

Network Working Group  
Internet Draft  
Intended status: Proposed Standard  
Expires: November 2011

S. Giacalone  
Thomson Reuters

D. Ward  
Juniper Networks

J. Drake  
Juniper Networks

A. Atlas  
Juniper Networks

S. Previdi  
Cisco Systems

May 30, 2011

OSPF Traffic Engineering (TE) Express Path  
draft-giacalone-ospf-te-express-path-01.txt

## Abstract

In certain networks, such as, but not limited to, financial information networks (e.g. stock market data providers), network performance criteria (e.g. latency) are becoming as critical to data path selection as other metrics.

This document describes extensions to OSPF TE [RFC3630] such that network performance information can be distributed and collected in a scalable fashion. The information distributed using OSPF TE Express Path can then be used to make path selection decisions based on network performance.

Note that this document only covers the mechanisms with which network performance information is distributed. The mechanisms for measuring network performance or acting on that information, once distributed, are outside the scope of this document.

## Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

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/lid-abstracts.txt>

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

This Internet-Draft will expire on November 31, 2011.

#### Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

#### Table of Contents

1. Introduction.....	3
2. Conventions used in this document.....	4
3. Express Path Extensions to OSPF TE.....	4
4. Sub TLV Details.....	6
4.1. Unidirectional Link Delay Sub-TLV.....	6
4.1.1. Type.....	6
4.1.2. Length.....	6

4.1.3. A bit.....	7
4.1.4. Reserved.....	7
4.1.5. Delay Value.....	7
4.2. Unidirectional Delay Variation Sub-TLV.....	7
4.2.1. Type.....	7
4.2.2. Length.....	7
4.2.3. Reserved.....	8
4.2.4. Delay Variation.....	8
4.3. Unidirectional Link Loss Sub-TLV.....	8
4.3.1. Type.....	8
4.3.2. Length.....	8
4.3.3. A bit.....	8
4.3.4. Reserved.....	9
4.3.5. Link Loss.....	9
4.4. Unidirectional Residual Bandwidth Sub-TLV.....	9
4.4.1. Type.....	9
4.4.2. Length.....	10
4.4.3. Residual Bandwidth.....	10
4.5. Unidirectional Available Bandwidth Sub-TLV.....	10
4.4.4. Type.....	10
4.4.5. Length.....	10
4.4.6. Available Bandwidth.....	10
5. Announcement Thresholds and Filters.....	11
6. Announcement Suppression.....	11
7. Network Stability and Announcement Periodicity.....	11
8. Compatibility.....	12
9. Security Considerations.....	12
10. IANA Considerations.....	12
11. References.....	12
11.1. Normative References.....	12
11.2. Informative References.....	12
12. Acknowledgments.....	13
13. Author's Addresses.....	13

## 1. Introduction

In certain networks, such as, but not limited to, financial information networks (e.g. stock market data providers), network performance information (e.g. latency) is becoming as critical to data path selection as other metrics.

In these networks, extremely large amounts of money rest on the ability to access market data in "real time" and to predictably make trades faster than the competition. Because of this, using metrics such as hop count or cost as routing metrics is becoming only

tangentially important. Rather, it would be beneficial to be able to make path selection decisions based on performance data (such as latency) in a cost-effective and scalable way.

This document describes extensions to OSPF TE (hereafter called "OSPF TE Express Path"), that can be used to distribute network performance information (such as link delay, delay variation, packet loss, residual bandwidth, and available bandwidth).

The data distributed by OSPF TE Express Path is meant to be used as part of the operation of the routing protocol (e.g. by replacing cost with latency or considering bandwidth as well as cost), by enhancing CSPF, or for other uses such as supplementing the data used by an Alto server [Alto]. With respect to CSPF, the data distributed by OSPF TE Express Path can be used to setup, fail over, and fail back data paths using protocols such as RSVP-TE [RFC3209].

Note that the mechanisms described in this document only disseminate performance information. The methods for initially gathering that performance information, such as [Frost], or acting on it once it is distributed are outside the scope of this document.

## 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 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

## 3. Express Path Extensions to OSPF TE

This document proposes new OSPF TE sub-TLVs that can be announced in OSPF TE LSAs to distribute network performance information. The extensions in this document build on the ones provided in OSPF TE [RFC3630] and GMPLS [RFC4203].

OSPF TE LSAs [RFC3630] are opaque LSAs [RFC5250] with area flooding scope. Each TLV has one or more nested sub-TLVs which permit the TE LSA to be readily extended. There are two main types of OSPF TE LSA; the Router Address or Link TE LSA. Like the extensions in GMPLS

(RFC4203), this document proposes several additional sub-TLVs for the Link TE LSA:

Type	Length	Value
TBD1	4	Unidirectional Link Delay
TBD2	4	Unidirectional Delay Variation
TBD3	4	Unidirectional Packet Loss
TBD4	4	Unidirectional Residual Bandwidth Sub TLV
TBD5	4	Unidirectional Available Bandwidth Sub TLV

As can be seen in the list above, the sub-TLVs described in this document carry different types of network performance information. Many (but not all) of the sub-TLVs include a bit called the Anomalous (or "A") bit. When the A bit is clear (or when the sub-TLV does not include an A bit), the sub-TLV describes steady state link performance. This information could conceivably be used to construct a steady state performance topology for initial tunnel path computation, or to verify alternative failover paths.

When network performance violates configurable link-local thresholds a sub-TLV with the A bit set is advertised. These sub-TLVs could be used by the receiving node to determine whether to fail traffic to a backup path, or whether to calculate an entirely new path. From an MPLS perspective, the intent of the A bit is to permit LSP ingress nodes to:

- A) Determine whether the link referenced in the sub-TLV affects any of the LSPs for which it is ingress. If there are, then:
- B) Determine whether those LSPs still meet end-to-end performance objectives. If not, then:
- C) The node could then conceivably move affected traffic to a pre-established protection LSP or establish a new LSP and place the traffic in it.

If link performance then improves beyond a configurable minimum value (reuse threshold), that sub-TLV can be re-advertised with the Anomalous bit cleared. In this case, a receiving node can conceivably do whatever re-optimization (or fallback) it wishes to do (including nothing).

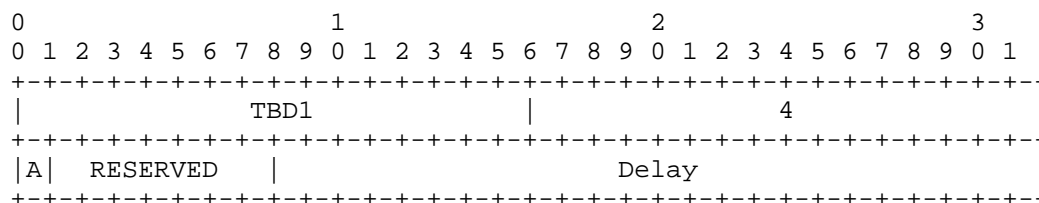
Note that when a sub-TLV does not include the A bit, that sub-TLV cannot be used for failover purposes. The A bit was intentionally omitted from some sub-TLVs to help mitigate oscillations. See section 7. 1. for more information.

Consistent with existing OSPF TE specifications (RFC3630), the bandwidth advertisements defined in this draft MUST be encoded as IEEE floating point values. The delay and delay variation advertisements defined in this draft MUST be encoded as integer values. Delay values MUST be quantified in units of microseconds, packet loss MUST be quantified as a percentage of packets sent, and bandwidth MUST be sent as bytes per second. All values (except residual bandwidth) MUST be calculated as rolling averages where the averaging period MUST be a configurable period of time. See section 5. for more information.

#### 4. Sub TLV Details

##### 4.1. Unidirectional Link Delay Sub-TLV

This sub-TLV advertises the average link delay between two directly connected OSPF neighbors. The delay advertised by this sub-TLV MUST be the delay from the local neighbor to the remote one (i.e. the forward path latency). The format of this sub-TLV is shown in the following diagram:



##### 4.1.1. Type

This sub-TLV has a type of TBD1.

##### 4.1.2. Length

The length is 4.

#### 4.1.3. A bit

This field represents the Anomalous (A) bit. The A bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A bit is clear, the sub-TLV represents steady state link performance.

