Network Working Group J. Moy (Sycamore Networks)
Internet Draft Padma Pillay-Esnault (Juniper Networks)
Expiration Date: April 2003 Acee Lindem, Editor (Redback Networks)
File name: draft-ietf-ospf-hitless-restart-04.txt October 2002

Hitless OSPF Restart
draft-ietf-ospf-hitless-restart-04.txt

Status of this Memo

Abstract

This memo documents an enhancement to the OSPF routing protocol, whereby an OSPF router can stay on the forwarding path even as its OSPF software is restarted. This is called "hitless restart" or "non-stop forwarding". A restarting router may not be capable of adjusting its forwarding in a timely manner when the network topology changes. In order to avoid the possible resulting routing loops the procedure in this memo automatically reverts to a normal OSPF restart when such a topology change is detected, or when one or more of the restarting router's neighbors do not support the enhancements in this memo. Proper network operation during a hitless restart makes assumptions upon the operating environment of the restarting router; these assumptions are also documented.

Table of Contents

1. Overview

Today many Internet routers implement a separation of control and forwarding functions. Certain processors are dedicated to control and management tasks such as OSPF routing, while other processors perform the data forwarding tasks. This separation creates the possibility of maintaining a router's data forwarding capability while the router's control software is restarted/reloaded. We call such a possibility "hitless restart" or "non-stop forwarding".

The problem that the OSPF protocol presents to hitless restart is that, under normal operation, OSPF intentionally routes around a restarting router while it rebuilds its link-state database. OSPF avoids the restarting router to minimize the possibility of routing loops and/or black holes caused by lack of database synchronization. Avoidance is accomplished by having the router's neighbors reissue their LSAs, omitting links to the restarting router.

However, if (a) the network topology remains stable and (b) the restarting router is able to keep its forwarding table(s) across the restart, it would be safe to keep the restarting router on the forwarding path. This memo documents an enhancement to OSPF that

makes such hitless restart possible, and one that automatically reverts back to a standard OSPF restart for safety when network topology changes are detected.

In a nutshell, the OSPF enhancements for hitless restart are as follows. The router attempting a hitless restart originates link-

local Opaque-LSAs, herein called Grace-LSAs, announcing the intention to perform a hitless restart, and asking for a "grace period". During the grace period its neighbors continue to announce the restarting router in their LSAs as if it were fully adjacent (i.e., OSPF neighbor state Full), but only if the network topology remains static (i.e, the contents of the LSAs in the link-state database having LS types 1-5,7 remain unchanged; periodic refreshes are allowed).

There are two roles being played by OSPF routers during hitless restart. First there is the router that is being restarted. The operation of this router during hitless restart, including how the router enters and leaves hitless restart, is the subject of Section 2. Then there are the router's neighbors, which must cooperate in order for the restart to be hitless. During hitless restart we say that the neighbors are executing in "helper mode". Section 3 covers the responsibilities of a router executing in helper mode, including entering and leaving helper mode.

1.1. Acknowledgments

The authors wish to thank John Drake, Vishwas Manral, Kent Wong and Don Goodspeed for their helpful comments.

2. Operation of restarting router

After the router restarts/reloads, it must change its OSPF processing somewhat until it re-establishes full adjacencies with all its previously fully-adjacent neighbors. This time period, between the restart/reload and the reestablishment of adjacencies, is called "hitless restart". During hitless restart:

(1) The restarting router does not originate LSAs with LS types 1-5,7. Instead, the restarting router wants the other routers in the OSPF domain to calculate routes using the LSAs that it

had originated prior to its restart. During this time, the restarting router does not modify or flush received self-originated LSAs, (see Section 13.4 of [1]) but instead accepts them as valid. In particular, the grace-LSAs that the restarting router had originated before the restart are left in place. Received self-originated LSAs will be dealt with when the router exits hitless restart (see Section 2.3).

(2) The restarting router runs its OSPF routing calculations, as specified in Section 16 of [1]. This is necessary to return any OSPF virtual links to operation. However, the restarting router does *not* install OSPF routes into the system's forwarding table(s), instead relying on the forwarding entries that it had installed prior to the restart.

(3) If the restarting router determines that it was Designated Router on a given segment immediately prior to the restart, it elects itself as Designated Router again. The restarting router knows that it was Designated Router if, while the associated interface is in Waiting state, an Hello packet is received from a neighbor listing the router as Designated Router.

Otherwise, the restarting router operates the same as any other OSPF router. It discovers neighbors using OSPF's Hello protocol, elects Designated and Backup Designated Routers, performs the Database Exchange procedure to initially synchronize link-state databases with its neighbors, and maintains this synchronization through flooding.

