Some Thoughts on Multi-Domain Routing

Ross Callon
Digital Equipment Corporation
July, 1991

Status of this Memo

This internet draft is an informal proposal. It is hoped that this work may influence future Internet Standards, but this draft does not represent a standard. This paper represents the personal opinion of the author, although some of the ideas presented are borrowed from other sources.

Abstract

Frequently a router may have multiple sources of information about paths through the network. For example, a router may be simultaneously learning routes from multiple routing domains, and may have static routes configured by the network manager.

This paper presents a proposal for control of interactions between multiple routing protocols, including choice of routes when multiple IP routing protocols are being used simultaneously, as well of exchange of routes between routing protocols. This paper has been inspired by the two papers "Ruminations on the Next Hop" and "Ruminations on Route Leaking" by Philip Almquist, as well as by many conversations with Philip and with Chris Gunner.

Comments should be sent to “callon@bigfut.enet.dec.com”.

Contents

Figures

1 Introduction: Multi-Domain Routing Interactions

Frequently a router may have multiple sources of information about paths through the network. For example, a router may be simultaneously learning routes from multiple routing domains, (running multiple routing protocols, and/or multiple instances of the same routing protocol) and may have static routes configured by the network manager.

This paper presents a proposal for control of interactions between multiple routing protocols, including choice of routes for routers in multiple domains, as well of leakage of routes between domains. This paper has been inspired by the two papers "Ruminations on the Next Hop" [1] and "Ruminations on Route Leaking" [2] by Philip Almquist, as well as by multiple conversations with Philip and with Chris Gunner (from whom I also stole some descriptive text).

This is a DRAFT paper, and will need further revision and enhancement.

2 An Abstract Model for Correct Routing Protocol Interactions

This section proposes an abstract model which may be used to help ensure correct behavior of multiple interacting routing protocols.

The "Safety Principle" described in section 2.1, if followed, will guarantee that the interactions between multiple routing protocols will not create route loops. The "Routing Progress Principle", if followed completely, will ensure against "black holes". Note that in practice, it should be possible (in the absence of errors) to ensure that the safety principle is followed. However, it will not be possible to ensure that the routing progress principle is always followed completely. These principles, even when not followed completely, may be used to help determine under what circumstances looping or black holes might be possible.

Note that the model presented in this section is intended as an **abstract model** only. This model is not suitable for direct use in any implementation. The primary value of this model is therefore to aid in the understanding of multi-domain routing interactions.

The abstract model presented in this section does not directly consider the internal behavior of any one instance of one routing protocol, but rather applies to the interactions between different routing domains running different instances of routing protocols. It will be assumed that each routing protocol itself behaves in an internally consistent and correct manner. Also, dynamic aspects of routing behaviour, relating to the speed of convergence of intra-domain and inter-domain routing are not explicitly considered.

A basis for the model is to observe that a network in which multiple routing protocols are operating can be partitioned into ***domains***. Each domain is defined to consist of those routers which are exchanging routing information through one instance of one routing protocol. A ***border router*** is defined as one which exists in more than one domain. A border router therefore runs more than one instance of a routing protocol. In many cases each domain makes use of a different routing protocol (e.g., one domain may run OSPF while another runs Integrated IS-IS). However, the model is generalized to the case where multiple instances of the same routing protocol are run, each in a separate domain.

