6. Decision Process

The Link State protocol allows each router to construct a graph representing the routing area. The vertices of the graph are the nodes and pseudonodes—one for every LSP series. The arcs of the graph are the links between (pseudo)nodes. They are reported in the Link Information fields of LSPs. The Decision Process operates on this graph. It uses Dijkstra's algorithm to produce the Forwarding database. For each IPX™ network number in the Link State database, the Forwarding database indicates the next hop from this router toward that destination network number. The network numbers are in the LSPs—and consequently attached to nodes of the link state graph—in the Management Information field and the External Routes field.

The Decision Process is also responsible for determining equal cost paths, computing the throughput and delay characteristics of the end-to-end path, and calculating the RIP delay (Ticks) to report to RIP routers and to end nodes.

6.1. Running the Decision Process

The Decision Process runs when a change occurs to the Link State database. It is not run immediately, but rather the router waits for five seconds after detecting the change. This allows several changes clustered in time to be dealt with at once.

Implementations should ensure that the Link State database is not modified while the Decision Process is running. There is more than one way to accomplish this, and the approach chosen is influenced by the operating environment surrounding the routing software.

One approach is for the Decision Process to signal when it is running. During this time, incoming LSPs are queued but not inserted in the Link State database. If more LSPs arrive than can fit in the space allotted for the queue, the router drops any excess LSPs without acknowledging them; that is, without including their sequence numbers in any SNPs transmitted.

A second approach is to use two copies of the Link State database. When the Decision Process starts, it takes a snapshot copy of the database, keeping the snapshot static until completion of the process. Meanwhile, updates continue to be incorporated into the dynamic copy. If there is a separate memory area for each copy, the areas can alternate between the static and dynamic roles.

The first approach uses less memory, but the second approach avoids dropping incoming LSPs.

Once started, an execution of the Decision Process should not be abandoned due to new information arriving. To do so could lead to starvation; the process might never be run to completion. If an event occurs that would change the Link State database while the Decision Process is running, the five-second timer restarts when the Decision Process is completed.

6.2. Dijkstra's Algorithm in Pseudocode

The description of Dijkstra's algorithm in Section 2 is intended to appeal to your graphical intuition. Here is another description of the same algorithm, more akin to pseudocode.

```
START:

  Keep a list containing all "unplaced" nodes, and associate a cost with
  each.

  Start with all nodes in the unplaced list.

  Set the cost of all nodes to infinite except the local node, which is set
  to 0.


  Keep a list of nodes that are already in the "shortest path" spanning tree.
  (This is the Known Set of the description in Section 2.)

  This set starts empty.


NEXT_NODE:

  Select a node in the unplaced list with lowest cost that is not infinite.

  Remove the selected node from the unplaced list and add it to the shortest-
  path list.

  If there is no such node, the algorithm is complete and all nodes remaining
  in the unplaced list are unreachable.

  In the unplaced list, update the costs of all nodes adjacent to the node
  just removed. The new cost is the minimum of
          (a) the previous cost of the adjacent node, and
          (b) the sum of
               (i) the cost of the node just removed, and
               (ii) the link cost of the adjacency.

  goto NEXT_NODE
```

For example, at Step 6 in Section 2.1.7, the unplaced list looks like this, after removing V from the list:

S	40
C	50
W	infinite

Because S has the lowest cost, it is next to be removed.

When there is a tie comparing costs, additional criteria are applied, in a specified order, until the tie is broken. The list of criteria is called the *order of preference of routes.*

(a) End-to-end path with the lowest cost

(b) End-to-end path providing the highest net Throughput

(c) End-to-end path providing the lowest total Delay

(d) End-to-end path supporting the largest MTU size

(e) First Hop node with the Lowest System ID

(f) Circuit with the lowest localCircuitID on the local router

(g) Neighbor with the lowest LAN MAC address on the remote router

The MTU Size, Delay, and Throughput are in the Link Information of LSPs. System IDs are assigned by an administrator to each router; circuit IDs are assigned by router software. The reason for these criteria are (a) to provide the best routes, and (b) to provide deterministic routes (for reproducibility in problem resolution).

6.3. Load-splitting

The NLSP™ specification supports load-splitting. That is, if there are equal cost paths, the traffic can be divided among them to make fuller use of the internetwork.

First, define a *maximal splitting degree*. This is a number, MSD, assigned by the network administrator. It can be different for different routers. If there are equal cost paths, MSD is an upper limit on the number used for routing. If MSD=1, the router does not do load-splitting. Load-splitting focuses on the step in Dijkstra's algorithm where you choose which node to add to the known set.