The processes of entering hitless restart, and of exiting hitless restart (either successfully or not) are covered in the following sections.

2.1. Entering hitless restart

The router (call it Router X) is informed of the desire for its hitless restart when an appropriate command is issued by the network operator. The network operator may also specify the length of the grace period, or the necessary grace period may be calculated by the router's OSPF software. In order to avoid the restarting router's LSAs from aging out, the grace period

should not exceed LSRefreshTime (1800 second) [1].

In preparation for the hitless restart, Router X must perform the following actions before its software is restarted/reloaded. Note that common OSPF shutdown procedures are *not* performed, since we want the other OSPF routers to act as if Router X remains in continuous service. For example, Router X does not flush its locally originated LSAs, since we want them to remain in other routers' link-state databases throughout the restart period.

(1) Router X must ensure that its forwarding table(s) is/are up-to-date and will remain in place across the restart.

(2) The router may need to preserve the cryptographic sequence numbers being used on each interface in non-volatile storage. An alternative is to use the router's clock for cryptographic sequence number generation and ensure the clock is preserved across restarts (either on the same or redundant route processors). If neither of these can be guarenteed, it can take up to RouterDeadInterval seconds after the restart before adjacencies can be reestablished and this would force the grace period to be lengthened greatly.

Router X then originates the grace-LSAs. These are link-local Opaque-LSAs (see Appendix A). Their LS Age field is set to 0, and the requested grace period (in seconds) is inserted into the body of the grace-LSA. The precise contents of the grace-LSA are described in Appendix A.

A grace-LSA is originated for each of the router's OSPF interfaces. If Router X wants to ensure that its neighbors receive the grace-LSAs, it should retransmit the grace-LSAs until they are acknowledged (i.e, perform standard OSPF reliable flooding of the grace-LSAs). If one or more fully adjacent neighbors do not receive grace-LSAs, they will more than likely cause premature termination of the hitless restart procedure (see Section 4).

After the grace-LSAs have been sent, the router should store the fact that it is performing hitless restart along with the length of the requested grace period in non-volatile storage. (Note to implementors: It may be easiest to simply store the absolute

time of the end of the grace period). The OSPF software should then be restarted/reloaded, and when the reloaded software starts executing the hitless restart modifications in Section 2 above are followed. (Note that prior to the restart, the router does not know whether its neighbors are going to cooperate as "helpers"; the mere reception of grace-LSAs does not imply acceptance of helper responsibilities. This memo assumes that the router would want to restart anyway, even if the restart is not going to be hitless).

2.2. When to exit hitless restart

A Router X exits hitless restart when any of the following occurs:

(1) Router X has reestablished all its adjacencies. Router X can determine this by examining the router-LSAs that it had last originated before the restart (called the "pre-restart router-LSA"), and, on those segments where the router is Designated Router, the pre-restart network-LSAs. These LSAs will have been received from the helping neighbors, and need not have been stored in non-volatile storage across the restart. All previous adjacencies will be listed as type-1 and type 2 links in the router-LSA, and as neighbors in the body of the network-LSA.

(2) Router X receives an LSA that is inconsistent with its pre-restart router-LSA. For example, X receives a router-LSA originated by router Y that does not contain a link

to X, even though X's pre-start router-LSA did contain a link to Y. This indicates that either a) Y does not support hitless restart, b) Y never received the grace-LSA or c) Y has terminated its helper mode for some reason (Section 3.2).

(3) The grace period expires.

2.3. Actions on exiting hitless restart

On exiting "hitless restart", the reloaded router reverts back to completely normal OSPF operation, reoriginating LSAs based on

the router's current state and updating its forwarding table(s) based on the current contents of the link-state database. In particular, the following actions should be performed when exiting, either successfully or unsuccessfully, hitless restart.

(1) The router should reoriginate its router-LSAs for all attached areas, to make sure they have the correct contents.

(2) The router should reoriginate network-LSAs on all segments where it is Designated Router.

(3) The router reruns its OSPF routing calculations (Section 16 of [1]), this time installing the results into the system forwarding table, and originating summary-LSAs, Type-7 LSAs and AS-external-LSAs as necessary.

(4) Any remnant entries in the system forwarding table that were installed before the restart, but that are no longer valid, should be removed.

(5) Any received self-originated LSAs that are no longer valid should be flushed.

(6) Any grace-LSAs that the router had originated should be flushed.

