Title: Additional IDRP Functions (Revised)

Source: C. A. Kunzinger

Reference: This document supersedes X3S3.3/90-260

This is the revised list of possible IDRP changes, based on the discussions that took place during the X3S3.3 meeting in December 1990. Within the Task Group, we should reach consensus on their merits, and the worthwhile ones should be included in the USA comments on SC6 N6387.

1. **More Descriptive Title**:

   ----ACCEPTED----

   The current title for SC6 N6387 can in many cases lead to confusion on the part of new readers: it is long, awkwardly constructed, and almost identical to the title of DIS 10589 (″inter″ vs. ″intra″).

   SC6 N6387 deals primarily with the protocol for exchanging routeing information, and it defines the rules which must be followed for advertising locally selected routes to other systems that are participating in an instance of IDRP.

   A slight rewording of the title can clarify this. The USA suggests that the title of SC6 N6387 be changed to *Protocol for the Exchange of Inter-domain Routeing Information among Intermediate Systems*.

2. **Abbreviated form of an unfeasible route**

   ----ACCEPTED----

   SC6 N6387 says that an unfeasible route is created by attaching the UNREACHABLE attribute to the previously feasible route. This implies that all the path attributes that have been present in the feasible route should be present in the unfeasible route as well.

   However, in IDRP each Adj-RIB has at most one route to a particular destination as specified in the Network Layer Reachability Information, and the distinguishing attributes of the route unambiguously identify a particular Adj-RIB. Thus, since the combination of the distinguishing attributes and the Network Layer Reachability Information allows one to unambiguously identify a previously feasible route the following shortcut is possible: to announce that some previously feasible route becomes unfeasible it is sufficient to supply only the Network Layer Reachability Information and the distinguishing attributes (rather than all the attributes) of the previously feasible route.

3. **Usage of the MULTI_EXIT_DISC attribute**

   ----ACCEPTED----

   SC6 N6387 provides no clarifying examples for the use of the MULTI_EXIT_DISC attribute. The USA submits the following illustrative example, and suggests that the editor include it in an informative annex:

   EXAMPLE OF MULTI-EXIT_DISC USAGE

   The MULTI-EXIT DISC attribute can be used to provide a limited form of multi-path (load-splitting), as is shown in the following examples.

   - Example 1 (see Figure 1 on page 2):

