Accredited Standards Committee
X3, INFORMATION PROCESSING SYSTEMS*

Doc. No.: X3S3.3/92-162R2
Date: 22 June 1992
Project:
Ref. Doc.:
Reply to: David Oran
Digital Equipment Corp.
MS LKG1-2/A19
550 King Street
Littleton, MA 01460
oran@sneezy.lkg.dec.com

To: X3S3.3

From: Dave Oran (DEC), Yakov Rekhter(IBM), Sue Hares(Merit)

Subject: Technical Design for Routeing Information Exchange between ISIS and IDRP

Enclosed is a technical proposal developed by the three of us which gives a design for how ISIS and IDRP interact to exchange routeing information at a routeing domain boundary. It also addresses a number of configuration and management issues such as how BISs can use ISIS to find each other and how resource overloads introduced by their interaction can be dealt with.

I would like to see this material discussed with the goal of submitting a contribution to the July SC6 meeting, proposing modifications to CD 10747, and necessary tweaks to ISO 10589 (via a technical amendment). This could come as a US contribution, or if deadlines are a problem, as a joint submission by the ISO 10589 and DIS 10747 editors.

Interactions between ISIS and IDRP

1 Introduction

This document covers the following technical issues concerning the interaction and routeing information exchange between ISIS and IDRP:

1) Extraction of Intra-domain Routeing information (specifically Area Addresses and Reachable Address prefixes) from ISO 10589 for use by IDRP as NLRI (Network Layer Reachability Information).

    a) Policy on using the ISIS internal metric to create the IDRP multi-exit discriminator.
    b) Handling of statically-configured Inter-domain routeing information expressed as ISO 10589 Reachable Address Prefixes.

2) Tunneling (encapsulation) of packets BIS-to-BIS through an RD using ISIS

3) Importation of Routeing information from IDRP's NLRI into ISIS's Reachable Address Prefixes.

    a) Policy for which and how many NSAP address prefixes in IDRP's NLRI will be imported into ISIS as Reachable Address Prefixes.
    b) Policy for whether NLRI is used by ISIS with internal or external metrics.

4) BIS Discovery using ISIS

5) Decision on whether/how to handle partitions of a Routeing Domain.

Other possible areas for work are not addressed here. They include:

1) Possible piggybacking of IDRP on ISIS L2 LSPs, thus using the ISIS flooding mechanism for all BIS-BIS communication inside an RD. This might be a performance win on transit RDs with few interior ISs.

2 Extraction and Summarization of ISIS Information into IDRP

In order for IDRP to route to destinations inside its local routeing domain, IDRP needs some source of information which represents the dynamic state of routeing domain, i.e. what destinations exist in the domain and of those, which are currently reachable. In the case of a system which is both an ISO 10589 Level 2 IS *and* an IDRP BIS, IDRP can obtain this information directly from the routeing information maintained by ISIS. ISIS provides this routeing information in the form of Area Addresses and Reachable Address prefixes carried in the Level 2 LSPs of ISIS. IDRP can use this information to construct NLRI to represent destinations in the local routeing domain.

The extraction of ISIS routeing information for propagation by IDRP as NLRI is controlled by one policy variable, a set of pre-configured prefixes, and a L2 cost threshold for each supported ISIS metric. ***NOTE: Charlie Kunzinger has some reservations about the specificity of this design which we should discuss — see his comments send via email.***

The policy variable is represented as an attribute of the MO containing the IDRP global management parameters. It has three settings:

a) Automatic summarization
b) Pre-configured Summarization
c) No Summarization

If the setting is "No Summarization", then IDRP extracts all of the Area Addresses in the RD from the "Destination Area" managed objects of the IS and announces them as IDRP NLRI. [1]

The default setting of the policy variable should be "No Summarization" for safety of not reporting routes to destinations that are in fact not in the local routeing domain.[2]

If policy variable says "Automatic", IDRP uses the RDI for the local routeing domain (the localRDI attribute of the iDRPConfig MO) as a template for what destinations to announce as NLRI. IDRP scans all of the area addresses in the L2 LSPs and announces:

1) a prefix equal to the longest common prefix of all of area addresses which match the localRDI of the routeing domain. This covers the common case of a routeing domain which has area addresses taken from a common addressing assignment from one authority, and uses one of its these addresses as its own RDI.