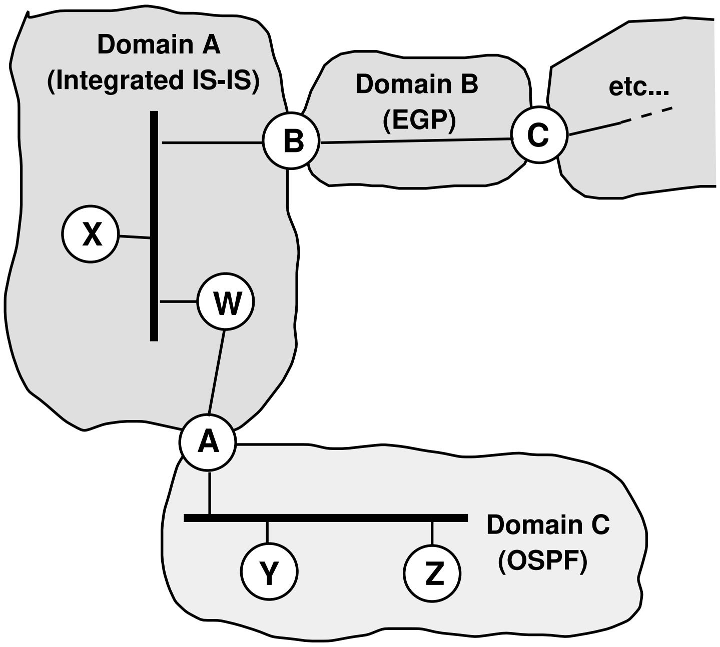

Figure 1 - Network Domain Example

Figure 1 illustrates our use of the term "domain". In this example, routers A, B, W, and X are in the Domain A, and are running one instance of Integrated IS-IS. Similarly, routers A, Y, and Z, are in routing domain C and are running one instance of OSPF. We have modelled use of EGP, such as running EGP between two routers B and C in the example, as a routing domain.In this example, routers A, B, and C are border routers, and are therefore in more than one domain (i.e. running more than one instance of one routing protocol). A non-border router is one which exists only in one domain and which therefore runs only one instance of one routing protocol.

2.1 The Safety Principle

The safety principle may be used to check for the possibility of routing loops between a set of interconnected routing domains.

In order to describe the safety principle, we need to consider the set of all route entries that may apply for any one destination. The term ***route entry*** is used to describe the routes advertised by routing protocols. In principle, each route entry consists of a {IP address, subnet mask} combination, although with some routing protocols (such as RIP) only the IP address is explicitly provided (the subnet mask is implied, and may be known to routers on the basis of configuration information maintained by the routers).

With link state protocols (such as OSPF and Integrated IS-IS) a route entry exists for each router which is advertising in its LSPs that it can reach a particular destination. Generally, the route entry includes the IP address and subnet mask, as well as a set of one or more metrics that describe the distance to the destination. For distance vector protocols (such as RIP), a separate route entry occurs for each router which is advertising that it can reach some destination at some given metric cost. In general, with distance vector routing protocols every router in a domain will be advertising routes to all destinations, with the metric values appropriately adjusted to account for the distance to each destination.

This implies that for any particular destination, there are a very large number of route entries available in any one routing domain. However, the model described in this chapter is an abstract model intended solely to describe routing correctness, and is not intended to be implementable. Thus an explosion in the number of route entries is not a problem.

For any particular set of interconnected routing domains, for any one destination IP address and requested TOS class, we can consider the complete set of all route entries which apply to that destination. In principle, we could create a large list of such entries, ordered by the relative preference of each route. This would create a partially ordered list, since there will in general be some pairs of entries which will never need to be compared directly, and for which an ordering does not exist (such as intra-area entries in two distant routing domains).

Each route entry on the list includes an IP address, a subnet mask, a TOS class, a metric value, an indication of which routing domain, area (if the domain is hierarchically organized) and router the entry comes from, and the "route class" (as defined in "Ruminations on the Next Hop" [1]).

In order for the list to be meaningful, the following conditions must be met:

1) The list must be ***complete***: for any routing entry *A* which applies to the desired IP address (and TOS), occuring in any routing domain under consideration, an entry corresponding to *A* must exist on the list.

2) The list must be ***sufficiently ordered***: for any two routing entries *A* and *B*, if these two entries will ever be compared by any router, then either *A* must be prefered to *B*, or *B* must be prefered to *A*.

3) The list must be ***internally consistent***: For any three entries *A*, *B*, and *C* on the list, if *A* is prefered to *B*, and *B* is prefered to *C*, then *A* must be prefered to *C*.

4) The list must be ***compatible with routing protocols***: The order of preference of routes in the list must be compatible with the order of preference of routes specified for each routing protocol in use in any of the associated routing domains.

5) The router implementations must be ***correct***: Any router in any of the routing domains under consideration must forward user data packets in a manner which is consistent with the preference order of routes on the list.