#### 4.1.4. Reserved

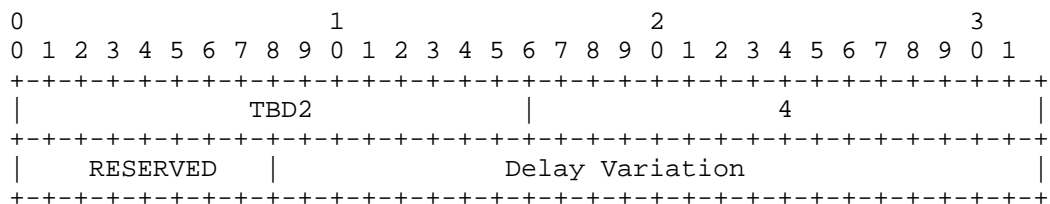
This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

#### 4.1.5. Delay Value

This 24-bit field carries the average link delay over a configurable interval in micro-seconds, encoded as an integer value. When set to 0, it has not been measured. When set to the maximum value 16,777,215 (16.777215 sec), then the delay is at least that value and may be larger.

### 4.2. Unidirectional Delay Variation Sub-TLV

This sub-TLV advertises the average link delay variation between two directly connected OSPF neighbors. The delay variation advertised by this sub-TLV MUST be the delay from the local neighbor to the remote one (i.e. the forward path latency). The format of this sub-TLV is shown in the following diagram:



#### 4.2.1. Type

This sub-TLV has a type of TBD2.

#### 4.2.2. Length

The length is 4.

#### 4.2.3. Reserved

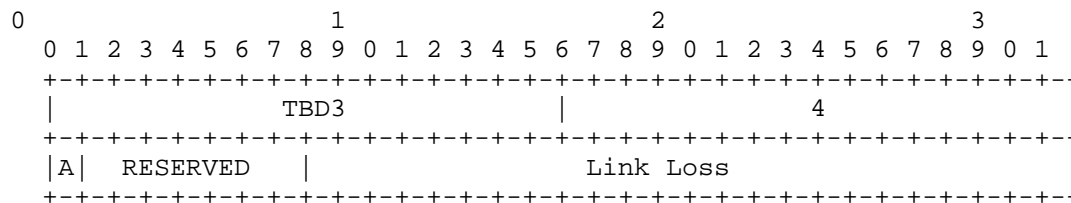
This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

#### 4.2.4. Delay Variation

This 24-bit field carries the average link delay variation over a configurable interval in micro-seconds, encoded as an integer value. When set to 0, it has not been measured. When set to the maximum value 16,777,215 (16.777215 sec), then the delay is at least that value and may be larger.

#### 4.3. Unidirectional Link Loss Sub-TLV

This sub-TLV advertises the loss (as a packet percentage) between two directly connected OSPF neighbors. The link loss advertised by this sub-TLV MUST be the packet loss from the local neighbor to the remote one (i.e. the forward path loss). The format of this sub-TLV is shown in the following diagram:



##### 4.3.1. Type

This sub-TLV has a type of TBD3

##### 4.3.2. Length

The length is 4

##### 4.3.3. A bit

This field represents the Anomalous (A) bit. The A bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below



its configured reuse threshold. If the A bit is clear, the sub-TLV represents steady state link performance.

#### 4.3.4. Reserved

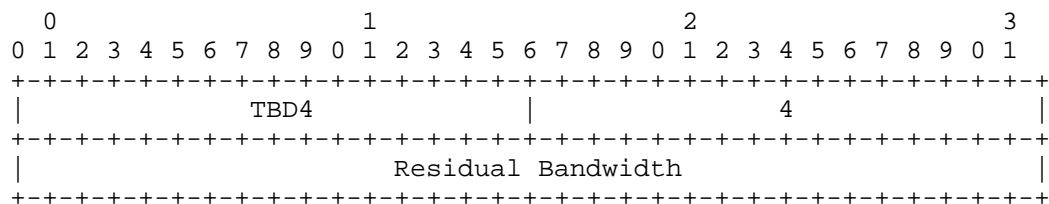
This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

#### 4.3.5. Link Loss

This 24-bit field carries link packet loss as a percentage of the total traffic sent over a configurable interval. The basic unit is 0.000003%, where  $(2^{24} - 2)$  is 50.331642%. This value is the highest packet loss percentage that can be expressed (the assumption being that precision is more important on high speed links than the ability to advertise loss rates greater than this, and that high speed links with over 50% loss are unusable). Therefore, measured values that are larger than the field maximum SHOULD be encoded as the maximum value. When set to a value of all 1s ( $2^{24} - 1$ ), the link packet loss has not been measured.

### 4.4. Unidirectional Residual Bandwidth Sub-TLV

This TLV advertises the residual bandwidth (defined in section 4.4.3. between two directly connected OSPF neighbors. The residual bandwidth advertised by this sub-TLV MUST be the residual bandwidth from the system originating the LSA to its neighbor. The format of this sub-TLV is shown in the following diagram:



#### 4.4.1. Type

This sub-TLV has a type of TBD4.

#### 4.4.2. Length

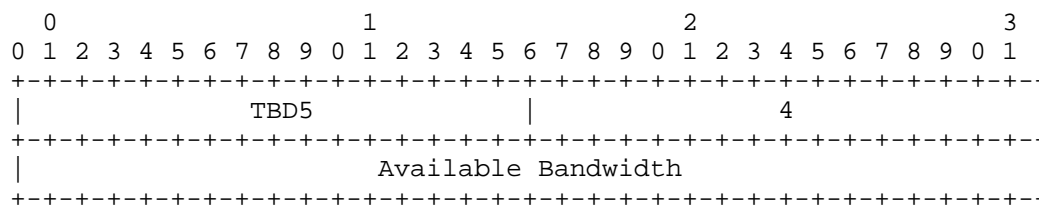
The length is 4.

#### 4.4.3. Residual Bandwidth

This field carries the residual bandwidth on a link, forwarding adjacency [RFC4206], or bundled link in IEEE floating point format with units of bytes per second. For a link or forwarding adjacency, residual bandwidth is defined to be Maximum Bandwidth [RFC3630] minus the bandwidth currently allocated to RSVP-TE LSPs. For a bundled link, residual bandwidth is defined to be the sum of the component link residual bandwidths.

#### 4.5. Unidirectional Available Bandwidth Sub-TLV

This TLV advertises the available bandwidth (defined in section 4.4.6. ) between two directly connected OSPF neighbors. The available bandwidth advertised by this sub-TLV MUST be the available bandwidth from the system originating the LSA to its neighbor. The format of this sub-TLV is shown in the following diagram:



#### 4.4.4. Type

This sub-TLV has a type of TBD5.

#### 4.4.5. Length

The length is 4.

#### 4.4.6. Available Bandwidth

This field carries the available bandwidth on a link, forwarding adjacency, or bundled link in IEEE floating point format with units of bytes per second. For a link or forwarding adjacency, available

bandwidth is defined to be residual bandwidth (see section 4.4. ) minus the measured bandwidth used for the actual forwarding of non-RSVP-TE LSP packets. For a bundled link, available bandwidth is defined to be the sum of the component link available bandwidths.

## 5. Announcement Thresholds and Filters

The values advertised in all sub-TLVs MUST be controlled using an exponential filter (i.e. a rolling average) with a configurable measurement interval and filter coefficient.

Implementations are expected to provide separately configurable advertisement thresholds. All thresholds MUST be configurable on a per sub-TLV basis.

The announcement of all sub-TLVs that do not include the A bit SHOULD be controlled by variation thresholds that govern when they are sent.

Sub-TLV that include the A bit are governed by several thresholds. Firstly, a threshold SHOULD be implemented to govern the announcement of sub-TLVs that advertise a change in performance, but not an SLA violation (i.e. when the A bit is not set). Secondly, implementations MUST provide configurable thresholds that govern the announcement of sub-TLVs with the A bit set (for the indication of a performance violation). Thirdly, implementations SHOULD provide reuse thresholds. These thresholds govern sub-TLV re-announcement with the A bit cleared to permit fail back.

## 6. Announcement Suppression

When link performance average values change, but fall under the threshold that would cause the announcement of a sub-TLV with the A bit set, implementations MAY suppress or throttle sub-TLV announcements. All suppression features and thresholds SHOULD be configurable.