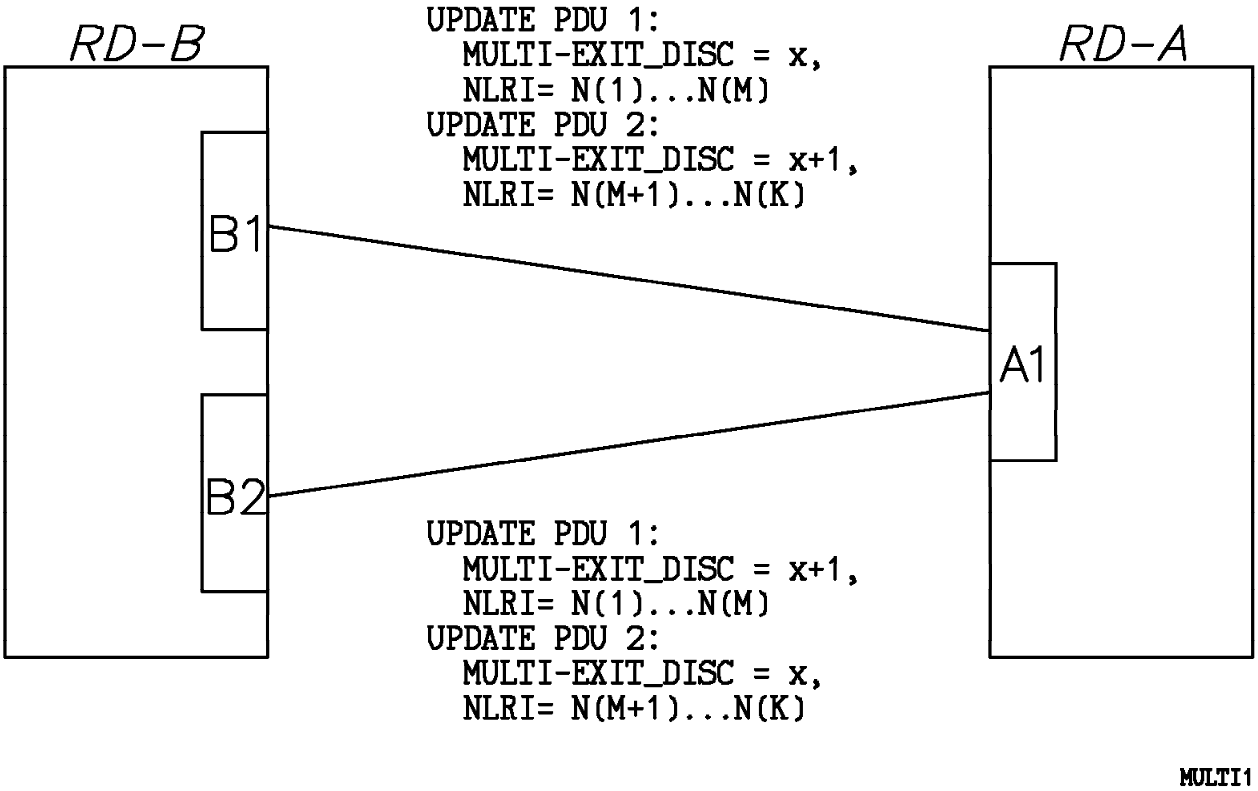

Figure 1. Example 1 Configuration

Consider the case when a BIS A located in routeing domain RD-A has two adjacent BISs (B1 and B2) that belong to the routeing domain RD-B. Assume that RD-B has Network Layer Reachability information about NSAPs N1, N2, ... Nk, and it wants to advertise this information to RD-A. By using the MULTI-EXIT_DISC attribute RD-B may do selective load splitting (based on NSAP addresses) between B1 and B2.

For example, BIS B1 advertises to BIS A Network Layer Reachability information N1, N2, ... Nm with the MULTI_EXIT_DISC set to X, and advertises N(m+1), ... Nk with the MULTI_EXIT_DISC set to X + 1.

Similarly, BIS B2 advertises to BIS A Network Layer Reachability information N1, N2, ... Nm with the MULTI_EXIT_DISC set to X + 1, and advertises N(m+1), ... Nk with the MULTI_EXIT_DISC set to X.

As a result, traffic from BIS A that destined to N1, N2, ... Nm will flow through BIS B1, while traffic from BIS A that destined to N(m+1), ... Nk will flow through BIS B2. This scenario illustrates the simplest way of doing limited multipath with IDRP.

- Example 2 (see Figure 2 on page 3):

  Next consider more complex case where there is a multihomed routeing domain RD-A that has only slow speed links. RD-A is connected at several points to a transit routeing domain RD-B that has only high speed links; BIS A1 is adjacent to BIS B1, and BIS A2 is adjacent to BIS B2.