In order to ensure that routing loops do not occur, Route Leaking (exchange of routes learned from one routing protocol into another protocol) must meet the the following additional condition:

6) Route leaking must follow the ***safety principle***: Routes may only be leaked from a higher preference entry on the list to a lower preference entry.

These conditions are sufficient to ensure that route looping will not occur. This set of criteria therefore provide a method for evaluating whether any particular configuration might cause a loop. Also note that this requires something that resembles global coordination (between any set of domains which may be parts of a single multi-domain loop)., However, in the absence of coordnation, it is clear that looping will be a possibility.

Some examples of error situations (which violate the above principles, and in which loops occur) are illustrated in section 2.3 below.

2.2 The Routing Progress Principle

The Routing Progress Principle may be used to check for the possibility of black holes -- routers into which packets may disappear without being able to subsequently be forwarded towards the destination.

In principle, black holes cannot be completely eliminated. In particular, some addresses are not reachable, and some addresses will not correspond to any existent end system. In many cases, due to the use of hierarchical addressing, the non-existence of the end system corresponding to a particular address will not be know to routers outside of a limited area. This implies that a packet may need to travel some distance before being discarded.

In order to ensure against black holes, the following conditions must be met:

1) Routing must follow the safety principle, described above

2) For every route entry being advertised by a router, one of the following must be true:

   a) the entry specifies a host route, and the router has a direct route to that host (either the host is actually part of the router, or the routers has a link to the host, with no intervening routers); or:

   b) the entry specifies a subnet route corresponding to a real physical subnet, and the router has an interface directly on the physical subnet (with no intervening routers); or:

   c) The router knows about a higher preference route to the specified destinations.

Note that the routing progress principle ensures that a packet cannot get "lost", it will always arrive at routers which know a higher preference route which will allow the packet to continue to make progress towards the destination, until it reaches a router which has a direct host or subnet route. Also, the safety principle outlined above ensures that the packet cannot loop.

However, in general the routing progress principle cannot always be followed completely. For example, a router advertising a default route does not really have a route to every IP address. Thus the routing progress principle is more useful in understanding where black holes may occur than in eliminating black holes.

2.3 Examples of Error Conditions

This section illustrates some simple examples where things don't work. The examples shown here are somewhat over-simplified, but should suffice to illustrate the safety principle and the routing progress principle.

2.3.1 RIP Infinity Problem

Figure 3 illustrates a problem with the interactions between a routing domain running RIP (in which a metric value of 16 is considered infinity), and a routing domain running a routing protocol using a metric with a larger dynamic range. Here an over-simplified example is chosen in order to illustrate the safety principle.

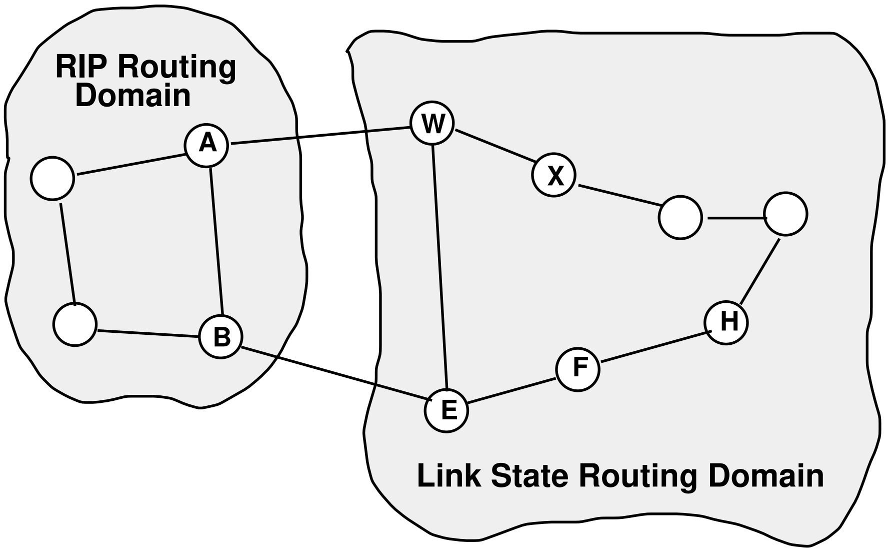

Figure 2 - Example with Different Metric Dynamic Ranges

In this example, let's suppose that the two domains have the same overall preference level for some particular destination, and that the relative preference of any particular route entry is determined by looking at the advertised metric value. Also, lets simplify the example by assuming that neither routing domain makes use of TOS routing.