## 7. Network Stability and Announcement Periodicity

To mitigate concerns about stability, all values (except residual bandwidth) MUST be calculated as rolling averages where the averaging

period MUST be a configurable period of time, rather than instantaneous measurements.

Announcements MUST also be able to be throttled using configurable inter-update throttle timers. The minimum announcement periodicity is 1 announcement per second.

## 8. Compatibility

As per (RFC3630), unrecognized TLVs should be silently ignored

## 9. Security Considerations

This document does not introduce security issues beyond those discussed in [RFC3630] and [RFC5329].

## 10. IANA Considerations

IANA maintains the registry for the sub-TLVs. OSPF TE Express Path will require one new type code per sub-TLV defined in this document.

## 11. References

### 11.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3630] Katz, D., Kompella, K., Yeung, D., "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.

### 11.2. Informative References

[RFC2328] Moy, J., "OSPF Version 2", RFC 2328, April 1998

- [RFC3031] Rosen, E., Viswanathan, A., Callon, R., "Multiprotocol Label Switching Architecture", January 2001
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC5250] Berger, L., Bryskin I., Zinin, A., Coltun, R., "The OSPF Opaque LSA Option", RFC 5250, July 2008.
- [Frost] D. Frost, S. Bryant "A Packet Loss and Delay Measurement Profile for MPLS-based Transport Networks"
- [Alto] R. Alimi R. Penno Y. Yang, "ALTO Protocol"

## 12. Acknowledgments

The authors would like to recognize Ayman Soliman for his contributions.

This document was prepared using 2-Word-v2.0.template.dot.

## 13. Author's Addresses

Spencer Giacalone  
Thomson Reuters  
195 Broadway  
New York NY 10007, USA

Email: [Spencer.giacalone@thomsonreuters.com](mailto:Spencer.giacalone@thomsonreuters.com)

Dave Ward  
Juniper Networks  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089, USA

Email: [dward@juniper.net](mailto:dward@juniper.net)

John Drake  
Juniper Networks

1194 N. Mathilda Ave.  
Sunnyvale, CA 94089, USA

Email: [jdrake@juniper.net](mailto:jdrake@juniper.net)

Alia Atlas  
Juniper Networks  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089, USA

Email: [akatlas@juniper.net](mailto:akatlas@juniper.net)

Stefano Previdi  
Cisco Systems  
Via Del Serafico 200  
00142 Rome  
Italy

Email: [sprevidi@cisco.com](mailto:sprevidi@cisco.com)



Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: October 15, 2011

D. Cheng  
Huawei Technologies  
M. Boucadair  
France Telecom  
April 13, 2011

Routing for IPv4-embedded IPv6 Packets  
draft-ietf-ospf-ipv4-embedded-ipv6-routing-00

## Abstract

This document describes routing packets destined to IPv4-embedded IPv6 addresses across IPv6 transit core using OSPFv3 with a separate routing table.

## Requirements Language

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 [RFC2119].

## Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on October 15, 2011.

## Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents



carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
1.1. The Scenario . . . . .	3
1.2. Routing Solution per RFC5565 . . . . .	4
1.3. An Alternative Routing Solution with OSPFv3 . . . . .	4
1.4. OSPFv3 Routing with a Specific Topology . . . . .	5
2. Provisioning . . . . .	6
2.1. Deciding the IPv4-embedded IPv6 Topology . . . . .	6
2.2. Maintaining a Dedicated IPv4-embedded IPv6 Routing Table . . . . .	6
2.3. OSPFv3 Topology with a Separate Instance ID . . . . .	6
2.4. OSPFv3 Topology with the Default Instance . . . . .	7
3. IP Packets Translation . . . . .	7
3.1. Address Translation . . . . .	8
4. Advertising IPv4-embedded IPv6 Routes . . . . .	8
4.1. Advertising IPv4-embedded IPv6 Routes into IPv6 Transit Network . . . . .	8
4.1.1. Routing Metrics . . . . .	9
4.1.2. Forwarding Address . . . . .	9
4.2. Advertising IPv4 Addresses into Client Networks . . . . .	9
5. Aggregation on IPv4 Addresses and Prefixes . . . . .	9
6. Forwarding . . . . .	10
7. MTU Issues . . . . .	10
8. Backdoor Connections . . . . .	11
9. Security Considerations . . . . .	11
10. IANA Considerations . . . . .	11
11. Acknowledgements . . . . .	11
12. References . . . . .	11
12.1. Normative References . . . . .	11
12.2. Informative References . . . . .	12
Authors' Addresses . . . . .	12

## 1. Introduction

This document describes a routing scenario where IPv4 packets are transported over IPv6 network.

In this document the following terminology is used:

- o An IPv4-embedded IPv6 address denotes an IPv6 address which contains an embedded 32-bit IPv4 address constructed according to the rules defined in [RFC6052].
- o IPv4-embedded IPv6 packets are packets of which destination addresses are IPv4-embedded IPv6 addresses.
- o AFBR (Address Family Border Router, [RFC5565]) refers to an edge router, which supports both IPv4 and IPv6 address families, of a backbone that supports only IPv4 or IPv6 address family.
- o AFXLBR (Address Family Translation Border Router) is defined in this document. It refers to a border router that supports both IPv4 and IPv6 address families, located on the boundary of IPv4-only network and IPv6-only network, and is capable of performing IP header translation between IPv4 and IPv6 according to [I-D.ietf-behave-v6v4-xlate].

### 1.1. The Scenario

Due to exhaustion of public IPv4 addresses, there has been continuing effort within IETF on IPv6 transitional techniques. In the course of transition, it is certain that networks based on IPv4 and IPv6 transfer capabilities, respectively, will co-exist at least for some time. One scenario of the co-existence is that IPv4-only networks inter-connecting with IPv6-only networks, and in particular, when an IPv6-only network serves as a transit network that inter-connects several segregated IPv4-only networks. In this scenario, IPv4 packets are transported over the IPv6 transit network between IPv4 networks. In order to forward an IPv4 packet from a source IPv4 network to the destination IPv4 network, IPv4 reachability information must be exchanged among involved networks by dedicated means.

Unlike dual-stack networks, operating an IPv6-only network would allow optimize OPEX and maintenance operations in particular. Some solutions have been proposed to allow delivery of IPv4 services over an IPv6-only network. This document focuses on an engineering techniques which aims to separate the routing instance dedicated to IPv4-embedded IPv6 destination from native IPv6 ones.

The purpose of running separate instances or topologies for IPv4-embedded IPv6 traffic is to distinguish from the native IPv6 routing topology, and the topology that is used for routing IPv4-embedded IPv6 datagram only. Separate instances/topologies are also meant to prevent any overload of the native IPv6 routing tables by IPv4-embedded IPv6 routes.

### 1.2. Routing Solution per RFC5565

The aforementioned scenario is described in [RFC5565], i.e.- IPv4-over-IPv6 scenario, where the network core is IPv6-only, and the inter-connected IPv4 networks are called IPv4 client networks. The P routers in the core only support IPv6 but the AFRs (Address Family Border Routers) support IPv4 on interface facing IPv4 client networks, and IPv6 on interface facing the core. The routing solution defined in [RFC5565] for this scenario is to run i-BGP among AFRs to exchange IPv4 routing information with each other, and the IPv4 packets are forwarded from one IPv4 client network to the other through a software using tunneling technology such as MPLS LSP, GRE, L2TPv3, etc.

### 1.3. An Alternative Routing Solution with OSPFv3

In this document, we propose an alternative routing solution for the scenario described in Section 1.1, where several segregated IPv4 networks, called IPv4 client networks, are interconnected by an IPv6 transit network, and in particular, we name the border node on the boundary of an IPv4 client network and the IPv6 transit network as Address Family Translation Border Router, or AFXLBR, which supports both IPv4 and IPv6 address families, and is capable of translating an IPv4 packet to an IPv6 packet, and vice versa, according to [I-D.ietf-behave-v6v4-xlate].