3. Operation of helper neighbor

The helper relationship is per network segment. As a "helper neighbor" on a segment S for a restarting router X, router Y has several duties. It monitors the network for topology changes, and as long as there are none, continues to its advertise its LSAs as if X had remained in continuous OSPF operation. This means that Y's LSAs continue to list an adjacency to X over network segment S, regardless of the adjacency's current synchronization state. This

logic affects the contents of both router-LSAs and network-LSAs, and also depends on the type of network segment S (see Sections 12.4.1.1 through 12.4.1.5 and Section 12.4.2 of [1]). When helping over a virtual link, the helper must also continue to set bit V in its router-LSA for the virtual link's transit area (Section 12.4.1 of

[1]).

Also, if X was the Designated Router on network segment S when the helping relationship began, Y maintains X as Designated router until the helping relationship is terminated.

3.1. Entering helper mode

When a router Y receives a grace-LSA from router X, it enters helper mode for X, on the associated network segment, as long as all the following checks pass:

(1) Y currently has a full adjacency with X (neighbor state Full) over the associated network segment. On broadcast, NBMA and Point-to-MultiPoint segments, the neighbor relationship with X is identified by the IP interface address in the body of the grace-LSA (see Appendix A). On all other segment types X is identified by the grace-LSA's Advertising Router field.

(2) There have been no changes in content to the link-state database (LS types 1-5,7) since router X restarted. This is determined as follows. Router Y examines the link-state retransmission list for X over the associated network segment. If there are any LSAs with LS types 1-5,7 on the list, then they all must be periodic refreshes. If there are instead LSAs on the list whose contents have changed (see Section 3.3 of [8]), Y must refuse to enter helper mode.

(3) The grace period has not yet expired. This means that the LS age of the grace-LSA is less than the grace period specified in the body of the grace-LSA (Appendix A).

(4) Local policy allows Y to act as the helper for X. Examples of configured policies might be a) never act as helper, b) never allow the grace period to exceed a Time T, c) only help on software reloads/upgrades, or d) never act as a helper for certain specific routers (specified by OSPF Router ID).

There is one exception to the above requirements. If Y was already helping X on the associated network segment, the new

grace-LSA should be accepted and the grace period should be updated accordingly.

Note that Router Y may be helping X on some network segments, and not on others. However, that circumstance will probably lead to the premature termination of X's hitless restart, as Y will not continue to advertise adjacencies on the segments where it is not helping (see Section 2.2).

Alternately, Router Y may choose to enter enter helper mode when a grace LSA is received and the above checks pass for all adjacencies with Router X. This implemenation alternative of aggregating the adjacencies with respect to helper mode is compatible with implementations considering each adjacency independently.

A single router is allowed to simultaneously serve as a helper for multiple restarting neighbors.

3.2. Exiting helper mode

Router Y ceases to perform the helper function for its neighbor Router X on a given segment when one of the following events occurs.

(1) The grace-LSA originated by X on the segment is flushed. This is the successful termination of hitless restart.

(2) The grace-LSA's grace period expires.

(3) A change in link-state database contents indicates a network topology change, which forces termination of a hitless restart. Specifically, if router Y installs a new LSA in its database with LS types 1-5,7 and having the following two properties, it should cease helping X. The two properties of the LSA are a) the contents of the LSA have changed; this includes LSAs with no previous link-state database instance and the flushing of LSAs from the database, but excludes periodic LSA refreshes (see Section 3.3 of [8]), and b) the LSA would have been flooded to X, had Y and X been fully adjacent. As an example of the second property, if Y installs a changed AS-external-LSA, it should not terminate a helping relationship with a neighbor belonging to a stub area, as that neighbor would not see the AS-external-LSA in any case. An implementation MAY provide a configuration option to disable link-state database options from terminating hitless restart. Such an option will, however, increase the risk of routing loops and

black holes.

When Router Y exits helper mode for X on a given network segment, it reoriginates its LSAs based on the current state of its adjacency to Router X over the segment. In detail, Y takes the following actions: (a) Y recalculates the Designated Router for the segment, (b) Y reoriginates its router-LSA for the segment's OSPF area, (c) if Y is Designated Router for the segment, it reoriginates the network-LSA for the segment and (d) if the segment was a virtual link, Y reoriginates its router-LSA for the virtual link's transit area.

If Router Y aggregated adjacencies with Router X when entering helper mode (as described in section 3.1), it must also exit helper mode for all adjacencies with Router X when any one of the exit events occurs for of adjacency with Router X.

4. Backward compatibility

Backward-compatibility with unmodified OSPF routers is an automatic consequence of the functionality documented above. If one or more neighbors of a router requesting hitless restart are unmodified, or if they do not received the grace-LSA, the hitless restart converts to a normal OSPF restart.