2) In addition, any area addresses which do not match the localRDI are announced individually.

If policy variable says "preconfigured", IDRP has a set of preconfigured prefixes which it is willing to announce as NLRI[3]. If any area address in the set of L2 LSPs matches a prefix, then IDRP announces that prefix; otherwise it does not. This deals with the case where a RD has address assignments taken from a number addressing authorities (e.g. a corporate network with area addresses taken from the French and Botswanan addressing authorities).

In addition to announcing the prefixes obtained via automatic or preconfigured summarization, it is possible to also announce individual area addresses. This is accomplished using the L2 Cost Threshold attributes. By discriminating the announcement of individual area addresses using the ISIS L2 metric(s) for the area, IDRP can direct the entry of traffic into the RD for areas that are "close"[4].

If the internal ISIS metric for an area falls below the threshold, IDRP advertises the area address individually, in order to optimize the entry of traffic. If the metric falls above the threshold, IDRP does not advertise that individual area address and lets it be covered by the summarization prefix. Longest match routeing will ensure that the traffic comes in the better way. This solves the well-known "east coast/west coast" problem illustrated below in figure 1. A, B, and G are BISs each in its own area $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{G}$ respectively. C, D, E, and F represent both L2 ISs and the areas in which they reside. If the L2 Cost Threshold is set to 7 on each of the BISs, then the following announcements will be made in IDRP NLRI (considering only the case of the ISIS default routeing metric):

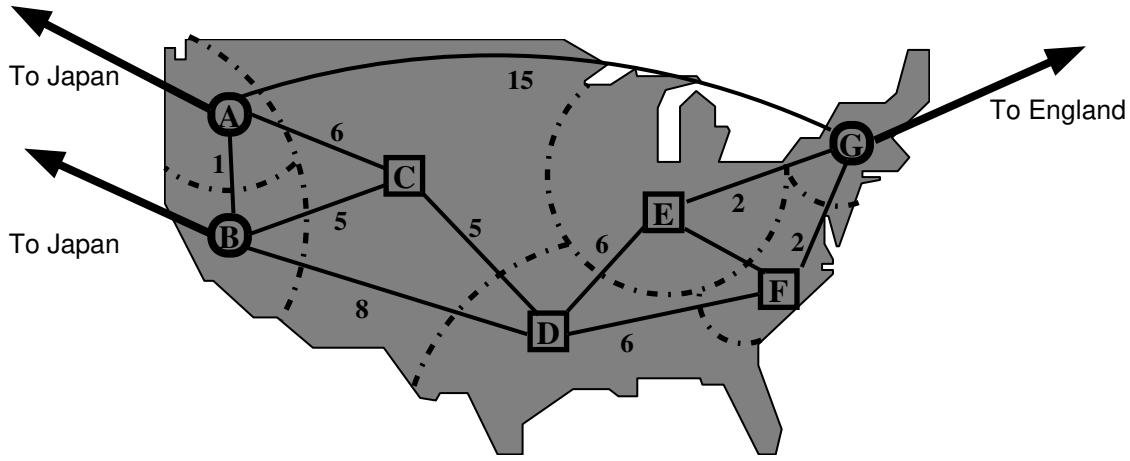

Figure 1 - Route Advertisement using Level 2 Cost Threshold

[1]This isn't currently possible with the ISO 10589 MO definition, since the Destination Area MOs contain both real areas and external destinations with no easy way to tell them apart. Therefore, either we fix this by adding a new "External Destination" MO and put the information from the Reachable Address MOs in there, or resort to pawing through the LSP database directly. For the remainder of this document, we'll assume we have to paw through the LSPs.

[2]If you don't summarize, the protocol will be able to reach most internal destinations even if the RD is partitioned but this is not deemed sufficient reason to avoid summarization.

