Title: Details on Usage of NEXT_HOP Attribute

Source: Tsuchiya, Rekhter, and Kunzinger

Reference: X3S3.3/91-6, ″Proposal to Keep NEXT_HOP in IDRP″

This paper expands on the proposed use of the NEXT_HOP attribute that was described in X3S3.3/91-6. This paper proposes text for a USA comment on IDRP which will define new encoding and usage rules for IDRP′s NEXT_HOP attribute to permit a ″route server″ function to be implemented on fully connected subnetworks (for example, public data networks or broadcast subnetworks). It will also permit a BIS to explicitly identify multiple SNPAs, and the order in which it would prefer its neighbor BISs to direct incoming traffic to each of them.

PROPOSED USA COMMENT

The NEXT_HOP attribute as it is defined in SC6 N6387 (IDRP) is a well-known mandatory attribute whose value is the NET of the BIS that originates the UPDATE PDU in which the attribute is contained (see clause 7.11.3 of SC6 N6387). Since the NET is either known at the time the BIS-BIS connection is established or can be ascertained from the NUNITDATA.INDICATION primitive, the attribute is redundant, and should not be retained in IDRP in its present form.

During discussions about this attribute, US experts developed a new optional use for this attribute which will require a new encoding and new usage rules. The USA recommends that the current NEXT_HOP description in SC6 N6387 should be replaced by new text which will define a route server function for IDRP, and will allow a BIS to explicitly list its SNPAs.

Rationale for New NEXT_HOP Attribute

A large public data network or a large broadcast subnetwork might contain hundreds or even thousands of BISs. Theoretically, each may want routeing information from any of the others. However, it may not be efficient for each BIS to establish a BIS-BIS connection with every other BIS in the subnetwork. It may be more efficient to establish a small number of ″IDRP route servers″. Then, each BIS can establish an IDRP connection with an IDRP route server, which will be capable of re-distributing the routeing information which it has learned from the other BISs.

For this scheme to provide the desired efficiencies, the IDRP route server must be capable of advertising a next-hop other than itself. In all other respects, it functions as a conventional BIS: it computes routes by applying its local policies to the information in its routeing information base, and it advertises those routes in accordance with IDRP procedures.

A Typical Application of NEXT_HOP

The new route server function is applicable to systems that are located on transitive fully connected subnetworks. We say that a subnetwork is transitive with respect to system reachability if all of the following are true:

1. Systems A, B, and C are all attached to the same subnetwork,

2. When A can reach B directly, and B can reach C directly, it follows that A can reach C directly.

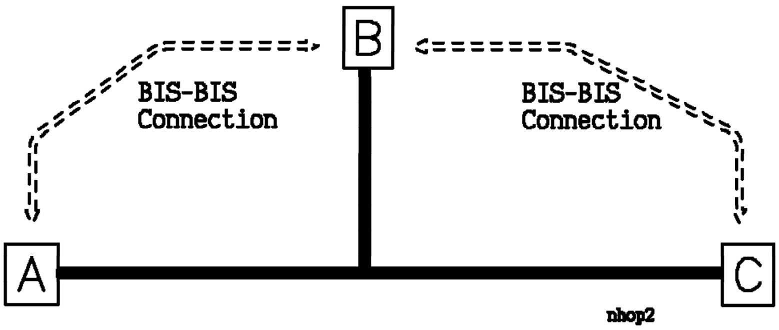
Verification of direct reachability can be accomplished by means outside of IDRP. For example, systems located on a common subnetwork could use an ES-IS protocol (such as IS 9542 or IS 10030) to ascertain if there is direct reachability between them. Examples of such media are IEEE 802.2, SMDS, and X.25.

Example Usages of the New NEXT_HOP

1. Route Server:

   Consider three BISs attached to a common subnetwork that has the transitivity property. Assume that A has an IDRP connection with B, and B has an IDRP connection with C, but A and C do not have an IDRP connection with each other. Let us call BIS-C the source BIS, BIS-B the first recipient, and BIS-A the subsequent recipient of the UPDATE PDUs. Clearly, IDRP in its current definition allows the first recipient (B) to receive an UPDATE PDU with routeing information from the source (C), and then advertise it in its own UPDATE PDU to a subsequent recipient (A), subject to B′s routeing policies and any propagation constraints that the source included in the path attributes of its UPDATE PDU.