Since the scenario occurs most in a single ISP operating environment, an IPv6 prefix can be locally allocated and used to construct IPv4-embedded IPv6 addresses according to [RFC6052] by each AFXLBR, where the embedded IPv4 addresses are associated with an IPv4 client network that is connected to the AFXLBR, and each IPv4 address is an individual IPv4 address or prefix. An AFXLBR injects IPv4-embedded IPv6 addresses/prefixes into the IPv6 transit network using OSPFv3 and also installs those advertised by other AFXLBRs. When an IPv4 packet is sent from one IPv4 client network to the other, it is first encapsulated with an IPv6 header, where the source and destination IPv6 address are constructed, in a stateless manner, as IPv4-embedded IPv6 address, respectively, and then forwarded to the destination AFXLBR that is the advertising router of the destination IPv4-embedded IPv6 address. The destination AFXLBR replaces the IPv6 header by the corresponding IPv4 header, where the source and

destination IPv4 addresses are derived from the IPv4-embedded IPv6 source and destination addresses, respectively, and then forwards the IPv4 packet using its IPv4 routing table in the attached IPv4 client network.

There are use cases where the proposed routing solution is useful. One case is that some border nodes do not participate in i-BGP for routes exchange (one example is documented in [I-D.boucadair-softwire-dslite-v6only]), or i-BGP is not used at all. Another case is that tunnel mechanism is not used in the IPv6 transit network, or native IPv6 forwarding is preferred. Note also that with this routing solution, the IPv4-IPv6 inter-connection and associated header translation that occurs at an AFXLBR is stateless.

#### 1.4. OSPFv3 Routing with a Specific Topology

Routing IPv4-embedded IPv6 packets in the IPv6 transit network using OSPFv3, in general, may be performed by the OSPFv3 operation that is already running in the IPv6 network. One concern however, is that IPv4-embedded IPv6 routes would flood throughout the entire transit network and stored on every router, which may not be desirable. Also, since all IPv6 routes are stored in the same routing table, it might be more difficult to manage the resource required for routing and forwarding based on traffic category, if so desired. To solve this problem and to ease the separation between native IPv6 and IPv4-inferred routing policies, a separate OSPFv3 routing table can be constructed that is dedicated to IPv4-embedded IPv6 topology, and that table is solely used for routing IPv4-embedded IPv6 packets (i.e., IPv4 part of the Internet) in the transit network. Further, only a set of routers in the transit network are required to be involved in such routing scheme, including AFXLBRs that connect to IPv4 client networks along with a set of P routers in the core for routing path.

There are two methods to build a separate OSPFv3 routing table for IPv4-embedded IPv6 routing.

- o The first one is to run a separate OSPFv3 instance for IPv4-embedded IPv6 topology in the IPv6 transit network according to [RFC5838],
- o The second one is to stay with the existing OSPFv3 instance that already operates in the transit network, but maintain a separate IPv4-embedded topology, according to [I-D.ietf-ospf-mt-ospfv3].

With both methods, there would be a dedicated IPv4-embedded IPv6 topology that is maintained by OSPFv3 speakers and thus a dedicated IPv4-embedded IPv6 routing table, which is then used for routing

IPv4-embedded IPv6 packets (i.e., packets destined to an IPv4 destination). It would be operators' preference as which method is going to be used. This document elaborates on how configuration is done for each method and related routing issues that is common to both.

This document only focuses on unicast routing for IPv4-embedded IPv6 packets using OSPFv3.

## 2. Provisioning

### 2.1. Deciding the IPv4-embedded IPv6 Topology

Before making appropriate configuration in order to generate a separate OSPFv3 routing table for IPv4-embedded IPv6 addresses/prefixes, decision must be made on the set of routers and their interfaces in the IPv6 transit network that should be on the IPv4-embedded IPv6 topology.

For the purpose of this topology, all AFXLBRs that connect to IPv4 client networks should be members of this topology, and also at least some of their network core facing interfaces, which depends on which P routers in the IPv6 transit network would be on this topology.

The IPv4-embedded IPv6 topology is a sub-topology of the entire IPv6 transit network, and if all routers (including AFXLBRs and P-routers) and their interfaces are included, the two topologies converge. In general, as more P routers and their interfaces are configured on this sub-topology, it would increase the inter-connectivity and potentially, there would be more routing paths cross the transit network from one IPv4 client network to the other, at the cost that more routers need to participate the IPv4-embedded IPv6 routing. In any case, the IPv4-embedded IPv6 topology must be continuous with no partitions.

### 2.2. Maintaining a Dedicated IPv4-embedded IPv6 Routing Table

In an IPv6 transit network, in order to maintain a separate IPv6 routing table that contains routes for IPv4-embedded IPv6 destinations only, OSPFv3 needs to use the mechanism defined either in [RFC5838] or [I-D.ietf-ospf-mt-ospfv3] with required configuration tasks, as described in the following sub-sections.

### 2.3. OSPFv3 Topology with a Separate Instance ID

It is assumed that the scenario as described in this document is under a single ISP and as such, an OSPFv3 instance ID (IID) is

allocated locally and used for an OSPFv3 operation dedicated to unicast IPv4-embedded IPv6 routing in an IPv6 transit network. This IID is configured on each OSPFv3 interface of routers that participates in this routing instance.

The range for a locally configured OSPFv3 IID is from 128 to 255, inclusively, and this number must be used to encode the "Instance ID" field in the OSPFv3 packet header on every router that executes this instance in the IPv6 transit network.

In addition, the "AF" bit in the OSPFv3 Option field must be set.

During the Hello packets processing, adjacency may only be established when received Hello packets contain the same Instance ID as configured on the receiving interface for OSPFv3 instance dedicated to the IPv4-embedded IPv6 routing.

For more details, the reader is referred to [RFC5838].

#### 2.4. OSPFv3 Topology with the Default Instance

Similar to that as described in the previous section, an OSPFv3 multi-topology ID (MT-ID) is locally allocated and used for an OSPFv3 operation including unicast IPv4-embedded IPv6 routing in an IPv6 transit network. This MTID is configured on each OSPFv3 interface of routers that participates in this routing topology.

The range for a locally configured OSPFv3 MT-ID is from 32 to 255, inclusively, and this number must be used to encode the "MT-ID" field that is included in some of the extended LSAs as documented in [I-D.ietf-ospf-mt-ospfv3].

In addition, the MT bit in the OSPFv3 Option field must be set.

For more details, the reader is referred to [I-D.ietf-ospf-mt-ospfv3].

### 3. IP Packets Translation

When transporting IPv4 packets across an IPv6 transit network with the mechanism described above, an IPv4 packet is translated to an IPv6 packet at ingress AFXLBR, and the IPv6 packet is translated back to the original IPv4 packet at egress AFXLBR. The IP packet translation is accomplished in stateless manner according to rules specified in [I-D.ietf-behave-v6v4-xlate], with the address translation detail explained in the next sub-section.

### 3.1. Address Translation

Prior to the operation, an IPv6 prefix is allocated by the ISP and it is used to form an IPv4-embedded IPv6 address.

The IPv6 prefix can either be a well-known IPv6 prefix (WKP) 64:ff9b::/96, or a network-specific prefix that is unique to the ISP, and for the later case, the IPv6 prefix length may be 32, 40, 48, 56 or 64. In either case, this IPv6 prefix is used during the address translation between an IPv4 address and an IPv4-embedded IPv6 address, which is performed according to [RFC6052].

During translation from an IPv4 header to an IPv6 header at an ingress AFXLBR, the source IPv4 address and destination IPv4 address are translated into the corresponding IPv6 source address and destination IPv6 address, respectively, and during translation from an IPv6 header to an IPv4 header at an egress AFXLBR, the source IPv6 address and destination IPv6 address are translated into the corresponding IPv4 source address and destination IPv4 address, respectively. Note that the address translation is accomplished in a stateless manner.

## 4. Advertising IPv4-embedded IPv6 Routes

In order to forward IPv4 packets to the proper destination across IPv6 transit network, IPv4 reachability needs to be disseminated throughout the IPv6 transit network and this work is performed by AFXLBRs that connect to IPv4 client networks using OSPFv3.

With the scenario described in this document, i.e. - a set of AFXLBRs that inter-connect a bunch of IPv4 client networks with an IPv6 transit network, we view that IPv4 networks and IPv6 networks belong to separate Autonomous Systems, and as such, these AFXLBRs are OSPFv3 ASBRs.

