Network Working Group M. Shand
Internet Draft Cisco Systems
Expiration Date: November 2002

May 2002

Restart signaling for ISIS
draft-ietf-isis-restart-01.txt

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026 [1].

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts. Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

1. Abstract

The IS-IS routing protocol (RFC 1142 [2], ISO/IEC 10589 [3]) is a link state intra-domain routing protocol. Normally, when an IS-IS router is re-started, the neighboring routers detect the restart event and cycle their adjacencies with the restarting router through the down state. This is necessary in order to invoke the protocol mechanisms to ensure correct re-synchronization of the LSP database. However, the cycling of the adjacency state causes the neighbors to regenerate their LSPs describing the adjacency concerned. This in turn causes temporary disruption of routes passing through the restarting router.

In certain scenarios such temporary disruption of the routes is highly undesirable.

This draft describes a mechanism for a restarting router to signal that it is restarting to its neighbors, and allow them to re-establish their adjacencies without cycling through the down state, while still correctly initiating database synchronization.

When such a router is restarted, it is highly desirable that it does not re-compute its own routes until it has achieved database synchronization with its neighbors. Re-computing its routes before synchronization is achieved will result in its own routes being temporarily incorrect.

This draft additionally describes a mechanism for a restarting router to determine when it has achieved synchronization with its neighbors.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [4].

3. Overview

There are two related problems with the existing specification of IS-IS with regard to re-synchronization of LSP databases when a router is re-started.

Firstly, when a routing process restarts, and an adjacency to a neighboring router is re-initialized the neighboring routing process does three things

1. It re-initializes the adjacency and causes its own LSP(s) to be regenerated, thus triggering SPF runs throughout the area (or in the case of Level 2, throughout the domain).

2. It sets SRMflags on its own LSP database on the adjacency concerned.

3. In the case of a Point-to-Point link it transmits a (set of) CSNP(s) over the adjacency.

In the case of a restarting router process, the first of these is highly undesirable, but the second is essential in order to ensure re-synchronization of the LSP database.

Secondly, whether or not the router is being re-started, it is desirable to be able to determine when the LSP databases of the neighboring routers have been synchronized (so that the overload bit can be cleared in the router's own LSP, for example). This document

describes modifications to achieve this.

It is assumed that the three-way handshake [5] is being used on Point-to-Point circuits.

4. Approach

4.1 Timers

A router that is restart capable maintains three additional timers, T1, T2 and T3.

An instance of T1 is maintained per interface, and indicates the time after which an unacknowledged restart attempt will be repeated. A typical value might be 3 seconds.

An instance of T2 is maintained for each LSP database present in the system. I.e. for a level1/2 system, there will be an instance of T2 for Level 1 and one for level 2. This is the maximum time that the system will wait for LSPDB synchronization. A typical value might be 60 seconds.

A single instance of T3 is maintained for the entire system. It indicates the time after which the router will declare that it has failed to achieve database synchronization (by setting the overload bit in its own LSP). This is initialized to 65535 seconds, but is set to the minimum of the remaining times of received IIHs containing a restart TLV with RA set.

4.2 Adjacency re-acquisition

Adjacency re-acquisition is the first step in re-initialization. The restarting router explicitly notifies its neighbor that the adjacency is being re-acquired, and hence that it should not re-initialize the adjacency. This is achieved by the inclusion of a new "re-start" option (TLV) in the IIH PDU. The presence of this TLV indicates that the sender supports the new restart capability and it carries flags that are used to convey information during a restart. All IIHs transmitted by a router that supports this capability MUST include this TLV.

```
  Type   211
  Length 3
  Value (3 octets)
     Flags (1 octet)
            Bit 1 - Restart Request (RR)
            Bit 2 - Restart Acknowledgment (RA)
            Bits 3-8 û Reserved
    Remaining Time (2 octets)
```

Remaining holding time (in seconds)
(note: only required when RA bit is set)