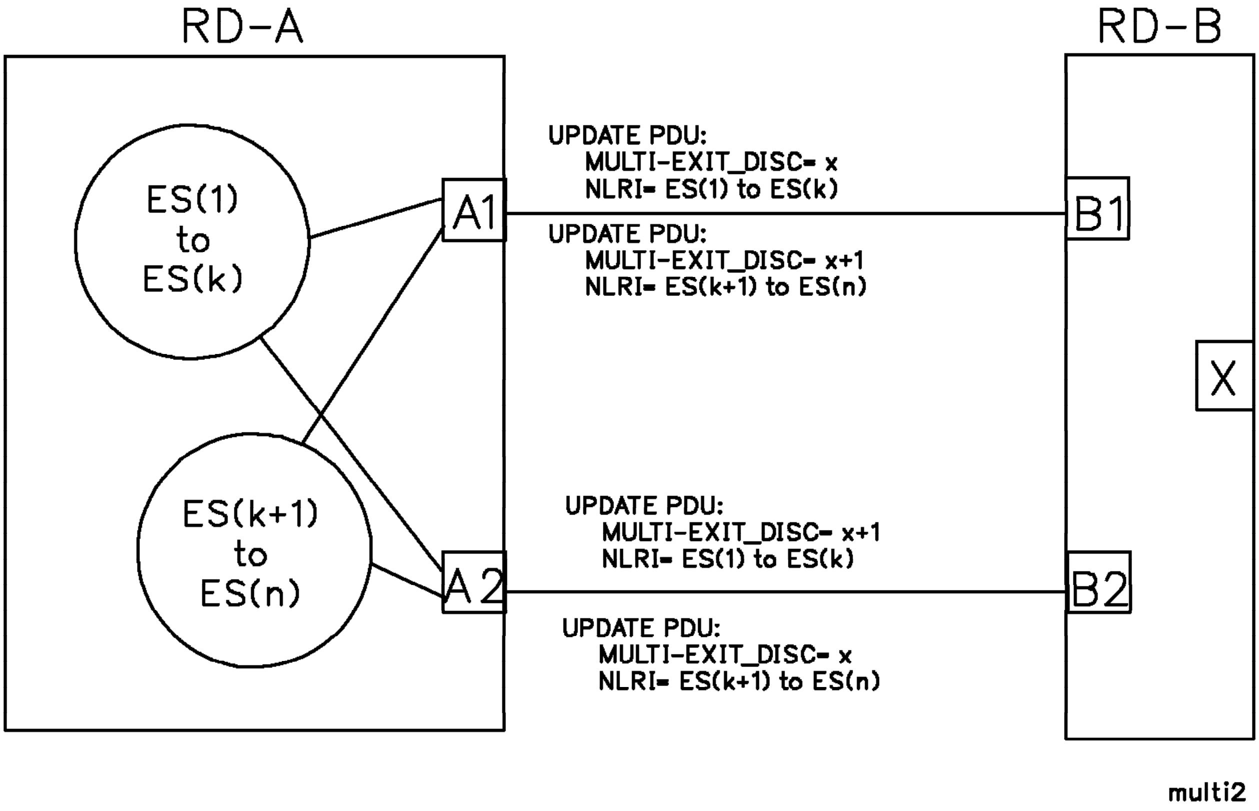

Figure 2. Example 2 Configuration

RD-A wants to minimize the distance that incoming NPDUs addressed to certain ESs—say ES(1) through ES(k)—will have to travel within RD-A.

One way of doing this is by making BIS A1 to announce to BIS B1 destinations ES(1) – ES(k) with a lower MULTI_EXIT_DISC, as compared to the MULTI_EXIT_DISC that BIS A2 will use when announcing the same destinations to the BIS B2. Similarly, BIS A2 would announce to BIS B2 destinations ES(k=1)&ndash,ES(n) within the RD-A that are closer to the BIS A2 (than to the BIS A1) with the lower MULTI_EXIT_DISC, as compared to the MULTI_EXIT_DISC that the BIS A1 will use when announcing the same destinations to the BIS B1.

When traffic that destined to some ES within RD-A enters RD-B on its way to RD-A via BIS X, X picks up the exit BIS that has the lowest MULTI_EXIT_DISC value for that destination. For example, X may pick up BIS A2 as an exit, even if the distance between A2 and X is greater than the distance between A1 and X.

4. **Source Routeing of 8473 NPDUs**:

----ACCEPTED----

The IDRP text on the forwarding process should be expanded to address the source route parameters of 8473 NPDUs. For example, there should be normative text stating that a complete source route (if present) takes precedence over the next IS contained in IDRP's FIB.

5. **NEXT_HOP attribute**

----REJECTED, BUT FURTHER DISCUSSION EXPECTED----

The task group recommended that this attribute be dropped from IDRP. However, a new contribution is expected which will advocate retaining this attribute, and expanding its role. Hence, further discussion is anticipated

6. **Use of ranges:**

----REVISED----

The USA recommends that ranges should not be supported in IDRP.

7. **Use of Masks:**

----NEW ITEM----

The task group was amenable to considering a recommendation to replace prefixes with masks, pending further study to establish that masks provided advantages without introducing undue complexity. A new contribution on masks is expected for our February meeting, and will be generated by Joel Halpern.

8. **Handling of the 8473 Security parameter in NPDUs:**

----ACCEPTED, MORE DETAIL REQUESTED----

The USA asks the editor to include material in IDRP on the handling of the 8473 Security parameter, and suggests that a new distinguishing attribute and associated usage rules be included.

The task group asked to see some more details on the method being proposed. The method is outlined in more detail below:

> *The usage rules should be similar to those already defined for SS-QOS and DS-QOS routeing: that is, it can parallel the text of clause 7.11.13-14 of SC6 N6387. The new path attributes would be type-value specific, and the matching process of attributes to the parameters in the 8473 NPDU would be based on the methods of IDRP's clause 7.15.3. It would be necessary to expand clause 7.15.2 so that NPDU-derived Distinguishing Attributes could be obtained from the 8473 security parameter as well as from its QOS-Maintenance parameter.*

9. **Interaction between Decision and Update Processes:**

----ACCEPTED IN PRINCIPLE, DETAIL TEXT REQUESTED----

There should be a constraint in IDRP that the Decision Process should run to completion based on currently available routes before any newly arrived routes are used in the computation. The following text takes an approach similar to that of DIS 10589, and the USA suggests that it be incorporated into the IDRP text.