Let's suppose that in the example, the link state routing domain is using some "modern" link state routing protocol (for example, OSPF or Integrated IS-IS) which provides a wide dynamic range for metric values, so each of the links is configured to have a cost of "10". Lets suppose that subnet "19.234.*" is reachable via router H via cost "1". Thus router E has a route to subnet

"19.234.*" with cost 21 (two hops of cost 10, plus the initial cost 1). However, router E cannot advertise this route to router B (using RIP) with cost 21, since this is greater than RIP's notion of infinity. Therefore, lets assume that router E does something very stupid, and, in its RIP packet sent to router B, it advertises the route to subnet "19.234.*" at cost one. This implies that router B will advertise the route with cost 2, and router A will advertise this route with cost 3. Now, let's assume that router W also does something stupid, believes what it hears from router A, and decides that it's lowest cost path to "19.234.*" is via router A. Router W therefore advertised this route into the link state routing domain with cost 4, which causes router E to decide that the best path to the subnet is via router W (with cost 14, which is lower than the cost of 21 via router F).

In this case, the routes within the two illustrated routing domains have a precedence which is determined solely via the associated metric values, and the metrics in the two domains are assumed to have comparable metrics. Thus, when router E leaks the route from the link state routing domain (in which it has a route with cost of 21), into the RIP domain (with cost 1), it is leaking a route from a lower precedence entry to a higher precedence entry. This illustrates a clear violation of the safety principle.

2.3.2 How a Single-Level Protocol and a Hierarchical Protocol Can Mess Up

Figure 3 illustrates an example of the interaction between a single level routing protocol (perhaps RIP), and a two-level hierarchical routing domain.

Let's suppose that the folks who had the network in the figure 2 figured out what they were doing wrong, and changed all of the configured link costs to 1, so that the RIP value of infinity would not be exceeded. This allowed the border routers to advertise the various entries into the RIP domain without having to reduce the metric. However, over time the link state routing domain grew to the point that it needed to be sub-divided hierarchically, creating the situation illustrated in figure 3.