On receipt of an IIH with the "re-start" TLV having the RR bit set, if there exists on this interface an adjacency in state "Up" with the same System ID, and in the case of a LAN circuit, with the same source LAN address, then, irrespective of the other contents of the "Intermediate System Neighbors" option (LAN circuits), or the "Point-to-Point Adjacency State" option (Point-to-Point circuits): -

a) DO NOT refresh the timer on the adjacency, but leave the adjacency in state "Up",

b) immediately (i.e. without waiting for any currently running timer interval to expire, but with a small random delay of a few 10s of milliseconds on LANs to avoid "storms"), transmit over the corresponding interface an IIH including the "re-start" TLV with the RR bit clear and the RA bit set, having updated the "Point-to-Point Adjacency State" option to reflect any new values received from the re-starting router. (This allows the restarting router to quickly acquire the correct information to place in its hellos.) The "Remaining Time" MUST be set to the current time (in seconds) before the holding timer on this adjacency is due to expire. This IIH SHOULD be transmitted before any LSPs or SNPs transmitted as a result of the receipt of the original IIH.

c) if the corresponding interface is a Point-to-Point interface, or if the receiving router has the highest LnRouterPriority (with highest source MAC address breaking ties) among those routers whose IIHs contain the restart TLV, excluding the transmitting router (note the actual DR is NOT changed by this process.), initiate the transmission over the corresponding interface of a complete set of CSNPs, and set SRMflags on the corresponding interface for all LSPs in the local LSP database.

Otherwise (i.e. if there was no adjacency in the "UP" state to the system ID in question), process the IIH as normal by re-initializing the adjacency, and setting the RA bit in the returned IIH.

A router that does not support the re-start capability will ignore the "re-start" TLV and re-initialize the adjacency as normal, returning an IIH without the "re-start" TLV.

On starting, a router initializes the timer T3, starts timer T2 for each LSPDB and for each interface (and in the case of a LAN circuit, for each level) starts a timer T1 and transmits an IIH containing the "re-start" TLV with the RR bit set.

On a Point-to-Point circuit the "Point-to-Point Adjacency State" SHOULD be set to "Init", because the receipt of the acknowledging IIH (with RA set) MUST cause the adjacency to enter "Up" state

immediately.

Transmission of "normal" IIHs is inhibited until the conditions described below are met (in order to avoid causing an unnecessary

adjacency re-initialization). On expiry of the timer T1, it is restarted and the IIH is re-transmitted as above.

On receipt of an IIH by the restarting router, a local adjacency is established as usual, and if the IIH contains a "re-start" TLV with the RA bit set, the receipt of the acknowledgement over that interface is noted.

T3 is set to the minimum of its current value and the value of the "Remaining Time" field in the received IIH.

Receipt of an IIH not containing the "re-start" option is also treated as an acknowledgement, since it indicates that the neighbor is not re-start capable. In this case the neighbor will have re-initialized the adjacency as normal, which in the case of a Point-to-Point link will guarantee that SRMflags have been set on its database, thus ensuring eventual LSPDB synchronization. In the case of a LAN interface, the usual operation of the update process will also ensure that synchronization is eventually achieved. However, since no CSNP is guaranteed to be received over this interface, T1 is cancelled immediately without waiting for a CSNP. Synchronization may therefore be deemed complete even though there are some LSPs which are held (only) by this neighbor (see section 4.3).

In the case of a Point-to-Point circuit, the "LocalCircuitID" and "Extended Local Circuit ID" information contained in the IIH can be used immediately to generate an IIH containing the correct 3-way handshake information. The presence of "Neighbor System ID" or "Neighbor Extended Local Circuit ID" information which does not match the values currently in use by the local system is ignored (since the IIH may have been transmitted before the neighbor had received the new values from the re-starting router), but the adjacency remains in the initializing state until the correct information is received.

In the case of a LAN circuit the information in the Intermediate Systems Neighbors option is recorded and used for the generation of subsequent IIHs as normal.

When BOTH a complete set of CSNP(s) (for each active level, in the case of a pt-pt circuit) and an acknowledgement have been received over the interface, the timer T1 is cancelled.