> *Since the Adj-RIBs are used both to receive inbound UPDATE PDUs and to provide input to the Decision Process, care must be taken that their contents are not modified while the Decision Process is running. That is, the input to the Decision Process shall not be changed while the computation is in progress.*
>
> *There are two approaches that could be taken:*
>
> 1. *The Decision Process can signal when it is running. During this time, any incoming UPDATE PDUs will be queued and will not be written into the Adj-RIBs. If more UPDATE PDUs arrive than can be fit into the allotted queue, they will be dropped and will not be acknowledged.*

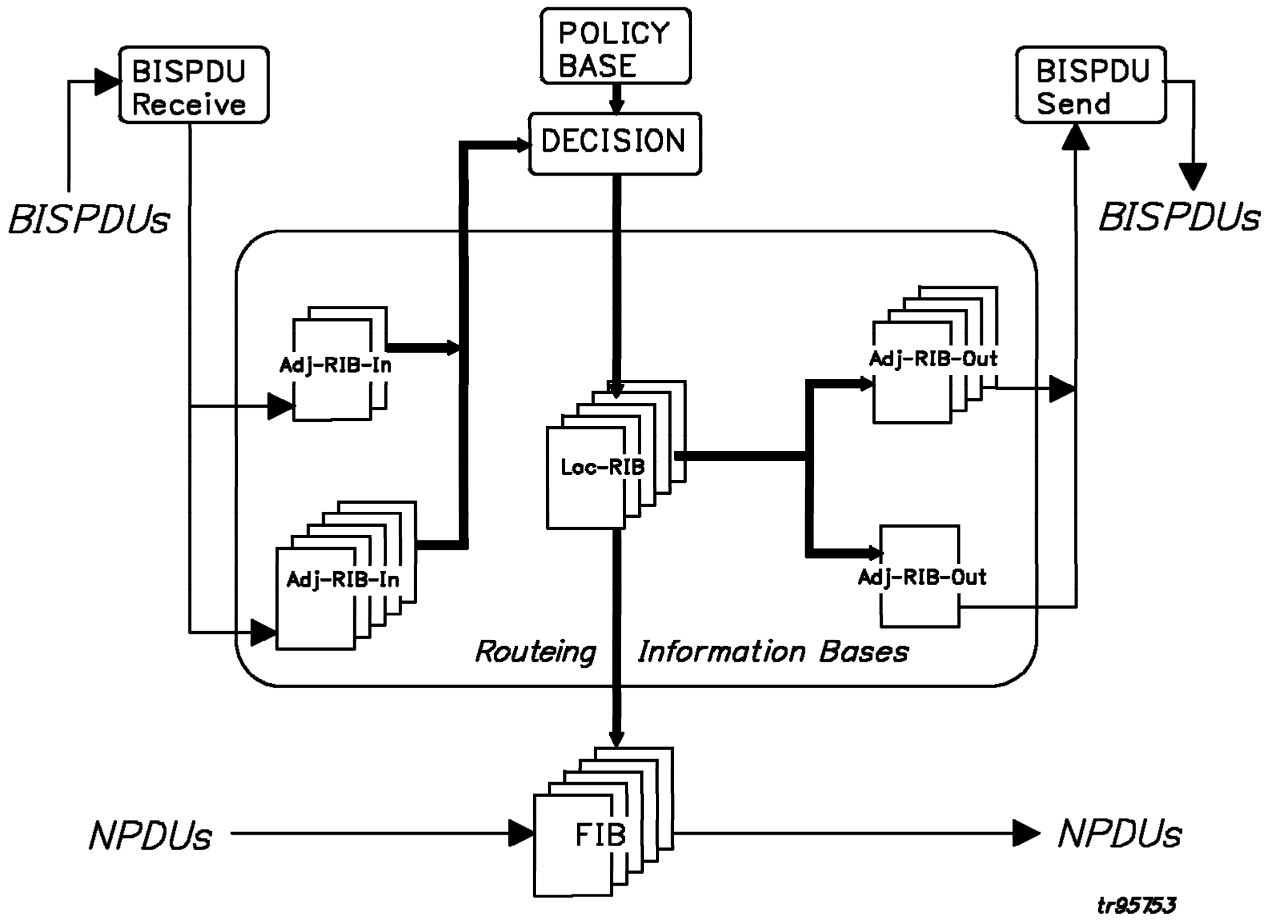

Figure 3. Replacement for SC6 N6387 Figure 4

2. *A BIS can maintain two copies of the Adj-RIBs—one used by the Decision Process for its computation (call this the Comp-Adj-RIB) and the other to receive inbound UPDATE PDUs (call this the Inbound-Adj-RIB). Each time the Decision begins a new computation, the contents of the Inbound-Adj-RIB will be copied to the Comp-Adj-RIB: that is, the a snapshot of the Comp-Adj-RIB is used as the input for the Decision Process. The contents of the Comp-Adj-RIB remain stable until a new computation is begun.*

The advantage of the first approach is that it takes less memory; the advantage of the second is that inbound UPDATE PDUs will not be dropped. This international standard does not require that either of these methods be used. Any method that guaranteees that the input data to the Decision Process will remain stable while a computation is in progress and that is consistent with the conformance requirements of this international standard may be used.