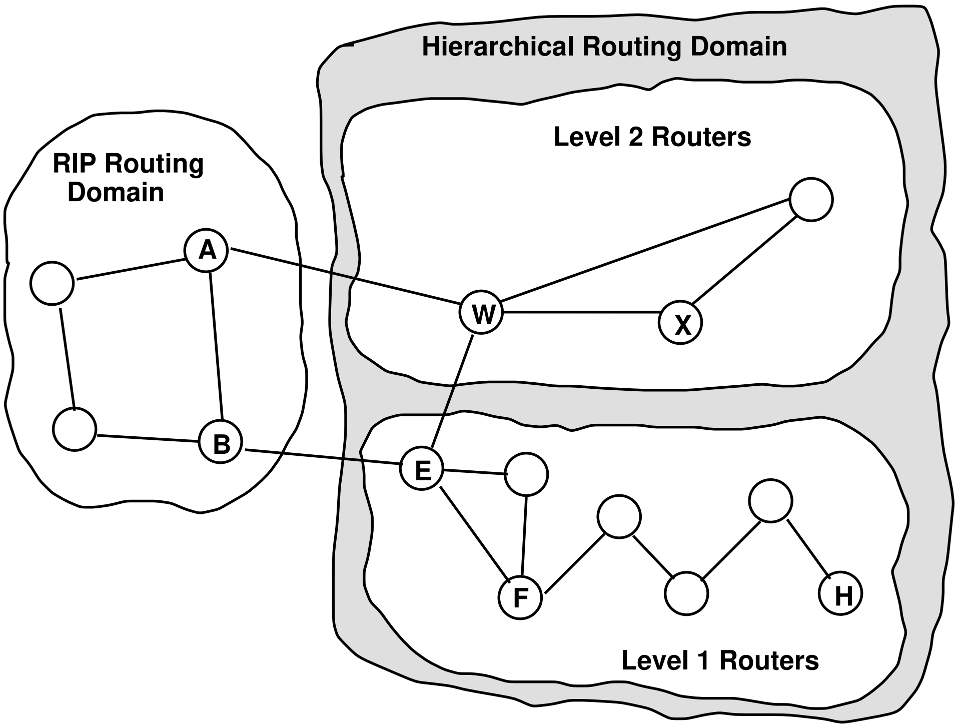

Figure 3 - Interaction between RIP and a Hieracrchical Routing Protocol

In this case, lets suppose that the metrics used in the level 2 backbone in the hierarchical routing domain are kept separate from the metrics used for level 1 (intra-area) routing. Thus the metrics are not comparable. Let’s also assume that the level 2 routers which have links directly into a level 1 area (such as router W in figure 3) receive the level 1 LSPs, extract the reachability entries to determine what is reachable in the area, and then include this information in their own level 2 LSPs with a metric cost of 1. Thus, if subnet "19.234.*" is reachable via router H, router W finds this out from the level 1 LSP transmitted by router H (forwarded via router E), and includes this in its own (level 2) LSP with cost 1. Router W therefore informs router A (in its RIP packet) that it has a route to "19.234.*" of cost 1. Router A therefore announces a route of cost 2. Router B announces a route of cost 3. Thus, router E discovers a route of cost 4 via router B, which is better than the route of cost 5 via router F. Again a loop is created.

In this case, because the metrics are not comparable between the level 1 and level 2 routers, a different means (other than metric comparison) must be used to determine the relative preference of routes announced by level 1 and level 2 routers. The method used here is that the routing protocol used in the hierarchical routing domain specifies that level 1 routes are of a higher preference than level 2 routes. The level 1 routes and level 2 routes are therefore separate route classes (as defined in [1]). The relative values of metrics associated with each route entry is therefore only considered when comparing routes from the same route class.

In this case, the mistake comes from attempting to consider metrics from the RIP routing domain to be comparable with metrics used within both level 1 and level 2 routing. This has resulted in a list of route preferences that is not internally consistent (and therefore fails the third criteria in the list in section 2.1). In particular, a (level 2) route advertised by router W of cost 1, is considered higher precedence than a (RIP) route advertised by router A of cost 2, which is considered preferable than a (level 1) route advertised by router E of cost 4, which in turn is considered preferable to a (level 2) route of metric 1. The safety principle is violated, resulting in a loop.

2.3.3 TOS and Summary Routes

Figure 4 gives an example in which the safety principle is followed (thus no loops will form), but in which the routing progress principle is violated (resulting in a black hole)..

Figure 4 represents the interconnection of a private corporate network with a public backbone (for example, the public backbone might represent the NSFNET backbone or an NSFNET regional at some time in the future, after TOS routing becomes deployed).

In this example, we are assuming that TOS routing is prefered over best match. Thus, a router which provides a poorer match route (e.g., network number match only) but which advertises the ability to support a particular TOS class is preferred to a router which provides a better match (e.g., subnet match) but which advertises only the default metric.

In this example, the private network is running a link state routing protocol. The private routing domain is just being updated to support TOS routing, such that most but not all of the routers support TOS routing. In figure 4, routers R, S, and T (as well as the two un-named routers) support TOS routing, but the older routers X and Y do not (they route only according to a default metric, and ignore the TOS field in forwarding IP packets). The public network supports TOS routing.