Once T3 has expired or been cancelled, subsequent IIHs are transmitted according to the normal algorithms, but including the "re-start" TLV with both RR and RA clear.

If a LAN contains a mixture of systems, only some of which support the new algorithm, database synchronization is still guaranteed, but

the "old" systems will have re-initialized their adjacencies.

If an interface is active, but does not have any neighboring router reachable over that interface the timer T1 would never be cancelled, and according to clause 4.3.1.2 the SPF would never be run. Therefore timer T1 is cancelled after some pre-determined number of expirations (which MAY be 1). (By this time any existing adjacency on a remote system would probably have expired anyway.)

A router which supports re-start SHOULD ensure that the holding time of any IIHs it transmits is greater than the expected time to complete a re-start. However, where this is impracticable or undesirable a router MAY transmit one or more normal IIHs (containing a restart option, but with RR and RA clear) after the initial RR/RA exchange, but before synchronization has been achieved, in order to extend the holding time of the neighbors adjacencies, beyond that indicated in the remaining time field of the neighbors IIH with the RA bit set.

4.2.1 Multiple levels

A router which is operating as both a level 1 and a level 2 router on a particular interface MUST perform the above operations for each level.

On a LAN interface, it MUST send and receive both Level 1 and Level 2 IIHs and perform the CSNP synchronizations independently for each level.

On a pt-pt interface, only as single IIH (indicating support for both levels) is required, but it MUST perform the CSNP synchronizations independently for each level.

4.3 Database synchronization

When a router is started or re-started it can expect to receive a (set of) CSNP(s) over each interface. The arrival of the CSNP(s) is now guaranteed, since the "re-start" IIH with the RR bit set will be retransmitted until the CSNP(s) are correctly received.

The CSNPs describe the set of LSPs that are currently held by each neighbor. Synchronization will be complete when all these LSPs have been received.

On starting, a router starts the timer T3 and an instance of timer T2 for each LSPDB. In addition to normal processing of the CSNPs, the set of LSPIDs contained in the first complete set of CSNP(s) received over each interface is recorded, together with their remaining lifetime. If there are multiple interfaces on the restarting router, the recorded set of LSPIDs is the union of those received over each interface. LSPs with a remaining lifetime of zero

are NOT so recorded.

As LSPs are received (by the normal operation of the update process) over any interface, the corresponding LSPID entry is removed (it is also removed if the LSP had arrived before the CSNP containing the reference). When an LSPID has been held in the list for its indicated remaining lifetime, it is removed from the list. When the list of LSPIDs becomes empty, the timer T2 is cancelled.

At this point the local database is guaranteed to contain all the LSP(s) (either the same sequence number, or a more recent sequence number) which were present in the neighbors' databases at the time of re-starting. LSPs that arrived in a neighbor's database after the time of re-starting may, or may not, be present, but the normal operation of the update process will guarantee that they will eventually be received. At this point the local database is deemed to be "synchronized".

Since LSPs mentioned in the CSNP(s) with a zero remaining lifetime are not recorded, and those with a short remaining lifetime are deleted from the list when the lifetime expires, cancellation of the timer T2 will not be prevented by waiting for an LSP that will never arrive.

4.3.1 LSP generation and flooding and SPF computation

The operation of a router starting, as opposed to re-starting is somewhat different. These two cases are dealt with separately below.

4.3.1.1. Starting for the first time

In the case of a starting router, as soon as each adjacency is established, and before any CSNP exchanges, the router's own zeroth LSP is transmitted with the overload bit set. This prevents other routers from computing routes through the router until it has reliably acquired the complete set of LSPs. The overload bit remains set in subsequent transmissions of the zeroth LSP (such as will occur if a previous copy of the routers LSP is still present in the network) while any timer T2 is running.

When all the T2 timers have been cancelled, the own LSP(s) MAY be regenerated with the overload bit clear (assuming the router isn't in fact overloaded, and there is no other reason, such as incomplete BGP convergence, to keep the overload bit set), and flooded as normal.