10. **Conceptual Model for RIB(s)**

----ACCEPTED----

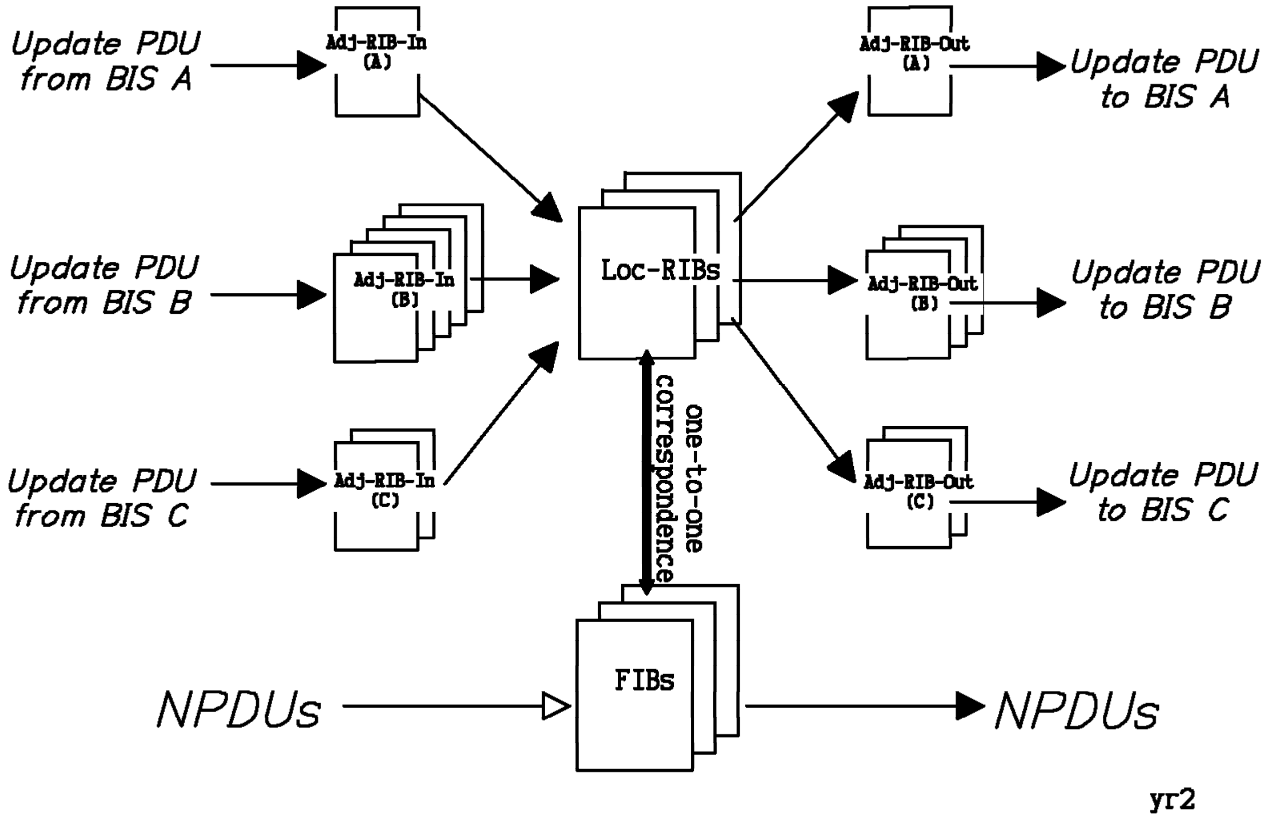

Figure 4. Replacement for SC6 N6387 Figure 8

SC6 N6387 describes a routeing information base as being composed of two conceptual parts: the Loc-RIB(s) and the Adj-RIB(s). In this model, the Adj-RIB(s) are the repository for routeing information received from other BISs.

Additional clarity can be obtained by revising the conceptual model into 3 parts:

- Loc-RIB(s): This part of the model remains unchanged from that of SC6 N6387.
- Adj-RIB(s)-In: This part of the model identifies the storage used to hold incoming routeing information received from other BISs. (It corresponds to "Adj-RIB(s)" of SC6 N6387.)
- Adj-RIB(s)-Out: This is the new part of the model. It identifies the portion of storage used to hold routes which will be advertised to other BISs. Note from the picture that it is legitimate, for example, to have only one Adj-RIB-Out for a given neighbor even if there are several Loc-RIBs. This could occur, for example, if the receiving BIS only supported the Default Attribute while the advertising BIS supported the Default Attribute plus others.