The unmodified routers will start routing around the restarted router X as it performs initial database synchronization, by reissuing their LSAs with links to X omitted. These LSAs will be interpreted by helper neighbors as a topology change, and by X as an LSA inconsistency, in either case reverting to normal OSPF operation.

5. Unplanned outages

The hitless restart mechanisms in this memo can be used for unplanned outages. (Examples of unplanned outages include the crash of a router's control software, an unexpected switchover to a redundant control processor, etc). However, implementors and network operators should note that attempting hitless restart from an unplanned outage may not be a good idea, owing to the router's inability to properly prepare for the restart (see Section 2.1). In particular, it seems unlikely that a router could guarantee the sanity of its forwarding table(s) across an unplanned restart. In

any event, implementors providing the option to recover hitlessly from unplanned outages must allow a network operator to turn the option off.

In contrast to the procedure for planned restart/reloads that was described in Section 2.1, a router attempting hitless restart after an unplanned outage must originate grace-LSAs *after* its control software resumes operation. The following points must be observed during this grace-LSA origination.

- The grace-LSAs must be originated and sent *before* the restarted router sends any OSPF Hello Packets. On broadcast networks, this LSA must be flooded to the AllSPFRouters multicast address (224.0.0.5) since the restarting router is not aware of its previous DR state.

- The grace-LSAs are encapsulated in Link State Update Packets and sent out all interfaces, even though the restarted router has no adjacencies and no knowledge of previous adjacencies.

- To improve the probability that grace-LSAs be delivered, an implementation may send them a number of times (see for example the Robustness Variable in [8]).

- The restart reason in the grace-LSAs must be set to unknown(0). This enables the neighbors to decide whether they want to help the router through an unplanned restart.

6. Interaction with Traffic Engineering

The operation of the Traffic Engineering Extensions to OSPF [4] during OSPF Hitless Restart is specified in [6].

7. Possible Future Work

Devise a less conservative algorithm for graceful restart helper termination that provides a comparable level of black hole and routing loop avoidance.

Normative References

[1] Moy, J., "OSPF Version 2", RFC 2328, April 1998.

[2] Coltun, R., "The OSPF Opaque LSA Option", RFC 2370, July 1998.

Informative References

[3] Murphy, S., M. Badger and B. Wellington, "OSPF with Digital Signatures", RFC 2154, June 1997.

[4] Katz, D., D. Yeung and K. Kompella, "Traffic Engineering Extensions to OSPF", work in progress.

[5] Coltun, R., V. Fuller and P. Murphy, "The OSPF NSSA Option", work in progress.

[6] Kompella, K., et. al., "Routing Extensions in Support of Generalized MPLS", work in progress.

[7] Moy, J., "Extending OSPF to Support Demand Circuits", RFC 1793, April 1995.

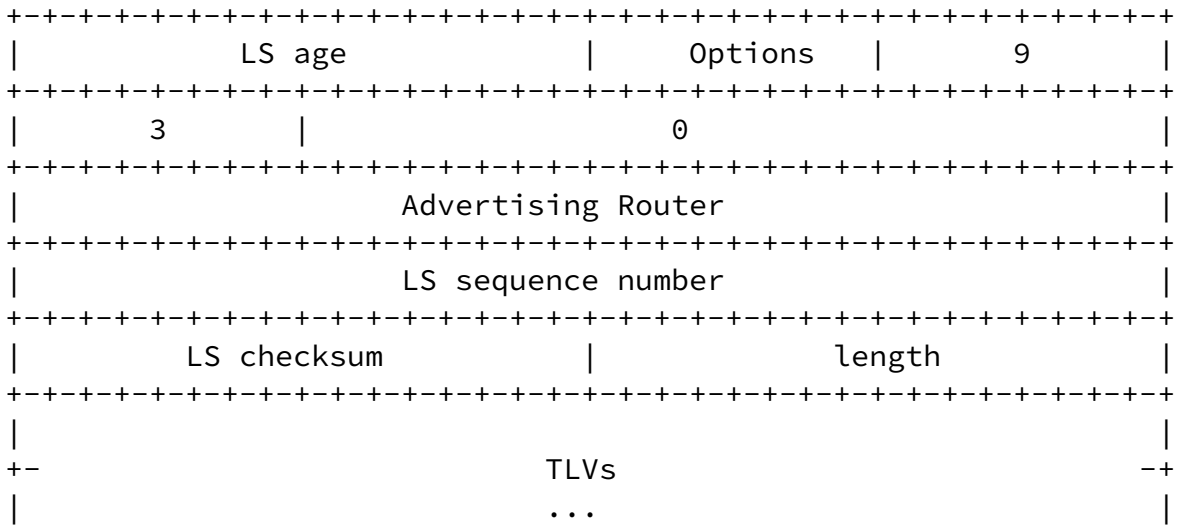
[8] Fenner, W., "Internet Group Membership Protocol, Version 2", RFC 2236, November 1997.