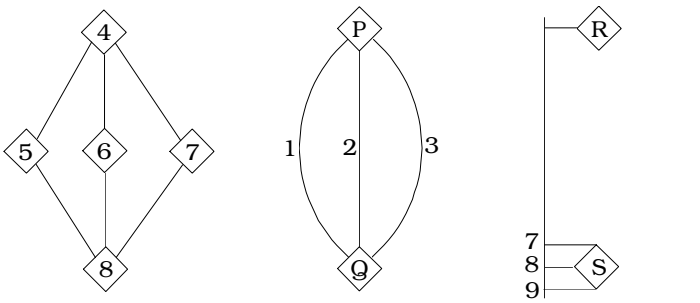
In case of a tie in cost to the same far node, the Forwarding database has more than one entry added to it, instead of just one. How many? If the tie is among *m* links, the number of entries added is the smaller of *m* and MSD. If *m* is greater than MSD, the choice of which to add is based on the order of preference of routes listed on page 6-2.

Figure 6-1 shows examples of rules (e), (f), and (g). (There are no parts (a)-(d) in the figure.) In each example, the links are of equal cost, and MSD is two.

Part (e) shows the system IDs of the routers (single-digit for brevity). When router #4 adds #8 to the Known Set, it uses the two links to #5 and #6, omitting the link to #7.

In part (f), the localCircuitIDs are shown. When router P adds Q to the known set, it uses links #1 and #2, but not #3.

Part (g) shows two routers connected by two LANs, with S having three connections to one of the LANs. (The MAC addresses are shown as single-digit numbers for brevity.) When router R adds S to the known set, it chooses S's two lowest-numbered MAC addresses, 7 and 8.



(e) system ID *(f) circuit ID* *(g) LAN address*

Figure 6-1: Load-splitting

Finally, when comparing the number of links with MSD, count the number of nodes traversed, not the number of partial paths. Figure 6-2 illustrates this point. When A adds D to the Known Set, it considers the number of paths to D to be two. So it splits the A-to-D load

between B and C. For the B part of that, it has already decided to split traffic between the two links.

There is an alternative acceptable implementation. Instead of focusing on the node being added, look at the set of next-hop adjacencies and prune these to, at most, MSD. With this approach, node A in Figure 6-2 splits traffic destined to D two ways rather than three.

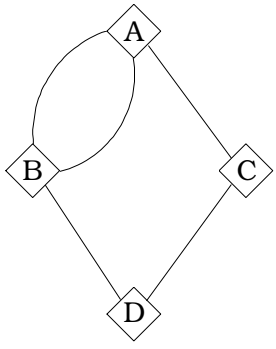


Figure 6-2: Count Nodes, Not Links

6.4. Information Used and Not Used

The Decision Process makes a special provision for an overloaded router; that is, for a router in the area having the LSP database overload bit set in its LSP number zero. Data traffic whose final destination is in the overloaded system itself is delivered to that system. (This leaves the door open for management action to diagnose the problem and perhaps even to reconfigure things and extricate the router from its predicament.) However, data traffic is not routed through an overloaded router to any other system—not to other routers and not to end nodes.

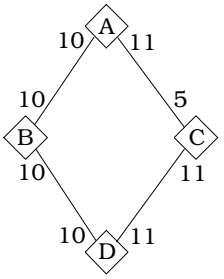
A system can indicate in its LSP that it is a multihomed nonrouting server. Such a server is attached to more than one network segment, but does not forward data traffic from one segment to another. However, it does participate with routers in NLSP exchanges to advertise its services and its internal network number. It also runs the decision process, so it can determine on which network segment to transmit outgoing data packets it originates. When a system running the decision process detects the LSP of a multihomed nonrouting server, it treats that server like an overloaded router. Data traffic whose final destination is in the server itself is delivered to that system. However, data traffic is not routed through that system to any other system.

Implementing this provision involves the step in Dijkstra's algorithm where a node is added to the shortest-path list. At this step, the router checks whether the node being added is overloaded. If so, it adds it anyway, as usual, but after that point all outgoing links from the overloaded router are disregarded for the remainder of the Decision Process.

The Decision Process does not use a link between two routers unless each router reports a link to the other router.

> **Note:** There is no way to determine that the links reported by the two routers are the same link. This is not a problem. Data traffic is still forwarded over the (possibly asymmetric) paths.

It is possible for the two routers to report different costs for the link. In this case, routes might be asymmetric. Figure 6-3 shows an example.



Numbers shown are link costs reported by the nearby router.

Note that A and C report different values for the same link.