[3]These can be represented as set of IDRP MOs, or a set-valued attribute of the global IDRP MO — it doesn't really matter

[4]If L2 Cost Threshold is set to zero, then only the area in which the BIS resides is announced individually. Some encoding needs to be defined to completely disable the use of this feature if announcement of the BISs own area is not desired.

BIS	Areas Announced
A	$\mathcal{A}, \mathcal{B}, \mathcal{C}$
B	$\mathcal{A}, \mathcal{B}, \mathcal{C}$
G	$\mathcal{G}, \mathcal{E}, \mathcal{F}$

Traffic inbound to the U.S. from Hong Kong might come in either to San Francisco (BIS A) or Los Angeles (BIS B) via Japan. Assuming the Inter-domain paths are of the same preference, then by announcing area $\mathcal{E}$ individually, traffic to Akron is directed through Boston rather than through San Francisco or Los Angeles. Conversely, by not announcing $\mathcal{D}$ traffic inbound to Dallas shows no preference for any of the individual entry points.

When announcing the individual areas, BISs A and B will include the MULTI_EXIT_DISC path attribute, so that the Japanese RD can discriminate between the inbound paths through San Francisco and Los Angeles. In the above example, the Japanese RD would choose BIS B over BIS A for traffic inbound to area $\mathcal{C}$ since it has the lower Intra-domain metric value. When reporting the MULTI_EXIT_DISC path attribute, IDRP multiplies the L2 ISIS metric by 4 to account for the difference in dynamic range of the two metrics.

IDRP is also permitted to announce NLRI for destinations that were statically configured via ISIS Reachable Address Prefixes. This is necessary, if only for the graceful introduction of IDRP into Routeing Domain, but may be needed until such time as all L2 ISIS implementation platforms support IDRP. Statically configured Reachable Address Prefixes from ISIS are imported into IDRP as follows[5]:

1) If an ISIS L2 LSP indicates that the IS is *not* a BIS (see BIS Discovery, below), then IDRP imports the prefixes in the L2 LSP into its decision process. If the LSP indicates that the L2 IS *is* a BIS, the information in the reachable address prefixes is not imported.[6]

2) The BIS sets the IDRP Path Attribute EXTERNAL_INFORMATION in indicate that the information describes external destinations not originating from IDRP.

3) If the BIS supports multiple routeing metrics, a separate path is imported for each of the supported routeing metrics. The degree of preference for each is set to the value of the corresponding ISIS metric assigned to the Reachable Address Prefix.

****NOTE: Charlie Kunzinger had some trouble understanding the above and tried to rewrite the last bit in his email. I found what's here easier to understand so we should try to see if there's a better way to write this that both of us understand! ***

3 Tunneling of Interdomain Traffic Through the RD

By default, IDRP and ISIS are totally independent of one another, in the sense that neither relies on the other to ensure correct routeing of data PDUs. IDRP BISs in different routeing domains are by definition connected to each other over a shared subnetwork. On the other hand, BISs in the same local routeing domain do not necessarily share a subnetwork and use the Intra-domain routeing protocol to ensure communication among themselves. IDRP therefore needs some way to get NPDUs from one BIS to another in the local RD without involving ISIS[7]. One way to accomplish this is to have IDRP provide the necessary routeing information to ISIS: this is discussed in clause 4. In many circumstance however, it is best to forward NPDUs through a routeing domain by encapsulating them inside another NPDU destined for the exit BIS as the network layer address. This form of Inter-domain forwarding is known as *tunneling*.

Tunneling is done using encapsulation as described in the current IDRP specification, with the exception that text needs to be added to IDRP to ensure that if you are using ISIS to do the tunneling, that you follow the procedures in ISIS for "encapsulation of traffic for partition repair".

[5]The representation of the ISIS static routes within IDRP can be accomplished by creating a pseudo-Adj-RIB-In and placing the NLRI there as if it had come from a real BIS. Other implementation techniques are of course possible.

[6]The reason is that both ISIS and IDRP in one BIS obtain statically configured routing information from the same Managed Objects — the Reachable Address MOs. Hence any information in the L2 LSP Reachable Address Prefix fields would be redundant with information already obtained via IDRP BIS-BIS protocol exchange.