Other 'own' LSPs (including pseudonodes) are generated and flooded as normal, irrespective of the timer T2. The SPF is also run as normal and the RIB and FIB updated as routes become available.

4.3.1.2. Re-starting

In order to avoid causing unnecessary routing churn in other routers, it is highly desirable that the own LSPs generated by the

restarting system are the same as those previously present in the network (assuming no other changes have taken place). It is important therefore not to regenerate and flood the LSPs until all the adjacencies have been re-established and any information required for propagation into the local LSPs is fully available. Ideally, the information should be loaded into the LSPs in a deterministic way, such that the same information occurs in the same place in the same LSP (and hence the LSPs are identical to their previous versions). If this can be achieved, the new versions will not even cause SPF to be run in other systems. However, provided the same information is included in the set of LSPs (albeit in a different order, and possibly different LSPs), the result of running the SPF will be the same and will not cause churn to the forwarding tables.

In the case of a re-starting router, none of the router's own non-pseudonode LSPs are transmitted, nor is the SPF run to update the forwarding tables while the timer T3 is running.

Redistribution of inter-level information must be regenerated before this router's LSP is flooded to other nodes. Therefore the level-n non-pseudonode LSP(s) should not be flooded until the other level's T2 timer has expired and its SPF has been run. This ensures that any inter-level information that should be propagated can be included in the level-n LSP(s).

During this period, if one of the router's own (including pseudonodes) LSPs is received, which the local router does not currently have in its own database, it is NOT purged. Under normal operation, such an LSP would be purged, since the LSP clearly should not be present in the global LSP database. However, in the present circumstances, this would be highly undesirable, because it could cause premature removal of an own LSP -- and hence churn in remote routers. Even if the local system has one or more own LSPs (which it has generated, but not yet transmitted) it is still not valid to compare the received LSP against this set, since it may be that as a result of propagation between level 1 and level 2 (or vice versa) a further own LSP will need to be generated when the LSP databases have synchronized.

When the timer T2 expires, or is cancelled, the SPF is run to update the RIB and FIB.

Once the other level's SPF has run and any inter-level propagation has been resolved, the 'own' LSPs can be generated and flooded. Any 'own' LSPs which were previously ignored, but which are not part of the current set of 'own' LSPs (including pseudonodes) should then be purged. Note that it is possible that a Designated Router change may

have taken place, and consequently the router should purge those pseudonode LSPs which it previously owned, but which are now no longer part of its set of pseudonode LSPs.

If the timer T3 expires before all the T2 timers have expired, this indicates that the synchronization process is taking longer than minimum holding time of the neighbors. The router's own LSP(s) for levels which have not yet completed their first SPF computation are then flooded with the overload bit set to indicate that the router's LSPDB is not yet synchronized (and other routers should therefore not compute routes through this router). In order to prevent the neighbor's adjacencies from expiring, IIHs with the normal interface value for the holding time are transmitted over all interfaces with neither RR nor RA set in the restart TLV. This will cause the neighbors to refresh their adjacencies. The own LSP(s) will continue to have the overload bit set until timer T2 has been cancelled as in the case of starting for the first time described in section 4.3.1.1

5. Security Considerations

This memo does not create any new security issues for the IS-IS protocol. Security considerations for the base IS-IS protocol are covered in [2] and [3].

6. References

1 Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

2 Callon, R., "OSI IS-IS for IP and Dual Environment," RFC 1195, December 1990.

3 ISO, "Intermediate system to Intermediate system routeing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)," ISO/IEC 10589:1992.

4 Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997

5 Katz, D., "Three-Way Handshake for IS-IS Point-to-Point Adjacencies", draft-ietf-isis-3way-03.txt, July 2000

7. Acknowledgments

The author would like to acknowledge contributions made by Radia Perlman, Mark Schaefer, Naiming Shen, Nischal Sheth, Russ White, and Rena Yang.

8. Author's Address

Mike Shand
Cisco Systems
4, The Square,
Stockley Park,
UXBRIDGE,
Middlesex
UB11 1BN, UK

Phone: +44 208 824 8690
Email: mshand@cisco.com