The consequence is that traffic from A to D goes through B, while traffic from D to A goes through C. The total cost from A to D is 20; for D to A, the total cost is 16.

Figure 6-3: Example of an Asymmetric Route.

The IPX network numbers in the routing area are accumulated from the Management Information and External Route options of LSPs in the Link State database. It is possible that a network number is accessible both through NLSP and through RIP. That is, it appears in the Management Information option of one LSP and the External Route option of another. The Decision Process ignores the external route in this case.

6.5. Products of the Decision Process

The Forwarding database consists of a set of pairs (network number, next hop). The next hop indicates on which circuit, and to which neighbor on that circuit, traffic to the network number should be immediately forwarded by this router. If load-splitting applies (page 6-3), there might be more than one next hop for a network number.

In certain cases (for example, LSP database overload, or a recent network partition), some network numbers might be known but unreachable.

If the Decision Process determines that certain nodes are unreachable, the router still keeps the associated LSPs in the Link State database. They persist until the LSP to which they refer expires or is deleted. The reason is that links can disappear and then reappear soon. Because a link can provide reachability to a large number of networks, it is inadvisable to delete them all only to re-create them soon after.

6.6. Routing in the Face of a LAN Partition

While running the Decision Process, a router might discover two Designated Routers for the same IPX network number. Consider Figure 6-4. The LAN with IPX network number *N* can become partitioned if the MAC-level bridge fails. Routers *A*, *B*, and *C* all still consider the LAN to have IPX network number *N*. By the election process, both *A* and *B* could become Designated Routers for *N*.

For traffic whose final destination is network *N*, the routers *R*, *S*, and *T* have no way to know whether *A* or *B* is the route to choose. By default, they choose whichever is closer to them. Reaching the destination is problematic. There is not much that can be done to improve on this situation.

For other routing decisions, however, the situation is better in the face of this partition. Consider traffic from *R* destined for network *M*. By tracing the adjacencies, *R* readily

determines that the path *R*-S-*B*-C-*T* will reach network *M*. Likewise, router *A* can reach network *M* by a circuitous path.

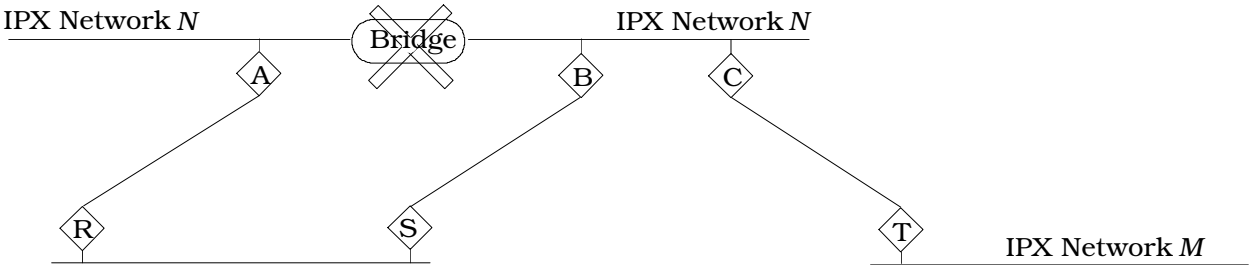


Figure 6-4: LAN Partition Example

Other hardware failure modes can cause partitions. Suppose, for example, that (instead of the bridge failing) the network interface connecting *B* to the LAN *N* were to fail. Again, both *A* and *B* could emerge as Designated Routers for *N*. But this time, *A* and C are adjacent, instead of *B* and C. *R* and *S* can route to anywhere through *A*. In addition, the *S*-*B* link is used from anywhere to reach SAP services residing on system B (because they are associated with *B*'s internal network number, not with network *N*).

6.7. Routing Outside the Routing Area

Although this specification encompasses Level 1 routing only, the Level 1 router must implement certain features to operate in an internetwork that includes both Level 1 and Level 2 routing.

6.7.1. Calculating the Actual Area Address

Each router reports its manualAreaAddresses in the Area Addresses option of its LSP number zero. The Decision Process collects these values. The collected list forms the set of synonymous area address values that define and delimit the routing area. The combination is called the actualAreaAddress.

Note: Level 1 routers from other areas can share circuits with routers of this area, but they fail to form adjacencies—an adjacency is formed only if the two routers have at least one manual area address in common. Consequently, LSPs from other routing areas do not appear in the Link State database.

Note: The collection of area addresses includes all systems in the Link State database, whether currently reachable or not. It also includes those in this router's own Level 1 LSP number zero.