[7]ISIS would not be able to forward these NPDUs because the destination NSAP is outside the local routeing domain and only IDRP has the necessary routeing information to determine the next Inter-domain hop for the NPDU.

IDRP needs to manipulate the ISIS management information to control the forwarding of packets through the domain, such that ISIS and IDRP can have a unified management representation of the FIBs used by the forwarding machine(s). The ISIS "Adjacency", "Circuit", "Virtual Adjacency" and "Reachable Address" managed objects are used to contain the shared state between ISIS and IDRP as follows:

1) Inter-RD BIS-BIS links are distinguished from Intra-RD links by marking the Circuit MO for Inter-RD links as "external domain". To handle the case of "DMZ" subnetworks (multi-access subnetworks with some internal and some external neighbors — DMZ is an acronym meaning “de-militarized zone”), the network manager can create multiple Circuit MOs and point them all at the same SNPA managed object.)

2) The Circuit MO has an Adjacency MO for each BIS neighbor on the circuit. These can be either manually created via configuration information (to control the Inter-RD topology directly), or could be created automatically by running ESIS on the circuit. (This latter method hasn’t been worked out yet — it isn’t clear if it would work properly on DMZ subnets.)

3) For each BIS reachable in the local RD which does not have a real adjacency (i.e. is not an ISIS neighbor of the BIS), IDRP creates a "Virtual Adjacency" MO to represent a tunnel to that BIS. (This is done even if the intent is to import all IDRP NLRI as ISIS Reachable Address Prefixes since under some conditions it may be necessary to stop importing the NLRI and use tunnels — see the discussion of ISIS overload below).[8]

4) IDRP create/updates Reachable Address MOs each time it recomputes its FIBs. It creates one or more Reachable Address MOs for each Inter-RD circuit, and puts in the "Address Prefix" attribute the prefixes for the destinations which are being forwarded to over that circuit. The Reachable Address MO is left "disabled" for destinations for which tunneling is being done. It is Enabled by IDRP when paths to the external destinations are to be computed by ISIS (see below for the details on how to decide whether to tunnel or import the NLRI). Note that this technique allows the existing management capability of ISIS to be used directly to model the complete forwarding state of the IS, since the union of the destinations in the Adjacency, Destination Area, and Reachable Address MOs represents the entire contents of the IS’s FIBs.

4 Importation of NLRI from IDRP into ISIS Reachable Address Prefixes

In order to avoid the processing and bandwidth overhead imposed by tunneling, it is desirable to allow IDRP to supply routeing information to ISIS. To accomplish this, ISIS must become aware of destinations outside the local routeing domain through a more robust means than statically-configured reachable address prefixes.

IDRP creates an ISIS Reachable Address MO as described above for each NLRI to be supplied to ISIS — L2 ISIS then simply computes Intra-domain routes to these destinations as with any other reachable address prefix. These Intra-domain routes allow the NPDUs to reach the proper exit BIS. The issue is to decide which NLRI to supply to ISIS. This is done as follows.

There is an IDRP MO attribute maximumNLRItoISIS which sets an upper bound on the NLRI which will be supplied to ISIS by the BIS. If the BIS has more destinations than maximumNLRItoISIS it is a local decision which ones get supplied to ISIS. ***NOTE: Charlie Kunzinger believes I have this backwards — that this attribute ought to belong to ISIS rather than IDRP. I have no strong feeling on this, except to say that it’s easier to modify IDRP than ISIS at this stage, and which one “owns” the attribute has no practical effect on the rest of the design or the implementation. What do the rest of you think? ***

The BIS has preconfigured information for potential NLRI to supply to ISIS, which consists of a set of tuples {prefix, supplyLonger,[metric type, [multiExit], [metric values]}. supplyLonger is a boolean. multiExit is an existing attribute of the iDRPConfig managed object. "Metric type" is an enumeration {default/internal/external}. "Metric Values" is a set of legal ISIS metric values, one for each supported ISIS metric. ***NOTE: Charlie Kunzinger believes that this is backwards — that this information belongs with the RAPs of ISIS rather than duplicated in IDRP. I don’t see how to accomplish the supplyLonger capability by simply adding attributes to RAPs. Perhaps we can work out a way that the actual prefixes don’t have to be duplicated in static information owned by both ISIS and IDRP. Any good design ideas here gratefully accepted!***

[8] It may also be useful to use the Adjacency or Virtual Adjacency MO as the management handle for BIS-BIS connection state, by attaching an IDRP conditional package to the Virtual Adjacency MO.