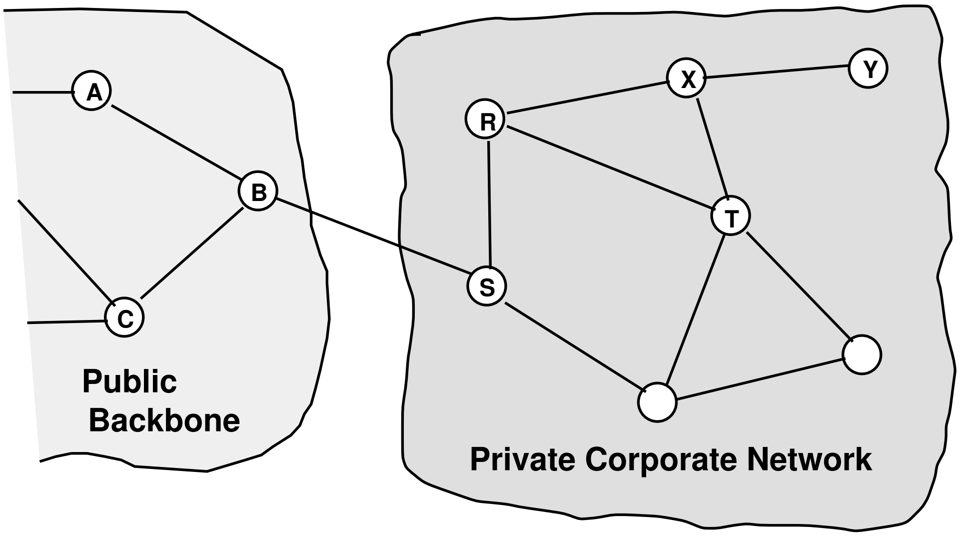

Figure 4 - TOS example

Within the private domain, a "default" route (with a subnet mask of all zero) is used to advertise the route to the public backbone. Given that the public backbone supports TOS routing and that most of the routers within the private corporate network supports TOS routing, naturally this default route is advertised with TOS-specific metrics for each of the TOS service classes supported.

Now, let's consider the subnet "21.15.*" which is reachable via router Y. In particular, lets consider a packet whose destination address is "21.15.127.234", which is requesting low-delay service, and which is being routed from router C in the public backbone. In this case, lets suppose that network number 21 corresponds to the private network, and therefore router S is advertising (in BGP packets sent to router B) that they can reach network 21. Router B therefore has included in its LSPs that it can reach network 21, and router C has no problem forwarding the IP packet to router B, which in turn forwards it to router S.

Now we discover the problem: Router S is advertising into the private corporate network that it can reach any IP address with a (relatively high precedence) TOS-specific route. Router Y can in fact reach the destination subnet directly, but only with a relatively lower precedence default-TOS route. Thus, the intermediate routers such as R will not forward the IP packet to the destination, but rather will forward any IP packet with a specfic (non-default) TOS request to S (unless they have a TOS-specific route with a better match). In this example, there is not way to forward any TOS-specific traffic to routers X or Y.

The problem here is that the routing progress principle is being violated. A higher preference route (TOS specific default route) is being advertised by a router which has a lower prefernce route to the destination (subnet route with the default TOS), but which has not higher preference route to the destination.

There are a lot of ways to fix this problem. The simplest is to state that best-match routing has preference over TOS routing (this approach is being adopted by both OSPF and Integrated IS-IS). Another simple, although more restrictive, solution is to not allow any summary TOS-specific routes to be advertised until ALL routers with more specific routes which match the summary route have been upgraded to support TOS routing, AND also have TOS-metrics defined for each of their links.

2.3.4 The Moral

These examples, although over-simplified, may be somewhat confusing to follow. However, they illustrate a rather simple point. The problems have occurred either because it was not possible to order the preference of routes to a single destination, or because routes are leaked from a lower preference entry to a higher preference entry, or because a router is advertising that it can reach a "summary address" corresponding to a range of destinations, when for some of the destinations the router had either no route available or only a lower precedence route available.

3 A Proposal for Router Configuration Options

It is not feasible to specify exactly what configuration options must be available for all IP routers. Rather, we propose that a "minimum subset" capability be defined. If accepted (and therefore specified in an Internet Standard) all IP routers would be required to support a minimum subnet of configuration options, and would be permitted to support other options as vendor-proprietary enhancements.

The proposed minimum subset provides a configuration method which provides routes consistent with the "safety principle" outlined above.