### 4.1. Advertising IPv4-embedded IPv6 Routes into IPv6 Transit Network

IPv4 addresses and prefixes in an IPv4 client network are translated into IPv4-embedded IPv6 addresses and prefixes, respectively, using the same IPv6 prefix allocated by the ISP and the method specified in [RFC6052], and then advertised by one or more attached ASBRs into the IPv6 transit network using AS External LSA [RFC5340], i.e. - with the advertising scope throughout the entire Autonomous System.

#### 4.1.1. Routing Metrics

By default, the metric in an AS External LSA that carries an IPv4-embedded IPv6 address or prefixes is a Type 1 external metric, which is then to be added to the metric of an intra-AS path during OSPFv3 routes calculation. By configuration on an ASBR, the metric can be set to a Type 2 external metric, which is considered much larger than that on any intra-AS path. The detail is referred to OSPFv3 specification [RFC5340]. In either case, an external metric may be exact the same unit as in an IPv4 network (running OSPFv2 or others), but may also be specified by a routing policy, the detail is outside of the scope of this document.

#### 4.1.2. Forwarding Address

If the "Forwarding Address" field of an OSPFv3 AS External LSA is used to carry an IPv6 address, that must also be an IPv4-embedded IPv6 address where the embedded IPv4 address is the actual address in an IPv4 client network to which, data traffic is forwarded to. However, since an AFXLBR sits on the border of an IPv4 network and an IPv6 network, it is recommended that the "Forwarding Address" field not to be used by setting the F bit in the associated OSPFv3 AS-external-LSA to zero, so that the AFXLBR can make the forwarding decision based on its own IPv4 routing table.

#### 4.2. Advertising IPv4 Addresses into Client Networks

IPv4-embedded IPv6 routes injected into the IPv6 transit network from one IPv4 client network may be advertised into another IPv4 client network, after the associated destination addresses/prefixes are translated back to IPv4 addresses/prefixes format. This operation is similar to the regular OSPFv3 operation, wherein an AS External LSA can be advertised in a non-backbone area by default.

An IPv4 client network that does not want to receive such advertisement can be configured as a stub area or with other routing policy.

For the purpose of this document, IPv4-embedded IPv6 routes must not advertised into any IPv6 client networks that also connected to the IPv6 transit network.

### 5. Aggregation on IPv4 Addresses and Prefixes

In order to reduce the amount of AS External LSAs that are injected to the IPv6 transit network, effort must be made to aggregate IPv4 addresses and prefixes at each AFXLBR before advertising.



## 6. Forwarding

There are three cases in forwarding IP packets in the scenario as described in this document, as follows:

1. On an AFXLBR, if an IPv4 packet that is received on an interface connecting to an IPv4 client network with the destination IPv4 address belong to another IPv4 client network, the header of the packet is translated to a corresponding IPv6 header as described in Section 3, and the packet is then forwarded to the destination AFXLBR that advertises the IPv4-embedded IPv6 address through the IPv6 transit network.
2. On an AFXLBR, if an IPv4-embedded IPv6 packet is received and the embedded destination IPv4 address is in its IPv4 routing table, the header of the packet is translated to a corresponding IPv4 header as described in Section 3, and the packet is then forwarded accordingly.
3. On any router that is within the IPv4-embedded IPv6 topology located in the IPv6 transit network, if an IPv4-embedded IPv6 packet is received and a route is found in the IPv4-embedded IPv6 routing table, the packet is forwarded accordingly.

The classification of IPv4-embedded IPv6 packet is according to the IPv6 prefix of the destination address, which is either the Well Known Prefix (i.e., 64:ff9b::/96) or locally allocated as defined in [RFC6052].

## 7. MTU Issues

In the IPv6 transit network, there is no new MTU issue introduced by this document. If a separate OSPFv3 instance (per [RFC5838]) is used for IPv4-embedded IPv6 routing, the MTU handling in the transit network is the same as that of the default OSPFv3 instance. If a separate OSPFv3 topology (per [I-D.ietf-ospf-mt-ospfv3]) is used for IPv4-embedded IPv6 routing, the MTU handling in the transit network is the same as that of the default OSPFv3 topology.

However, the MTU in the IPv6 transit network may be different than that of IPv4 client networks. Since an IPv6 router will never fragment a packet, the packet size of any IPv4-embedded IPv6 packet entering the IPv6 transit network must be equal to or smaller than the MTU of the IPv6 transit network. In order to achieve this requirement, it is recommended that AFXLBRs to perform IPv6 path discovery among themselves and the resulting MTU, after taking into account of the difference between IPv4 header length and IPv6 header

length, must be "propagated" into IPv4 client networks, e.g.- included in the OSPFv3 Database Description packet.

The detail of passing the proper MTU into IPv4 client networks is beyond the scope of this document.

## 8. Backdoor Connections

In some deployments, there may exist direct connections among IPv4 client networks themselves in addition to the IPv6 transit network, as "backdoor" connections referring to, where IPv4 packets can either be transported between those IPv4 client networks via backdoor connections, or through the IPv6 transit network. In general, routing policies should be as such that the "backdoor" path is preferred since the packet forwarding is within a single address family without the need for IP header translation, among other things.

## 9. Security Considerations

This document does not introduce any security issue than what has been identified in [RFC5838], [I-D.ietf-ospf-mt-ospfv3] and [RFC6052].

## 10. IANA Considerations

No new IANA assignments are required for this document.

## 11. Acknowledgements

Many thanks to Acee Lindem, Dan Wing and Joel Halpern for their comments.

## 12. References

### 12.1. Normative References

[I-D.ietf-behave-v6v4-xlate]  
Li, X., Bao, C., and F. Baker, "IP/ICMP Translation Algorithm", draft-ietf-behave-v6v4-xlate-23 (work in progress), September 2010.

[I-D.ietf-ospf-mt-ospfv3]

Mirtorabi, S. and A. Roy, "Multi-topology routing in OSPFv3 (MT-OSPFv3)", draft-ietf-ospf-mt-ospfv3-03 (work in progress), July 2007.

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.
- [RFC5838] Lindem, A., Mirtorabi, S., Roy, A., Barnes, M., and R. Aggarwal, "Support of Address Families in OSPFv3", RFC 5838, April 2010.
- [RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.

## 12.2. Informative References

- [I-D.boucadair-softwire-dslite-v6only]  
Boucadair, M., Jacquenet, C., Grimault, J., Kassi-Lahlou, M., Levis, P., Cheng, D., and Y. Lee, "Deploying Dual-Stack Lite in IPv6 Network", draft-boucadair-softwire-dslite-v6only-01 (work in progress), April 2011.
- [RFC5565] Wu, J., Cui, Y., Metz, C., and E. Rosen, "Softwire Mesh Framework", RFC 5565, June 2009.

## Authors' Addresses

Dean Cheng  
Huawei Technologies  
2330 Central Expressway  
95050  
USA  
  
Email: dean.cheng@huawei.com

Mohamed Boucadair  
France Telecom  
Rennes,    35000  
France

Email: mohamed.boucadair@orange-ftgroup.com



OSPF Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: November 7, 2011

M. Bhatia  
Alcatel-Lucent  
S. Hartman  
Painless Security  
D. Zhang  
Huawei Technologies co., LTD.  
A. Lindem  
Ericsson  
May 6, 2011

Security Extension for OSPFv2 when using Manual Key Management  
draft-ietf-ospf-security-extension-manual-keying-00

Abstract

The current OSPFv2 cryptographic authentication mechanism as defined in the OSPF standards is vulnerable to both inter-session and intra-session replay attacks when it uses manual keying. Additionally, the existing cryptographic authentication schemes do not cover the IP header. This omission can be exploited to carry out various types of attacks.