Figures 4 and 8 of SC6 N6387 would be replaced with the Figure 3 on page 5 and Figure 4

Since this change is informative in nature and does not change any normative aspects of SC6 N6387, the editor should be free to make changes throughout SC6 N6387 in order to achieve consistent

11. **Supported RIB_Atts**:

----NOT SURE IF ACCEPTED OR NOT, RECOMMEND ACCEPTANCE----

The coding of the OPEN PDU's *RIB-ATTsSET* field depicts "all *RIB_Atts* that the local BIS is willing to support when communicating with the remote BIS." The value of 0 (for number of RIB-Atts) is used to denote that the local BIS will accept (from the remote BIS) only routes that are associated with the Default Attribute. It is suggested that the value X′FF″ be used to denote that the local BIS will accept routes (from the remote BIS) based on any distinguishing attribute. application of the new model throughout the body of SC6 N6387.

12. **Computing checksum over a single RIB**

----????----

SC6 N6387 a BIS computes a single RIB checksum for inclusion in a CHECKSUM PDU. This single checksum is computed over the portions of all of its local RIBs (Loc-RIBs) that have been advertised to a given adjacent BIS. SC6 N6387 assumes that the receiving BIS maintains multiple Adj-RIBs, one for each of the Loc-RIBs which were covered by the CHECKSUM PDU. Thus, the remote BIS (the one receiving the CHECKSUM PDU) must maintain at least the same number of multiple Adj-RIBs in order to insure the Database integrity.

However, if the other BIS does not want to maintain multiple Loc-RIBs (e.g. due to memory constraints), there is no need for it to maintain multiple Adj-RIBs as well. With the current scheme this is not possible without sacrificing the Database integrity checking.

Therefore, we propose that individual checksums will be separately computed over the portion of each Loc-RIB that has been advertised to a neighbor BIS, and that each one will be be passed to that BIS separately (either in a single CHECKSUM PDU, or in multiple CHECKSUM PDUs).

To allow the receiving BIS to identify the Adj-RIB to which the CHECKSUM PDU is related, the format of the CHECKSUM PDU should be modified to associate each checksum value with the RIB-Att of the Loc-RIB (expressed as a set of distinguishing attributes) over which the checksum was computed.

The proposed format allows several <RIB-Att, checksum> pairs to be included in a single CHECKSUM PDU; each RIB-Att field is encoded exactly the same as in an OPEN PDU.

Fixed Header
First RIB-Att
First Checksum (2 octets)
Second RIB-Att
Second Checksum (2 octets)
....
Last RIB-Att
Last Checksum (2 octets)

13. **Controlling Inter-Domain Routeing traffic overhead**

To ensure the overall stability of the Inter-Domain routeing in the global OSIE, and to avoid potentially excessive overhead associated with the Inter-Domain routeing traffic, IDRP should take certain measures that limit the amount of routeing traffic (that is, BISPDUs) generated by BISs participating in the IDRP. These measures can constrain both the link bandwidth that may potentially be consumed by the Inter-Domain routeing traffic, and the BIS′s processing time for handling the Inter-Domain routeing traffic. Several measures are discussed below:

- Frequency of Route Selection

  ----ACCEPTED, add text that Bad News Travels Fast----

  (Text below has been edited slightly for clarity.)

  To limit the amount of Inter-Domain routeing traffic generated by an individual BIS as a result of receiving new routeing information, each BIS is allowed to select a better route (as determined by the local BIS) no more often than once per period, where the period is defined by the predefined architectural constant **MinRouteSelectionInterval**. Since a BIS may generate an UPDATE PDU only as a result of route selection, limiting the frequency of route selection also limits the frequency of UPDATE PDUs generated by the BIS.

  To ensure fast convergence of IDRP within an RD and to preserve consistency among BISs within an RD, the rule above should not be applied if a better route is received from another BIS that belongs to the same RD: the local BIS must advertise such routes without any further delay. It should be applied only when a better route (as determined by the local BIS) is received from a BIS in adjacent RD.

  If a BIS has selected new routes that were received from BISs in adjacent RDs, and has not yet advertised the new routes because the **MinRouteSelectionInterval** has not yet expired, the reception of any routes from other BISs in the same RD forces **MinRouteSelectionInterval** timer to expire, and triggers a new selection process that includes both routes received from other BISs in the same RD and from other BISs in adjacent RDs.

  The constraints on frequency of route generation do not apply to the advertisement of previously selected routes which have become unreachable. Such routes should be advertised as soon as possible.

- Frequency of Route Origination

  ----ACCEPTED----

  To limit the amount of routeing traffic generated by a BIS as a result of changes to the information about the systems located within that BIS′s routeing domain, there should be a minimum amount of time that must elapse before a change to that information can be advertised. It is suggested that this can be accomplished by defining a new architectural constant **MinRDOriginationInterval** whose value specifies the minimum time interval suggested above.