3.1 Configuration of Route Preference

Rules for setting relative preference between routing domains

- Set relative preference for each {domain & class} combination (call this a "routing nugget")

  - general preference level for the nugget

  - preference level for a particular destination network only for a particular nugget (overrides above general level)

  - preference level for a particular destination subnet for a particular nugget (overrides above)

  - preference level for a particular host route for a particular nugget (again, overrides above)

- When preference levels are separately set for different nuggets corresponding to route classes within a single routing domain, the levels must be consistent with the rules for the routing protocol used in that domain. If set equal, then the rules for the routing protocol are used [relative preference levels for each route class will be in the same order as specified by the routing protocol]. If set unequal, then the preferences set for each route class must be consistent with the routing protocol.

  Example: In some cases, it might be desired to "interleave" route classes from multiple domains. For example, Integrated IS-IS uses the relative preference ordering of route classes: (i) level 1 routes; (ii) level 2 routes with internal metrics; (iii) level 2 routes with external metrics. If a particular router belongs to two routing domains A and B, then the relative preference might be set to: (i) (highest preference) level 1 routes from domain B; (ii) level 2 routes with internal metrics from domain B; (ii) level 1 routes from domain A; (iv) level 2 routes with internal metrics from domain A; (v) level 2 routes with external metrics from domain B; (vi) level 2 routes with external metrics from domain A.

- OPTIONAL EXTENSION: If two nuggets are set to an equal preference level, then metrics, best match, or other route characteristics may be compared directly to compare routes provided by each nugget.

  This optional extension might not be always feasible for use. If the preference levels are set equal for two different routing domains, which use different routing protocols, then it may be "non-trivial" to determine how to compare routes from the two routing protocols. More work is needed to specify how to do this.

3.2 Configuration of Route Leaking

<This section is for futher study>

4 Reduction in Forwarding Database Size

It is necessary to be able to create a forwarding database that is of reasonable size

Assume all routing protocols use best match, have consistent TOS approach (relative precedence of TOS versus best match). Assume all route classes from all routing domains can be ordered in preference (no two route classes have equal preference).

Assume with any two route entries ***A*** and ***B***; either range of ***A*** and ***B*** are disjoint, or range of ***A*** and ***B*** are identical, or ***A*** includes ***B***, or ***B*** includes ***A***. [explain]. This means that ***A*** and ***B*** can't overlap in a general manner. Note that this must be true if you use [address, subnet mask] for route ranges with contiguous subnet masks. This is not necessarily true if (i) use non-contiguous subnet masks (example from RFC 1195, Annex C); (ii) use address ranges (e.g; ***A*** = {networks 2, 3, and 4}; ***B*** = {networks 4, 5, and 6}).

How to create one forwarding database per TOS class:

1) initialize forwarding database :== Null

2) loop over all available route classes, starting with highest precedence route class

   a) loop over all routes ***i*** available in that route class

      (i) if (less specific route ***j*** is in forwarding database, such that ***j*** includes ***i***) ignore entry ***i*** (go to loop)

      (ii) else, if (equally specific route ***j*** is in forwarding database, such that range of ***j*** == range of ***i***) ignore entry ***i*** (go to loop)

      (iii) else, add entry ***i*** to the forwarding database

      end loop

   end loop

3) forwarding database is now correct, forward using best match and TOS policy

For more general case: The inner loop becomes more complex. For overlapping route classes: at each step, if there is an overlapping route class already in the forwarding database, it is necessary reduce the range of the new route entry ***i*** by the amount that it overlaps with existing entries before adding to forwarding tables. In general, this can be relatively complex.

For multiple route classes with equal precedence: This may be complex, depending on the resolution of the issue mentioned in the "Optional Extension" note at the end of section 3.1.

5 Security Considerations

Security considerations are not discussed in this document.

6 Author's Address

Ross Callon
Digital Equipment Corporation
550 King Street, LKG 1-2/A19
Littleton, MA 01460-1289
508-486-5009

7 References

[1] "Ruminations on the Next Hop", Philip Almquist, July 1991.

[2] "Ruminations on Route Leaking", Philip Almquist, July 1991.