If there are more than three area addresses in the collected list, the router retains three of them, as follows:

a) If the address parts are unequal, the lowest address is preferred.

b) Otherwise, the lowest mask is preferred.

For example, (0xC9000000,0xFF000000) is preferred over (0xC9000000, 0xFFFF0000).

If the preceding steps cause one of the router's own manualAreaAddresses to be dropped, the event manualAddressDroppedFromArea is generated. This is a means to detect misconfigurations.

Note: Implementors should document the following warning for the benefit of end users in administering networks based on NLSP.

The multiple area address support is not intended to permit all routers that share an area address to form a single area automatically in all situations. The intention is that under typical circumstances, all routers within the area have identical sets of manual area addresses. The ability to form adjacencies to systems that only have a subset of manual area addresses in common is intended to allow area addresses to be added or removed from the set of the whole area in a graceful manner; for example, given a set of routers all with area addresses A and B, it is possible to add area address C to routers in the area one by one.

The area is defined by the set of routers that are all connected by Level 1 adjacencies.

There are circumstances where divergence from this rule causes problems. If a LAN has three routers with area addresses A, AB, and B, then adjacencies are formed A-AB and AB-B, but not A-B. If the Designated Router priorities of A and B are higher than AB, then both A and B are elected Designated Router for the LAN, with the consequence that two sets of pseudonode LSPs are generated for the LAN. This does not compromise reachability, but if the LSPs are large (for example, if there are a large number of routers on the LAN), it is possible that the memory limits of routers in the area are exceeded, causing them to go into the overloaded state.

6.7.2. Routing to an Exit Router

The key to relaying data traffic to destinations outside the routing area is to locate the nearest *Exit Router*. An Exit Router could be either

- A Level 2 NLSP router

Or

- A router that attaches this area to other networks using a different method.

Either way, an Exit Router is recognized by examining the Attached Flag and Router Type fields in the LSPs of the Level 1 Link State database. An Exit Router has the Attached Flag set to "Yes" and the Router Type set to "Level 1 and Level 2 Router." These flags are significant only in a router's LSP number zero.

The Decision Process determines the Exit Router whose NLSP cost is smallest. Figure 6-5 illustrates an example of an Exit Router, E. Its Attached Flag is set to indicate that it can reach outside the routing area. It also is a Level 1 router, and, for example, conveys traffic from A to D as necessary. The links to the right of E are not in the Link State database. Exactly how E routes outside the area is not part of the NLSP specification. It could be by NLSP Level 2 routing, or it could use a different method.

Load-splitting can be applied, as with other routes. The router chooses one Exit Router and splits the load between paths to that router. (As an implementation option, a router can split the load between two or more equally distant Exit Routers.)

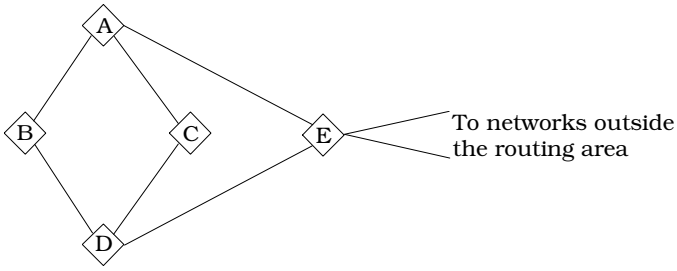


Figure 6-5: Exit Router Example

6.7.3. Forwarding Data Packets

When the router forwards an IPX data packet, it proceeds as follows:

a) Look up the route in the Level 1 Forwarding database. If found, use the information to identify the corresponding next hop.

b) If not found, forward the packet to the nearest Exit Router (do not check whether the destination is in the area).

Note: Lack of the destination check means that some packets destined inside the area might be sent to the Exit Router unnecessarily. However, checking for each packet would be a performance burden.

c) If in the preceding step there is no Exit Router accessible, discard the packet and generate an ipxOutNoRoutes event.

6.8. Decision Process Database

6.8.1. Configured Values

MSD
Maximal Splitting Degree: the maximum number of equal-cost paths the router uses to any one destination. Default is 1; no splitting.

6.8.2. Dynamic Values

actualAreaAddress
The collected manualAreaAddresses of the routers in the routing area. The combined set of up to three synonymous area addresses identifies the routing area.

6.8.3. NLSP Events

manualAddressDroppedFromArea
More than three manualAreaAddresses were detected in the area, and this router dropped one of its own when forming the actualAreaAddress.

ipxOutNoRoutes
An IPX packet could not be forwarded because the destination network number is unreachable and there is no reachable Exit Router.