- Introduction of Jitter

  ----ACCEPTED----

  When BISPDUs are transmitted by a BIS as a result of either receiving UPDATE PDUs or changes in the routeing information internal to the RD that the BIS belongs to, there is a danger that the BIS generated traffic distribution will contain peaks. Where there are a large number of BISs, this can cause overloading of both the transmission medium and the BISs.

  To prevent this from occurring, we propose that ″jitter″ (as defined in DIS 10589) be imposed upon **MinRouteSelectionInterval**, **MinRDOriginationInterval**, and the timer for generation of KEEPALIVE PDUs. Note that a given BIS will apply the same ″jitter″ to each of these quantities regardless of the destinations to which the updates are being sent: that is, jitter is not applied on a ″per peer″ basis.

- Order of PDU Processing

  ----OPEN, MORE DETAIL IS NEEDED----

To ensure correct IDRP operation, processing of the UPDATE PDU within a BIS must take priority over processing of the OPEN PDU. That is, processing of information about already existing BIS-BIS connections should take precedence over processing of new connections.

If a BIS also acts as an intra-domain IS, in order to ensure overall stability, processing of the intra-domain routeing protocol PDUs must take priority over processing of any of the BISPDUs.

14. **Handling BIS Overload**

----ACCEPTED, EMPHASIZE CPU OVERLOAD----

The task group agreed that it would be useful for IDRP to specify methods for reacting to both CPU overloads and memory overloads. The task group suggests that the discussion should be amended slightly to emphasize the fact that in a distance-vector protocol such as IDRP, CPU overloads are potentially more harmful than memory overloads. That is, in a CPU-overload situation, the UPDATE process should be halted and priority should be given to the Decision Process.

Since the remedies suggested for memory overload could be construed as local policies, it was suggested that such material should be placed in an informative annex rather than in the body of the text itself.

I have not had a chance to rework the text below. Volunteers will be welcome.

Due to misconfiguration or certain transitory conditions, it is possible that there may be insufficient resources available at a particular BIS to correctly handle the IDRP. In this situation we say that the BIS becomes overloaded.

- We say that a BIS becomes memory overloaded when there is not enough memory to store both the Adj-RIBs (that are used to store the routeing information as received from other BISs), and the Loc-RIBs (that are derived from the Adj-BISs).

  Since the Loc-RIBs form a subset of the Adj-RIBs, the amount of memory needed to store the Adj-RIBs is greater than or equal to the amount of memory needed to store the Loc-RIBs. Therefore, the first step to alleviate the memory overload condition would be to reduce the amount of information that is stored in Adj-RIBs. That can be accomplished by removing routes that are not in the Loc-RIBs.

  Clearly, all routes in Adj-RIBs to destinations that are not in the Loc-RIB may be removed without no negative impact. Even if Adj-RIBs have routes to destinations that are in the Loc-RIBs as well, it still may be possible to remove some of these routes from Adj-RIBs. Since these routes may potentially be used as a fallback routes (if the current route that is in the Loc-RIB becomes unfeasible), removing them from the Adj-RIB′s may cause at some point in the future suboptimal connectivity.

  If several Adj-RIB′s (that have the same RIB attribute) have routes to the same destination, then routes with higher degree of preference (as computed by the local BIS) should be retained, while routes with lower degree of preference may be deleted, thus reducing the amount of memory needed to store the Adj-RIBs. To ensure routeing consistency within an RD, the above procedure may be applied only to the Adj-RIBs associated with BISs in adjacent RDs.

  A more drastic measure would be to terminate one or more of the IDRP sessions with other BISs. That would result in releasing the memory that was previously used to store the Adj-RIB associated with that BIS. To ensure routeing consistency within an RD this measure may be applied only to the IDRP sessions with BISs in adjacent RDs. If the above measures do not alleviate the memory overload condition, the local BIS terminates all of its IDRP sessions.

- We say that a BIS becomes CPU overloaded when there is not enough CPU processing power to process incoming BISPDUs received from other BISs. In this situation BIS must continue to update the Adj-RIBs with information contained in BISPDUs received from other BISs but may not run the Decision Process over this information except when the route received in an UPDATE_PDU has the UNREACHABLE path attribute.

  If a route received in the UPDATE_PDU has the UNREACHABLE path attribute, the local BIS checks whether this route is currently installed in one of the Loc-RIBs; if so, it removes it from the appropriate Loc-RIB, updates the appropriate FIB, and generate (if necessary) an UPDATE-PDU to inform other BIS's of the change in the Loc-RIB. The Decision Process on the local BIS does not search Adj_RIBs for another route that may be installed to replace the one that becomes unfeasible.