If supplyLonger is TRUE, the BIS may supply any NLRI whose prefix matches the configured prefix. If supplyLonger is FALSE, the BIS may supply only NLRI whose prefix is identical in length (and value) to the configured prefix.

The BIS has preconfigured information of the form {RDI,Internal/External} for each neighbor RD which determines the default value for the metric type to use for routes whose next hop is that Routeing domain. When supplying NLRI to ISIS, the ISIS metric type is set according the "metric type" of the tuple:

1) If the "metric type" associated with the prefix is "default", then the ISIS metric type is taken from the preconfigured information for the RDI of the next hop routeing domain.

2) Otherwise, the ISIS metric type is set from the "metric type" associated with the prefix.

Note that supplying overlapping NLRI with inconsistent metric types (internal vs. external) may result in ISIS computing a sub-optimal exit point for the traffic, thus forcing IDRP to tunnel the traffic to the correct exit point.

The metric value is set as follows:

1) If multiExit is TRUE, then the metric value is taken from the MULTI_EXIT_DISC path attribute of the NLRI being supplied. When setting the ISIS metric value, the MULTI_EXIT_DISC value is divided by 4 and rounded to reflect that the IDRP value has four times the dynamic range of the ISIS metric value.

2) Otherwise, the metric value is set to the value associated with the preconfigured prefix.

The BIS has a preconfigured boolean expression which tells it whether or not to import an NLRI representing “all possible destinations” into ISIS. The expression is a list of prefixes connect by boolean AND or OR operators. The reachability information is matched against the prefixes using shorter prefix matching, and if the expression is TRUE, IDRP supplies a reachable address prefix for all possible external destinations (i.e. the null prefix).

A BIS may supply NLRI only for routes that it received from external BISs and that are present in the BIS’s Loc-RIB. Once a BIS determines (by means of IDRP) that a previously supplied NLRI is no longer available, the NLRI needs to be withdrawn from ISIS by disabling the reachable address managed object.

5 Policies for Deciding Which NLRI to Actually Import into ISIS

Amongst the routes in the RD, the BIS has to choose which ones, up to maximumNLRItoISIS to supply to ISIS. One approach is to sort the prefixes by length and prefer either shorter or longer prefixes. A second is to assign to each prefix a preference value and sort by preference. Neither of these techniques is coupled to the actual traffic matrix, however, and the preference value technique requires extra configuration information and traffic history analysis by the network manager in order to be more effective than a random selection technique. A more difficult, but adaptive technique is to actually observe the traffic to decide which NLRI to supply. This could be done as follows:

1) When a packet arrives from an Intra-RD circuit to be forwarded outside the RD, it is handed to the forwarding machinery, which does a FIB lookup to find the next hop. The FIB entry has a "back pointer" to the reachable address MO containing the longest matching prefix representing this destination (this needs to be known one way or another in order to guarantee longest match routeing as mandated already by ISIS).

2) Each reachable address MO has an LRU timer associated with it.

3) If the packet arrived directly (i.e. not over a virtual adjacency) forwarding proceeds normally, with the extra step of resetting the LRU timer.

4) If the packet arrived over a virtual adjacency and was decapsulated, then look at the state of the corresponding reachable address MO. If the reachable address is currently enabled, reset the LRU timer as described in (3).

5) If the reachable address is currently disabled, the value of maximumNLRItoISIS is not exceeded, and the overload state is not set (see below), then enable the reachable address prefix. This will cause the routeing information for this destination to be propagated through ISIS, and once ISIS converges subsequent packets will arrive directly and not over the virtual adjacency.

