for each IP type of service. Conceptually, this will allow networks where file transfers

Figure 7



Routing Different Types of Service

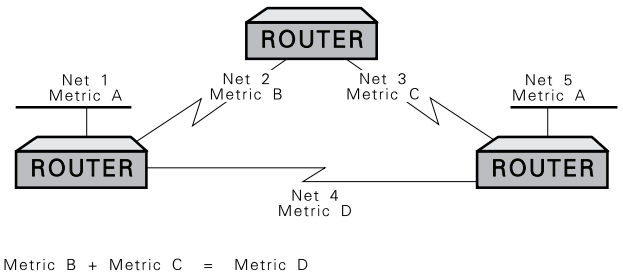
(using FTP) can be routed over high-throughput routes such as satellite circuits, and time-sensitive data (using Telnet) can be routed over low-delay terrestrial circuits (see figure 7).

Equal Cost Multipath

Equal-cost-multipath routing is not yet available on HP routers. The following discussion is intended only to explain the equal-cost-multipath routing concept.

Consider the network in figure 8. There are two equal-cost paths from net 1 to net 5. The sum of the metrics for net 2 plus net 3 is identical to the metric for net 4. When packets destined for net 5 are received on net 1 by the

Figure 8



Equal-Cost Routes from Net 1 to Net 5

router, packets will be forwarded on both paths. Obviously, careful attention must be paid to the assignment of link metrics to ensure that multiple paths are used to forward data when that is desired.

Conclusion

OSPF provides support for large networks using hierarchical network organization, improved metrics, and a link-state protocol. Additional features that help conserve IP address space, make addressing more flexible, and improve network security make OSPF attractive for use on many small and medium-sized internets as well.

OSPF features available in future router releases, such as type-of-service routing and equal-cost-multipath routing, will add even more value to an already valuable internetworking tool.