This draft proposes changes to the authentication sequence number mechanism that will protect OSPFv2 from both inter-session and intra-session replay attacks when it uses manual keys for securing its protocol packets. Additionally, we also describe some changes in the cryptographic hash computation so that we eliminate most attacks that result because OSPFv2 does not protect the IP header.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on November 7, 2011.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
1.1. Requirements Section . . . . .	4
2. Replay Protection using Extended Sequence Numbers . . . . .	4
3. OSPF Packet Extensions . . . . .	5
4. OSPF Packet Key Selection . . . . .	6
4.1. Key Selection for Unicast OSPF Packet Transmission . . . . .	7
4.2. Key Selection for Multicast OSPF Packet Transmission . . . . .	7
4.3. Key Selection for OSPF Packet Reception . . . . .	8
5. Mechanism to secure the IP header . . . . .	8
6. Security Considerations . . . . .	9
7. IANA Considerations . . . . .	9
8. References . . . . .	10
8.1. Normative References . . . . .	10
8.2. Informative References . . . . .	10
Authors' Addresses . . . . .	11

## 1. Introduction

The OSPFv2 cryptographic authentication mechanism as described in [[RFC2328]] uses per-packet sequence numbers to provide protection against replay attacks. The sequence numbers increase monotonically so that the attempts to replay the stale packets can be thwarted. The sequence number values are maintained as a part of adjacency states. Therefore, if an adjacency is broken down, the associated sequence numbers get reinitialized and the neighbors start all over again. Additionally, the cryptographic authentication mechanism does not specify how to deal with the rollover of a sequence number when its value would wrap. These omissions can be taken advantage of by attackers to implement various replay attacks ([RFC6039]). In order to address these issues, we propose extensions to the authentication sequence number mechanism. Compared with the cryptographic authentication mechanism proposed in [RFC5709], the solution proposed does not impose any more security presumption.

The cryptographic authentication as described in [RFC2328] and later updated in [RFC5709] does not include the IP header. This also can be exploited to launch several attacks as the source address in the IP header is no longer protected. The OSPF specification, for broadcast and NBMA (Non-Broadcast Multi-Access Networks), requires the implementations to look at the source address in the IP header to determine the neighbor from which the packet was received. Changing the IP source address of a packet which can confuse the receiver and can be exploited to produce a number of denial of service attacks [RFC6039]. If the packet is interpreted as coming from a different neighbor, the sequence number received from the neighbor may be updated. This may disrupt communication with the legitimate neighbor. Hello packets may be reflected to cause a neighbor to appear to have one-way communication. Old Database descriptions may be reflected in cases where the per-packet sequence numbers are sufficiently divergent in order to disrupt an adjacency [I-D.ietf-karp-ospf-analysis]. This is referred to as the IP layer issue in [I-D.ietf-karp-threats-reqs].

[RFC2328] states that implementations MUST offer keyed MD5 authentication. It is likely that this will be deprecated in favor of the stronger algorithms described in [RFC5709] in future deployments [I-D.ietf-opsec-igp-crypto-requirements].

This draft proposes a simple change in the cryptographic authentication mechanism, as currently described in [RFC5709], to prevent such IP layer attacks.



### 1.1. Requirements Section

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 RFC2119 [RFC2119].

When used in lowercase, these words convey their typical use in common language, and are not to be interpreted as described in RFC2119 [RFC2119].

## 2. Replay Protection using Extended Sequence Numbers

In order to provide replay protection against both inter-session and intra-session replay attacks, the OSPFv2 sequence number is expanded to 64-bits with the least significant 32-bit value containing a strictly increasing sequence number and the most significant 32-bit value containing the boot count. OSPFv2 implementations are required to retain the boot count in non-volatile storage for the deployment life the OSPF router. The requirement to preserve the boot count is also placed on SNMP agents by the SNMPv3 security architecture (refer to snmpEngineBoots in [RFC4222]).

Since there is no room in the OSPFv2 packet for a 64-bit sequence number, it will occupy the 8 octets following the OSPFv2 packet and MUST be included when calculating the OSPFv2 packet digest. These additional 8 bytes are not included in the OSPFv2 packet header length but are included in the OSPFv2 header Authentication Data length and the IPv4 packet header length.

The lower order 32-bit sequence number MUST be incremented for every OSPF packet sent by the OSPF router. Upon reception, the sequence number MUST be greater than the sequence number in the last OSPF packet of that type accepted from the sending OSPF neighbor. Otherwise, the OSPF packet is considered a replayed packet and dropped. OSPF packets of different types may arrive out of order if they are prioritized as recommended in [RFC3414].

OSPF routers implementing this specification MUST use available mechanisms to preserve the sequence number's strictly increasing property for the deployed life of the OSPFv3 router (including cold restarts). This is achieved by maintaining a boot count in non-volatile storage and incrementing it each time the OSPF router loses its prior sequence number state. The SNMPv3 snmpEngineBoots variable [RFC4222] MAY be used for this purpose. However, maintaining a separate boot count solely for OSPF sequence numbers has the advantage of decoupling SNMP reinitialization and OSPF reinitialization. Also, in the rare event that the lower order 32-

bit sequence number wraps, the boot count can be incremented to preserve the strictly increasing property of the aggregate sequence number. Hence, a separate OSPF boot count is RECOMMENDED.

### 3. OSPF Packet Extensions

The OSPF packet header includes an authentication type field, and 64-bits of data for use by the appropriate authentication scheme (determined by the type field). Authentication types 0, 1 and 2 are defined [RFC2328]. This section of this defines Authentication type TBD (3 is recommended).

When using this authentication scheme, the 64-bit Authentication field in the OSPF packet header as defined in section D.3 of [RFC2328] is changed as shown below. The sequence number is removed and the Key ID is extended to 32 bits and moved to the former position of the sequence number.

Additionally, the 64-bit sequence number is moved to the first 64-bits following the OSPFv2 packet and is protected by the authentication digest. These additional 64 bits or 8 octets are included in the IP header length but not the OSPF header packet length.

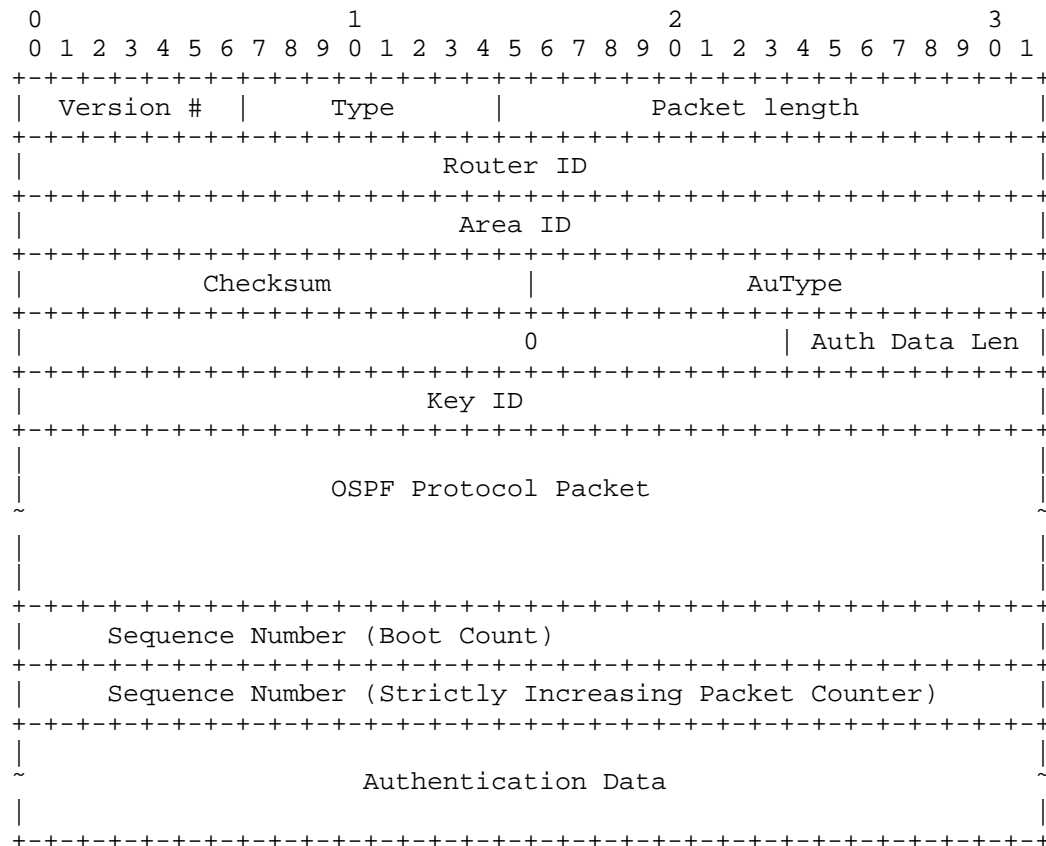


Figure 7 - Extended Sequence Number Packet Extensions

#### 4. OSPF Packet Key Selection

This section describes how the proposed security solution selects long-lived keys from key tables. [I-D.ietf-karp-crypto-key-table]. Generally, a key used for OSPFv2 packet authentication should satisfy the following requirements:

- o The key time period as defined by NotBefore and NotAfter must include the current time.
- o The key can be used for the desired security algorithm.