6) To garbage collect inactive NLRI, when the LRU timer exceeds the garbage collection threshold, the reachable address MO is disabled to remove the inactive routeing information.

6 Dealing with ISIS Memory Overload

It is possible that the union of all of the supplied routes from all of the BISs causes some L2 IS in the RD to become overloaded. (The overload might be due to some other transient/permanent problem, but the BIS can't distinguish *why* the overload happens, so we assume that the imported NLRI are at least part of the problem). In this case it is desirable that the BIS reduce the memory load on ISIS by reducing the number of reachable address prefixes enabled. The overload is detected by an interface between ISIS and IDRP to report the overload. A good way of doing this is to add an attribute to the ISIS package of the cLNS MO which is set true if a scan of the L2 LSP database by ISIS detects that at least one IS in the RD is overloaded, and to have ISIS signal IDRP when it detects the condition. At this point IDRP does the following:

1) Immediately removes some imported routeing information by disabling one or more of the Reachable address MOs. ISIS notices this immediately (as specified in ISIS) and will reissue the L2 LSP removing the corresponding NLRI.

2) IDRP sets a timer equal to the ISIS value "Waiting Time". If this timer goes off and the Overload state is still set, IDRP disables more Reachable Address MOs and resets the timer for another pass. If the overload state is no longer set, IDRP re-enables the Reachable Address MOs to start supplying the NLRI again.

An interesting design issue is just how elaborate and adaptive this machinery ought to be. The simplest approach is to disable all the reachable address MOs, and hence revert to tunneling for all destinations. Like all dynamic schemes, oscillation can result. The frequency of the oscillation is damped by the "Waiting Time" timer, which is sufficiently long to ensure that ISIS converges and some useful routeing occurs over the RD. A simple extension is, of course to first disable all Reachable Address Prefixes routes whose LRU timer (see above) is longer than a minute or so.

One possibility for a more elaborate adaptive scheme is to use the "Multiplicative Decrease/Additive Increase" technique defined to handling congestion control. If this is employed, the BIS, instead of immediately supplying all of the NLRI it is permitted to import (according to the maximumNLRItoISIS attribute), uses a "slow start" technique of supplying NLRI for one (or a few) destination(s), waiting for a while (a timer shorter than "Waiting Time" but longer than the L2 ISIS convergence time), and then additively supplying more NLRI. If the overload is detected, the BIS multiplicatively reduces the number of destinations it supplies, using the algorithm described above, and choosing the Reachable Address Prefixes with the oldest LRU timers first. The value of the decrease factor should probably be smaller than the .875 used by congestion control, since the effects of overload are *much* more serious than the effects of larger-than-optimum transport windows. It may need to be .5 or smaller, but we'd need analysis and simulation studies to pick a good value.

7 BIS Discovery

Auto-configuration of BISs in the same routeing domain can be done by piggybacking the knowledge of which L2 ISs in the domain are BISs on the normal ISIS flooding machinery. This is accomplished by defining a new option field in the level 2 LSPs to indicate that the reporting IS is a BIS. The presence of the option indicates that the system is a BIS; the value field carries the IDRP version number as an optimization to avoid version negotiation during IDRP connection establishment.

On receipt of a L2 LSP with the "I'm a BIS" option, the BIS creates an "adjacentBIS" managed object for that BIS if one does not exist already, and sets the bisNET attribute from the IS's source ID and lowest area number in the LSP. Then, open processing is initiated for that BIS, if necessary.

8 Routeing Domain Partitions

There are two sub-problems one could address:

1) Getting traffic from outside to the "correct" part of a partitioned routeing domain.

2) Using Inter-RD routes to heal an RD partition

There is general agreement, long-standing, that solving problem (2) is not necessary and would involve breaching the firewall between Intra- and Inter-domain routeing, which we don't want to do.

Solving problem (1) is potentially desirable, but you need first to be able to detect the partition. This is a hard problem. Therefore, all bets are off if an RD gets partitioned, except in the special case above where IDRP does no summarization.