A. Grace-LSA format

The grace-LSA is a link-local scoped Opaque-LSA [2] having Opaque Type of 3 and Opaque ID equal to 0. Grace-LSAs are originated by a router that wishes to execute a hitless restart of its OSPF software. A grace-LSA requests that the router's neighbors aid it in its hitless restart by continuing to advertise the router as fully adjacent during a specified grace period.

Each grace-LSA has LS age field set to 0 when the LSA is first originated; the current value of LS age then indicates how long ago the restarting router made its request. The body of the LSA is TLV-encoded. The TLV-encoded information includes the length of the grace period, the reason for the hitless restart and, when the grace-LSA is associated with a broadcast, NBMA or Point-to-MultiPoint network segment, the IP interface address of the restarting router.

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
```



The format of the TLVs within the body of a grace-LSA is the same as the TLV format used by the Traffic Engineering Extensions to OSPF [4]. The TLV header consists of a 16-bit Type field and a 16-bit length field, and is followed by zero or more bytes of value. The length field indicates the length of the value portion in bytes. The value portion is padded to four-octet alignment, but the padding is not included in the length field. For example, a one byte value would have the length field set to 1, and three bytes of padding would be added to the end of the value portion of the TLV.

The following is the list of TLVs that can appear in the body of a grace-LSA.

- Grace Period (Type=1, length=4). The number of seconds that the router's neighbors should continue to advertise the router as fully adjacent, regardless of the the state of database synchronization between the router and its neighbors. Since this time period began when grace-LSA's LS age was equal to 0, the grace period terminates when either a) the LS age of the grace-LSA exceeds the value of Grace Period or b) the grace-LSA is flushed. See Section 3.2 for other conditions which terminate the grace period. This TLV must always appear in a grace-LSA.

- Hitless restart reason (Type=2, length=1). Encodes the reason for the router restart, as one of the following: 0 (unknown), 1 (software restart), 2 (software reload/upgrade) or 3 (switch to redundant control processor). This TLV must always appear in a

grace-LSA.

- o IP interface address (Type=3, length=4). The router's IP interface address on the subnet associated with the grace-LSA. Required on broadcast, NBMA and Point-to-MultiPoint segments, where the helper uses the IP interface address to identify the restarting router (see Section 3.1).

DoNotAge is never set in a grace-LSA, even if the grace-LSA is flooded over a demand circuit [7]. This is because the grace-LSA's LS age field is used to calculate the extent of the grace period.

Grace-LSAs have link-local scope because they only need to be seen by the router's direct neighbors.

B. Change Log (To be removed prior to publication)

Changes from 02 to 03 version:

1. Add Padma Pillay-Esnault and Acee Lindem as authors to help finish up the draft.

Changes from 03 to 04 version:

1. Add change log (Appendix B).
2. Document that the grace period is restricted to LSRefreshTime (Section 2.1).
3. Document an alternative to saving cryptographic sequence numbers in non-volatile storage (Section 2.1).

4. Document that an implementation may aggregate multiple adjacencies with a restarting router when entering or exiting helper mode (Section 3.1 and 3.2).
5. Document that an implementation may disable graceful restart helper termination when the link-state database changes (Section 3.2).
6. In the case of an unplanned restart, document that grace LSAs should be flooded to AllSPFRouters on broadcast networks (Section 5).
7. Remove MOSPF from future work. Add Vishwas's suggested technique for less conservative helper mode termination as possible future work (Section 7).
8. Change references and citations to meet prevailing IETF standards.

Security Considerations

One of the ways to attack a link-state protocol such as OSPF is to inject false LSAs into, or corrupt existing LSAs in, the link-state database. Injecting a false grace-LSA would allow an attacker to spoof a router that, in reality, has been withdrawn from service. The standard way to prevent such corruption of the link-state database is to secure OSPF protocol exchanges using the Cryptographic authentication specified in [1]. An even stronger way of securing link-state database contents has been proposed in [3].

Authors' Addresses

J. Moy
Sycamore Networks, Inc.
150 Apollo Drive
Chelmsford, MA 01824
Phone: (978) 367-2505
Fax: (978) 256-4203
email: jmoy@sycamorenet.com

Padma Pillay-Esnault
Juniper Networks
1194 N, Mathilda Avenue
Sunnyvale, CA 94089-1206
Email: padma@juniper.net

Acee Lindem

Redback Networks
102 Carric Bend Court
Cary, NC 27519
Email: acee@redback.com