In the remainder of this section, additional requirements for keys are enumerated for different scenarios.

#### 4.1. Key Selection for Unicast OSPF Packet Transmission

Assume that a router R1 tries to send a unicast OSPF packet from its interface I1 to the interface R2 of a remote router R2 using security protocol P via interface I at time T. Firstly consider the circumstances where R1 and R2 are not connected with a virtual link. R1 then needs to select a long long-lived symmetric key from its key table. Because the key should be shared by the by both R1 and R2 to protect the communication between I1 and I2, the key should satisfy the following requirements:

- o The Peer field includes the router ID of R2.
- o the PeerKeyID field is not "unknown".
- o The Interfaces field includes I1.
- o the Direction field is either "out" or "both".

When R1 and R2 are connected to a virtual link, the third condition is a little more complex. Because the virtual link can be regarded as an unnumbered point-to-point network, the IP address of the interface actually used to send the packet (i.e., I1) is discovered during routing table calculation. Therefore, when the system operator configures keys to protect the virtual link, I1 is unknown and can be any OSPF interface in the OSPF virtual link's transit area. Therefore, the key should be identified solely by the local and remote router IDs rather than by the interface on which the packet is sent. The third requirement list above should be changed to "the Interface field includes the router ID".

#### 4.2. Key Selection for Multicast OSPF Packet Transmission

If a router R1 sends an OSPF packet from its interface I1 to a multicast address (e.g., AllSPFRouters, AllDRouters), it needs to select a key according to the following requirements:

- o The Peer field includes the multicast address.
- o The PeerKeyID field is "group".
- o The Interfaces field includes I1.
- o The Direction field is either "out" or "both".

#### 4.3. Key Selection for OSPF Packet Reception

When Cryptographic Authentication is employed, the ID of the authentication key is included in the authentication field of the OSPF packet header. Using this key ID, it is relatively easy for a receiver to locate the key. The simple requirements are:

- o The Peer field includes the router ID of the sender.
- o The PeerKeyID field includes the key ID obtained from the authentication field.
- o The Direction field is either "in" or "both".

#### 5. Mechanism to secure the IP header

This document updates the definition of Apad which is currently a constant defined in [RFC5709] to the source address from the IP header of the OSPFv2 protocol packet. The overall cryptographic authentication process defined in [RFC5709] remains unchanged. To reduce the potential for confusion, this section minimizes the repetition of text from RFC 5709 and is incorporated here by reference [RFC5709].

RFC 5709, Section 3.3, describes how the cryptographic authentication must be computed. It requires OSPFv2 packet's Authentication Trailer (which is the appendage described in RFC 2328, Section D.4.3, Page 233, items (6)(a) and (6)(d)) to be filled with the value Apad where Apad is a hexadecimal constant value 0x878FE1F3 repeated (L/4) times, where L is the length of the hash being used and is measured in octets rather than bits.

Routers at the sending side must initialize Apad to a value of the source address that would be used when sending out the OSPFv2 packet, repeated L/4 times, where L is the length of the hash, measured in octets. The basic idea is to incorporate the source address from the IP header in the cryptographic authentication computation so that any change of IP source address in a replayed packet can be detected.

At the receiving end, implementations MUST initialize Apad as the source address from IP Header of the incoming OSPFv2 packet, repeated L/4 times, instead of the constant that's currently defined in [RFC5709]. Besides changing the value of Apad, this document does not introduce any other changes to the authentication mechanism described in [RFC5709]. This would prevent all attacks where a rogue OSPF router changes the IP source address of an OSPFv2 packet and replays it on the same multi-access interface or another interface

since the IP source address is now protected and such changes would cause the authentication check to fail and the replayed packet to be rejected.

## 6. Security Considerations

This document attempts to fix the manual key management procedure that currently exists within OSPFv2, as part of the Phase 1 of the KARP Working Group. Therefore, only the OSPFv2 manual key management mechanism is considered. Any solution that takes advantage of the automatic key management mechanism is beyond the scope of this document.

The proposed sequence number extension offers most of the benefits of of more complicated mechanisms involving challenges. There are, however, a couple drawbacks to this approach. First, it requires the OSPF implementation to be able to save its boot count in non-volatile storage. If the non-volatile storage is ever repaired or upgraded such that the contents are lost or the OSPFv2 router is replaced with a model, the keys MUST be changed to prevent replay attacks.

Second, if a router is taken out of service completely (either intentionally or due to a persistent failure), the potential exists for reestablishment of an OSPFv2 adjacency by replaying the entire OSPFv2 session establishment. This scenario is however, extremely unlikely, since it would imply an identical OSPFv2 adjacency formation packet exchange. The replay of OSPFv2 hello packets alone for an OSPFv2 router that has been taken out of service should not result in any serious attack as the only consequence is superfluous processing. Of course, this attack could also be thwarted by changing the relevant manual keys.

This document also provides a solution to prevent certain denial of service attacks that can be launched by changing the source address in the IP header of the OSPFv2 protocol packet.

## 7. IANA Considerations

This document requests a new code point from the "OSPF Shortest Path First (OSPF) Authentication Codes" registry:

- o TBD - Cryptographic Authentication with Extended Sequence Numbers. The value 3 is recommended.

## 8. References

### 8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC5709] Bhatia, M., Manral, V., Fanto, M., White, R., Barnes, M., Li, T., and R. Atkinson, "OSPFv2 HMAC-SHA Cryptographic Authentication", RFC 5709, October 2009.

### 8.2. Informative References

- [I-D.ietf-karp-crypto-key-table]  
Housley, R. and T. Polk, "Database of Long-Lived Symmetric Cryptographic Keys", draft-ietf-karp-crypto-key-table-00 (work in progress), November 2010.
- [I-D.ietf-karp-ospf-analysis]  
Hartman, S. and D. Zhang, "Analysis of OSPF Security According to KARP Design Guide", draft-ietf-karp-ospf-analysis-00 (work in progress), March 2011.
- [I-D.ietf-karp-threats-reqs]  
Lebovitz, G., Bhatia, M., and R. White, "The Threat Analysis and Requirements for Cryptographic Authentication of Routing Protocols' Transports", draft-ietf-karp-threats-reqs-02 (work in progress), April 2011.
- [I-D.ietf-opsec-igp-crypto-requirements]  
Bhatia, M. and V. Manral, "Summary of Cryptographic Authentication Algorithm Implementation Requirements for Routing Protocols", draft-ietf-opsec-igp-crypto-requirements-04 (work in progress), October 2010.
- [RFC3414] Blumenthal, U. and B. Wijnen, "User-based Security Model (USM) for version 3 of the Simple Network Management Protocol (SNMPv3)", STD 62, RFC 3414, December 2002.
- [RFC4222] Choudhury, G., "Prioritized Treatment of Specific OSPF Version 2 Packets and Congestion Avoidance", BCP 112, RFC 4222, October 2005.
- [RFC6039] Manral, V., Bhatia, M., Jaeggli, J., and R. White, "Issues with Existing Cryptographic Protection Methods for Routing

Protocols", RFC 6039, October 2010.