  Since this procedure may decrease the size of the Loc-RIB, persistence of the CPU overload condition may result in the BIS that has no routes in the Loc-RIB, thus making itself unavailable as an intermediate system. If the CPU overload condition disappears, then the Decision Process and Update Process should run over all the new routes that were installed into the Adj-RIBs but had not been processed by the Decision Process. If the CPU overload condition persists for more than the predefined architectural constant **MaxCPUOverloadPeriod time interval**, the local BIS terminates its IDRP sessions.

  The order of termination of the IDRP sessions is significant. First the BIS may terminate one or more of the IDRP sessions with BISs in adjacent RDs. If after terminating IDRP sessions with all of the BISs in adjacent RDs the CPU overload still persists, the BIS terminates the rest of its IDRP sessions (with all the BISs within its own RD).

15. **Solicited and unsolicited refresh of Adj-RIBs:**

----OPEN, MY NOTES SHOW NO CLOSURE ON THIS TOPIC----

(*I edited the text of 90-260 somewhat to try to improve its clarity.*)

In certain situations (e.g. memory overload) a BIS may have to purge some of the routeing information stored in its Adj-RIBs. If such purges occur, the Database integrity scheme (CHECKSUM PDU) will not work correctly. Therefore, we propose to add the solicited and unsolicited Adj-RIB refresh capability to the IDRP. Addition of such capability requires a new type of BISPDU, called the RIB-REFRESH PDU, with the following format:

Fixed Header
Op Code (1 octet)
Variable part

Currently defined OpCode are:
1 - RIB-Refresh-Request
2 - RIB-Refresh-Start
3 - RIB-Refresh-End

- **Solicited Refresh**

  A BIS may request the refresh of one or more of its Adj-RIBs by sending a RIB-REFRESH PDU that contains the OpCode for RIB-Refresh-Request, and it can restrict the scope of refresh by specifying the RIB-Att of the Adj-RIBs that it wants to refresh.

When a BIS receives a RIB-REFRESH PDU with OpCode RIB-Refresh-Request, it sends back RIB-REFRESH PDU with OpCode RIB-Refresh-Start, followed by a sequence of UPDATE PDUs that contain that portion of its Loc-RIBs that have been advertised to the requesting BIS. The completion of the refresh procedure is indicated by sending the RIB-REFRESH PDU with OpCode RIB-Refresh-End.

- **Unsolicited Refresh**

  A BIS may initiate an unsolicited refresh by sending a RIB-REFRESH PDU with OpCode RIB-Request-Start, followed by a sequence of UPDATE PDUs that contain that portion of its Loc-RIBs that have been advertised to a given BIS. The completion of the refresh is indicated by sending the RIB-REFRESH PDU with OpCode RIB-Refresh-End.

If the refreshing BIS receives a RIB-Refresh-Request while it is in the middle of refresh (after sending RIB-REFRESH PDU with OpCode RIB-Refresh-Start, but before sending RIB-REFRESH PDU with OpCode RIB-Refresh-End), then the current refresh is aborted and the new refresh is initiated.

If the BIS being refreshed receives a RIB-Refresh-Start in the middle of refresh (after receiving a RIB-REFRESH PDU with OpCode RIB-Refresh-Start, but before receiving the RIB-REFRESH PDU with OpCode RIB-Refresh-End), then the current refresh is aborted and the new refresh is initiated. That is, a refresh cycle may be terminated either by receipt of RIB-Refresh-End or by receipt of a new RIB-Refresh-Start.

If the OpCode is RIB-Refresh-Request, then the variable part of the RIB-REFRESH PDU contains the RIB-Atts of the Adj-RIBs for which a refresh is being requested. For all other OpCodes the Variable Part is empty.

16. **CHecksum Details**

----NEW ITEM!!!----

SC6 N6387 does not provide clear procedural rules for computing the checksums. The following rules are suggested, and should be included in the normative sections of SC6 N6387.

Note: These rules are written in terms of the new conceptual model described in comment 10 above. If this model is not adopted, it would be necessary for the editor to recast them in terms of IDRP's current model.

When a BIS computes a checksum over an individual Adj-RIB-In or Adj-RIB-Out, the following rules shall be observed:

1. *A sequence number shall be associated with each route in an Adj-RIB-In or an ADj-RIB-Out. This number shall be the sequence number of the BISPDU used to transmit (for Adj-RIB-Out) or receive (for Adj-RIB-In) the UPDATE PDU that contains the given route.*
2. *Within a single Adj-RIB-In or Adj-RIB-Out, routes shall be sorted in a non-decreasing order of their sequence numbers.*
3. *Within each route, path attributes shall be sorted in a non-decreasing order based on their type codes.*
4. *Within each route, Network Layer Reachability Information shall be sorted in a non-decreasing order.*
5. *A checksum shall be computed according to the ISO 8473 algorithm. This algorithm shall be applied to the data as sorted by the previous rules, and the sorted data shall be treated as a sequence of octets.*