   With the currently defined NEXT_HOP attribute, if the source (C) informs B of a route to a destination X, and the first recipient (B) announces the same route to a subsequent recipient (A), the NPDUs from A destined to X will first be sent to B, and then B will forward them to C (subject of course to B′s own policies and to any constraints contained in the path attributes of C′s original UPDATE PDU).



   Under the new usage rules for NEXT_HOP, the source (C) would inform B of a route to a destination X, indicating its own NET and an associated SNPA in the NEXT_HOP field of its UPDATE PDU. The first recipient (B) may then, at its option, elect to advertise the route to the subsequent BIS (A), but B would list the NET and SNPA of the source (C) in the UPDATE PDU that it sends to A. As a result of this advertisement, BIS-A would then send NPDUs destined for X directly to BIS-C , bypassing BIS-B.

   Clearly, there are no compelling reasons why BIS-A should not send NPDUs directly to BIS-C, since this would not conflict with any path attributes that C had advertised in its UPDATE PDU. Sending NPDUs directly to C will avoid an extra hop at B, thus conserving network resources and eliminating the need for BIS-A and BIS-C to establish a BIS-BIS connection between themselves.

   One might argue that if A is going to send NPDUs directly to C, then A and C should establish an IDRP connection between themselves. This may be feasible when the number of BISs involved is not large. However, in a public data network or a large broadcast subnetwork, this may not be practical due to the large number of BIS-BIS connections that would be needed if each BIS were required to have a BIS-BIS connection to each other BIS. (Note, for example, that DIS 10589 uses the concept of a ″LAN Designated IS″, or ″pseudonode″, to minimize the amount of routeing traffic that flows on a LAN.)

In some public data networks, it may be desirable to allow one IDRP server to advertise routes to another IDRP server, which in turn advertises it to yet another IDRP server, etc. This would be useful, for example, so that all BISs in a public data network would not have to converge on a single server. As long as all the IDRP servers are on a common subnetwork and are directly reachable from one another, the rules given below do not preclude this sort of operation.

Thus, both performance (elimination of an intermediate hop) and efficient use of resource (minimum BIS-BIS connections) offer compelling reasons to allow a BIS to advertise the NET of some other BIS in its UPDATE PDUs. When the NEXT_HOP attribute is used to provide this sort of redirection capability, it is necessary that all three BISs are located on a common full-duplex subnetwork, and that they are all directly reachable by one another.

2. Advertising Multiple SNPAs:

   This proposal will also recommend that the NEXT_HOP field should allow a BIS to explicitly list its SNPA(s). Once the SNPAs are listed explicitly, then it follows that all other path attributes advertised in an UPDATE PDU apply to them. This allows a BIS to express its own preference as to which SNPAs it prefers its neighbors to direct inbound traffic to, in the case where it has several SNPAs attached to the same common subnetwork.

   For example, a given BIS may wish to receive inbound NPDUs over multiple SNPAs, perhaps as a form of load balancing. When multiple SNPAs are carried in a single UPDATE PDU, they are all considered to be equally preferable. In this case, a neighbor BIS may decide to send traffic to each of them on a round-robin basis.

   However, if all its SNPAs are not equivalent, then the BIS may choose to advertise several UPDATE PDUs, where each lists a different set of equivalent SNPAs. For example, by using different values of the MULTI-EXIT_DISC attribute in each UPDATE PDU, the advertising system could also express its preference about which SNPAs are most desirable. A neighbor BIS would then have this information available to its Decision Process, and could factor the advertising BIS′s preferences into its own local decision-making process.

   Thus, by allowing explicit advertisement of SNPAs, performance improvements can be obtained, since a preference among a set of SNPAs can now be expressed.

Proposed Changes to SC6 N6387

We propose that the NEXT_HOP attribute be retained, and that IDRP should support an optional function which allows a given BIS to advertise a BIS other than itself in this field. We also propose that the NEXT_HOP attribute allow explicit listing of SNPAs. The new usage rules for the NEXT_HOP attribute accomplish two purposes:

1. They allow a BIS to unambiguously identify the preferred SNPA(s) to be used for inbound traffic. When there are multiple SNPAs all attached to the same subnetwork, it can also be used to achieve a form of load-balancing, based on the value of the MULTI-EXIT_DISC attribute associated with a given SNPA.
2. They allow the source of an UPDATE PDU to specify whether or not the first recipient is allowed to advertise the source′s SNPA as part of its own UPDATE PDU.

