

Network Working Group

Internet Draft

Expiration Date: February 2002

File name: [draft-ietf-ospf-hitless-restart-01.txt](#)

J. Moy

Sycamore Networks, Inc.

August 2001

## Hitless OSPF Restart

[draft-ietf-ospf-hitless-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](#).

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>.

### 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 terminates when such a topology change is detected. The restart procedure is also backward-compatible, reverting to standard OSPF processing 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

<a href="#">1</a>	Overview .....	<a href="#">2</a>
<a href="#">2</a>	Operation of restarting router .....	<a href="#">3</a>
<a href="#">2.1</a>	Entering hitless restart .....	<a href="#">4</a>
<a href="#">2.2</a>	Exiting hitless restart .....	<a href="#">5</a>
<a href="#">3</a>	Operation of helper neighbor .....	<a href="#">6</a>
<a href="#">3.1</a>	Entering helper mode .....	<a href="#">7</a>
<a href="#">3.2</a>	Exiting helper mode .....	<a href="#">7</a>
<a href="#">4</a>	Backward compatibility .....	<a href="#">8</a>
<a href="#">5</a>	Notes .....	<a href="#">8</a>
<a href="#">6</a>	Future Work .....	<a href="#">9</a>
	References .....	<a href="#">9</a>
<a href="#">A</a>	Grace-LSA format .....	<a href="#">10</a>
	Security Considerations .....	<a href="#">12</a>
	Authors' Addresses .....	<a href="#">12</a>

## [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 standard OSPF 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; simple 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.

## **[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 [\[Ref1\]](#)) 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 2.2](#)).
- (2) The restarting router runs its OSPF routing calculations, as specified in Section 16 of [\[Ref1\]](#). 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 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 must note in non-volatile storage the cryptographic sequence numbers being used for each interface. Otherwise it will take up to RouterDeadInterval seconds after the restart before it can start to reestablish its adjacencies, which would force the grace period to be lengthened severely.

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. 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.

## 2.2. 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 [[Ref1](#)]), this time installing the results into the system forwarding table, and originating summary-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.





The router exits hitless restart when any of the following occurs:

- (1) Router X has reestablished all its adjacencies. Router X can determine this by building (but not installing or flooding) its router-LSAs, based on the current router state, and comparing it to the router-LSAs that it had last originated before the restart (called the "pre-restart router-LSA"). On those segments where the router is Designated Router, network-LSAs should also be built and compared to those received (if any). If the contents of the built and received LSAs are the same, all adjacencies have been reestablished.
- (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.

### **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 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 [[Ref1](#)]). 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 [[Ref1](#)]).

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 and NBMA 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 the beginning of the grace period specified by the grace-LSA. The grace period began N seconds ago, where N is the current LS age of the grace-LSA.
- (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).

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](#)).

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) Router Y receives an LSA with LS types 1-5,7 and whose contents have changed. This includes LSAs with no previous link-state database instance and the flushing of LSAs from the database, but excludes simple LSA refreshes. A change in LSA contents indicates a network topology change, which forces termination of a hitless restart.

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.

#### **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 is prematurely aborted.

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 aborting hitless restart and resuming normal OSPF operation.

#### **5. Notes**

Note the following details concerning the hitless OSPF restart mechanism described in this memo.

- o DoNotAge is never set in a grace-LSA, even if the grace-LSA is flooded over a demand circuit. This is because the grace-LSA's LS age field is used to calculate the extent of the grace period (see [Appendix A](#)).
- o Grace-LSAs have link-local scope because they only need to be seen by the router's direct neighbors.



- o It may be noted that the hitless restart mechanisms in this memo can also be used for unplanned outages. For example, after a crash of its control software, the router may come up and send grace-LSAs in an attempt to remain on the forwarding path while it regains its control state. This may not be a good idea, as it seems unlikely that such a router could guarantee the sanity of its forwarding table(s). However, if the router does attempt a hitless restart from an unplanned outage, it should at the least (a) allow the network operator to turn this feature off, (b) attempt to determine when its forwarding tables were last updated, setting the beginning of the grace period accordingly (this means originating the grace-LSA with LS age equal to the time that the forwarding tables were last updated), and (c) set the hitless restart reason in its grace-LSAs to unknown (0).
- o When comparing the LSAs that the router would build and those that the router has received, in order to determine whether hitless restart is complete, it is sufficient to compare the LSA's length, Options field, and the body of the LSA. The LS age, LS checksum and LS sequence number fields need not match. To cover the possibility that the body of the LSA may have been built in a different order, the standard Internet one's complement checksum of the LSA bodies can be calculated and compared, rather than comparing the bodies byte-for-byte.

## 6. Future Work

The interactions between OSPF Hitless Restart and the Traffic Engineering Extensions to OSPF [[Ref4](#)] and Multicast Extensions to OSPF [[Ref6](#)] have not been specified in this memo.

## References

- [Ref1] Moy, J., "OSPF Version 2", [RFC 2328](#), April 1998.
- [Ref2] Coltun, R., "The OSPF Opaque LSA Option", [RFC 2370](#), July 1998.
- [Ref3] Murphy, S., M. Badger and B. Wellington, "OSPF with Digital Signatures", [RFC 2154](#), June 1997.
- [Ref4] Katz, D., D. Yeung and K. Kompella, "Traffic Engineering Extensions to OSPF", work in progress.
- [Ref5] Coltun, R., V. Fuller and P. Murphy, "The OSPF NSSA Option", work in progress.



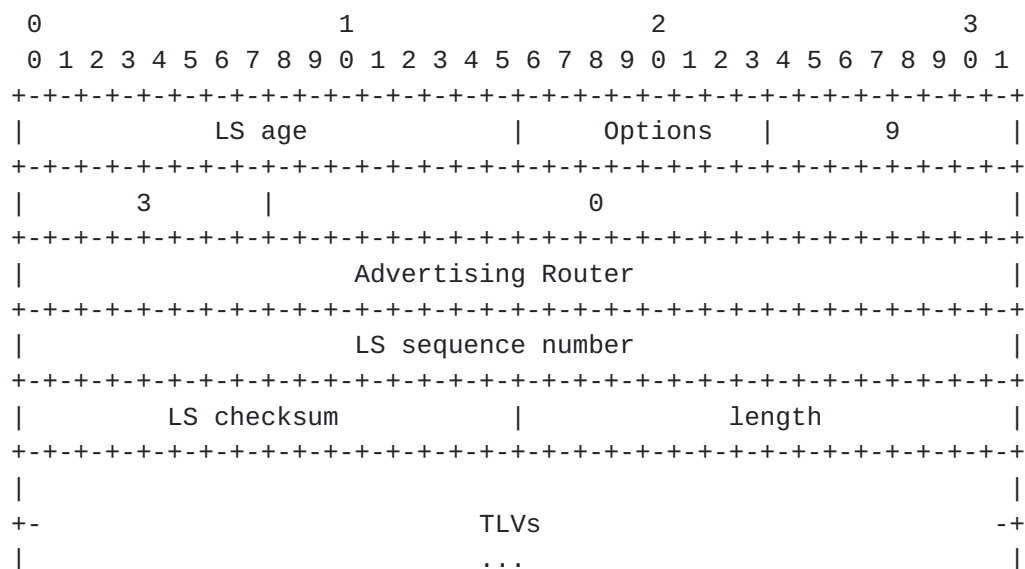


[Ref6] Moy, J., "Multicast Extensions to OSPF", [RFC 1584](#), March 1994.

## A. Grace-LSA format

The grace-LSA is a link-local scoped Opaque-LSA [[Ref2](#)] 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.

It is assumed that 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 or NBMA network segment, the IP interface address of the restarting router.



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 [[Ref4](#)]. 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.



- o 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.
- o 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 and NBMA segments, where the helper uses the IP interface address to identify the restarting router (see [Section 3.1](#)).



### 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 [[Ref1](#)]. An even stronger way of securing link-state database contents has been proposed in [[Ref3](#)].

### 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](mailto:jmoy@sycamorenet.com)