The following changes should be made in SC6 N6387:

1. The encoding of the NEXT_HOP attribute (in clause 6.3, item ″c″, on page 14) should be changed as shown below. In addition to listing the NET of the next-hop BIS, this attribute may contain one or more SNPAs. If multiple SNPAs are listed, it is understood that they are all equivalent from the viewpoint of the BIS that advertises them.

   IDRP_Server_Allowed: This is a one octet field. The value X′FF′ indicates the recipient of this UPDATE PDU has the option of advertising (in its own outbound UPDATE PDUs) the NET and SNPA information learned from this UPDATE PDU. If the value is not X′FF″, then the recipient of this UPDATE PDU shall not advertise the NET and SNPA information learned from this UPDATE PDU.

   Length of NET: A 1 octet field whose value expresses the length of the ″NET of Next Hop″ field as measured in octets

   NET of Next Hop: A variable length field that contains the NET of the next BIS on the path to the destination system

   Number of SNPAs: A 1 octet field which contains the number of distinct SNPAs to be listed in the following fields

   Length of First SNPA: A 1 octet field whose value expresses the length of the ″First SNPA of Next Hop″ field as measured in octets

   First SNPA of Next Hop: A variable length field that contains an SNPA of the BIS whose NET is contained in the ″NET of Next Hop″ field.

   ⋮

   ⋮

   Length of Last SNPA: A 1 octet field whose value expresses the length of the ″Last SNPA of Next Hop″ field as measured in octets

   Last SNPA of Next Hop: A variable length field that contains an SNPA of the BIS whose NET is contained in the ″NET of Next Hop″ field.

2. New usage rules should replace those currently listed in clause 7.11.3, as follows:

   NEXT_HOP is a well-known discretionary attribute that must be recognized upon receipt by all BISs.

   1. A BIS may choose not to include the NEXT_HOP attribute in its UPDATE PDU. In this case the SNPA must be learned from other means: for example, the value of the NET can be learned from the NUNITDATA.INDICATION, and IS 9542 can be used to associate an SNPA with that NET. When the NEXT_HOP field is not present, then it is understood that the BIS that advertises the UPDATE PDU is also the next-hop BIS.

   2. A BIS may choose to list its own NET and its own SNPAs in the appropriate fields of the NEXT_HOP attribute. It may set the value of the ″IDRP_Server_Allowed″ field in accordance with its local policies.

   3. A first recipient BIS that receives a route from a source BIS may advertise that route to a third subsequent recipient BIS. The first recipient BIS may list (in the NEXT_HOP field of its UPDATE PDU being sent to a subsequent recipient BIS) the NET of the source BIS and its associated SNPA (as contained in the UPDATE PDU received from the source BIS) when all of the following conditions are satisfied:

a. The ″IDRP_Server_Allowed″ field of the inbound UPDATE PDU was equal to X′FF′.

b. All three BISs (source, first recipient, and subsequent recipient) are located on a common subnetwork which is full-duplex.

c. The subnetwork has the transitivity property with respect to reachability of all three BISs.

d. The first recipient and subsequent recipient are located in different routeing domains.

e. Advertisement of this route to the subsequent recipient BIS does not conflict with any of the path attributes that were contained in the UPDATE PDU from the source BIS.

In this case, the advertising BIS may set the value of the ″IDRP_SERVER_Allowed″ field in the outbound UPDATE PDU in accordance with its own local policies.

A note should also be added to make it clear that the rules outlined above do not remove the requirement that there must be an IDRP connection between each pair of BISs located within a given routeing domain.

3. The text in clause 7.11.2, item ″b″ should be expanded to note that a BIS will always update the RD_PATH attribute with its own RDI regardless of the contents of the NEXT_HOP field. Since a BIS that receives the UPDATE PDU has enough information to determine if the NEXT_HOP BIS is located in the same RD as the BIS from whom the UPDATE PDU was received, it can factor this into its decision process when it determines its own degree of preference for the route.