Authors' Addresses

Manav Bhatia  
Alcatel-Lucent  
Bangalore,  
India

Phone:  
Email: manav.bhatia@alcatel-lucent.com

Sam Hartman  
Painless Security

Email: hartmans@painless-security.com

Dacheng Zhang  
Huawei Technologies co., LTD.  
Beijing,  
China

Phone:  
Fax:  
Email: zhangdacheng@huawei.com  
URI:

Acee Lindem  
Ericsson  
102 Carric Bend Court  
Cary, NC 27519  
USA

Phone:  
Email: acee.lindem@ericsson.com





OSPF  
Internet-Draft  
Intended status: Standards Track  
Expires: January 10, 2012

W. Lu  
Ericsson  
July 9, 2011

OSPF TE Extension for Area IDs  
draft-lu-ospf-area-tlv-01

Abstract

For multi-area path computation, it is desirable to have knowledge of area boundaries and the corresponding border routers which are capable of processing inter-area TE traffic. This memo defines a TLV to the OSPF TE extensions to meet such need.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on January 10, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
1.1. Background . . . . .	3
1.2. Current Solutions . . . . .	3
1.2.1. Global TED . . . . .	3
1.2.2. Stitch . . . . .	3
1.2.3. Crankback . . . . .	4
1.2.4. Distributed Path Computation . . . . .	4
1.3. Requirements Language . . . . .	4
1.4. Acronyms . . . . .	4
2. Definitions . . . . .	5
2.1. Exit Area . . . . .	5
2.2. TE-ABR . . . . .	5
3. Motivation . . . . .	6
3.1. The importance of ABR whereabouts . . . . .	6
3.2. Topic not Yet Addressed . . . . .	6
3.3. Benefits . . . . .	7
4. Scope of the proposal . . . . .	7
5. Area ID TLV . . . . .	7
5.1. TLV Origination Point . . . . .	8
5.2. TLV encoding . . . . .	9
5.3. Example . . . . .	9
6. Applications . . . . .	10
6.1. Use-Case 1 . . . . .	10
6.1.1. Crankback approach . . . . .	10
6.1.2. BRPC approach . . . . .	10
6.2. Use-Case 2 . . . . .	11
7. Acknowledgements . . . . .	12
8. IANA Considerations . . . . .	12
9. Security Considerations . . . . .	12
10. References . . . . .	12
10.1. Normative References . . . . .	12
10.2. Informative References . . . . .	12
Author's Address . . . . .	13

## 1. Introduction

### 1.1. Background

Traffic Engineering (TE) based networks are widely used by network operators. The provision and setup mechanisms work fine in a single IGP area thanks to the well defined TE extensions to the corresponding protocols, namely RSVP-TE [RFC3209], OSPF-TE [RFC3630], and ISIS-TE [RFC3784]. From the single area TE database, LSPs can be derived to meet various TE constraints using some Path Computation Element (PCE) methods such as CSPF.

The mechanisms however cannot be applied directly to multi-area networks, for which the path computation is one of the key applications of the PCE-based architecture [RFC4655].

It is highly desirable to compute inter-area shortest paths that satisfy some bandwidth constraints or any other constraints, with little manual intervention, as is possible within a single IGP area.

### 1.2. Current Solutions

Listed below are a few existing inter-area path computation mechanisms. As can be seen the ABR whereabouts are indispensable in computing inter-area LSPs.

#### 1.2.1. Global TED

A single TE database that contains all TE information of each and every area/domain is called the global TED. This certainly makes it easy to compute shortest path LSPs that meet all constraint requirements. The drawbacks nevertheless are apparent:

- a. IGP hierarchy enables improved IGP scalability by dividing the IGP domain into areas and limiting the flooding scope of topology information to within area boundaries. The global TED goes against this principle.
- b. Even if one is willing to compromise this principle, the LSPs created upon this global TED would lack of area information which may be required for enforcing path selection policies.

#### 1.2.2. Stitch

This method uses per-area path computation based on ERO expansion on the head-end LSR and on ABRs. ABRs can be selected through either:

- a. Static configuration of ABRs as loose hops at the head-end LSR;
- b. Dynamic ABR selection - Proprietary, knowledge of ABRs may be acquired through non-standard protocol modification.

#### 1.2.3. Crankback

Crankback method defined in [RFC4920] allows an LSP to be constructed to beyond the area scope provided some intermediate nodes, i.e. ABRs, are known. Crankback can probe the ABRs one after another till a viable path is found if it exists.

Note that this method does not allow computing an optimal path but just a feasible path. It may also have some non-negligible setup delay. These issues nevertheless are beyond the scope of this document.

#### 1.2.4. Distributed Path Computation

PCE architecture document [RFC4655] outlines a distributed PCE architecture. The idea is that various PCEs which have partial information of the topologies work together to conclude the best paths that meet the computation constraint requirements.

Individual PCEs may not have any visibility beyond the areas they are servicing. For inter-area path computation purpose, the knowledge of ABRs are essential for the neighboring PCEs to find out the overall best path over the combined areas.

BRPC as defined in [RFC5441] provides one of such methods. OSPF-CAP [RFC5088] on the other hand defines methods to make known distributed PCEs using OSPF capability TLVs.

Note that although [RFC5088] contains PCE-DOMAIN sub-TLV for OSPF Area ID, it however cannot be used for identifying area border LSRs. It is for locating PCEs, not LSRs.

#### 1.3. Requirements Language

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 [RFC2119].

#### 1.4. Acronyms

ABR - Area Border Router  
BRPC - Backward-Recursive Path Computation  
CSPF - Constrained Shortest Path First  
IGP - Interior Gateway Protocol  
LSA - Link State Advertisement  
LSDB - Link-State DataBase  
LSP - Label Switched Path  
LSR - Label Switching Router  
OSPF - Open Shortest Path First  
PCC - Path Computation Client  
PCE - Path Computation Element  
TE - Traffic Engineering  
TED - Traffic Engineering Database  
TLV - Type Length Value  
VSPT - Virtual Shortest Path Tree

## 2. Definitions

The following definitions are under the context where TE is enabled.

### 2.1. Exit Area

A neighboring OSPF area to which an inter-area path can possibly be extended is called Exit Area.

The ABR which connects the current area to an Exit Area is called exit ABR.

### 2.2. TE-ABR

An exit ABR is also referred to as TE-ABR.

### 3. Motivation

#### 3.1. The importance of ABR whereabouts

All solutions listed in Section 1.2, except the global TED approach, have to use the knowledge of exit ABRs to accomplish their inter-area path computation tasks.

Some methods require manual configuration which is costly and error prone. Others may have to use proprietary means to acquire the information.

It is desirable to have the exit ABR information available and make it conveniently accessible to the relevant PCEs without adding lot of complexity to the protocols nor too much burden to the participating routers.

#### 3.2. Topic not Yet Addressed

Apart from the manual provision, currently the exit ABR information is difficult to acquire. The reasons are:

1. OSPF TE Database (TED) does not contain information on exit ABRs;
2. CSPF operates on the TED and is therefore limited to the information the latter provides. Unless the critical information of the exit ABRs becomes available, the CSPF cannot operate optimally by seeing beyond the area scope.
3. One may argue that CSPF can dig into OSPF's other repositories, such as LSDB, to find out the ABR whereabouts. This is not advisable because it negates the purpose of keeping TED opaque and independent from the normal OSPF operation. This is also technically difficult because usually CSPF and OSPF are two different processes that they may not be running in the same nodes.
4. Even if one is willing for CSPF to intrude into OSPF space, and use the ABR bit (B bit) information for locating border routers, these ABRs are not necessarily the exit ABRs. In other words, even if CSPF learns which routers sit on area borders, it is still unable to ascertain whether the ABRs are supporting TE on the other side of the borders. OSPF's hierarchical design limits the topology sharing to within area boundaries.
5. And even if one is willing to jump across this limit and somehow manages to acquire the ABRs' TE ability onto other areas, there is still the need for one more key information, the Area IDs.

6. The Area IDs are critical to the distributed PCE architecture. They are essential in enforcing path computation area sequences and PCE policies. They are also useful for consolidating path computation jobs. Section 6.2 provides an example of ABR consolidation.
7. It is important to understand that the proposed TLV also implies TE capabilities. In other words the TLV signifies that the advertising router is not only an ABR, but also an LSR capable of handling TE traffic. Unless tied with TE knowledge, methods of discovering of ABRs will not be useful in locating TE-ABRs, i.e. the LSRs that can transit TE traffic across areas.
8. At last, since the TLV is TE based, it should be defined in the OSPF TE extensions and maintained similarly with its counterparts, Router Address TLV and Link TLV.

### 3.3. Benefits

The benefits of having an OSPF Area ID TLV are listed below:

1. It fits into OSPF TE architecture, preserves the protocol's hierarchy, and adds little burden to the OSPF process;
2. It automates the TE-ABR discovery, and eliminates the need of manual provisioning;
3. Very often an LSR is also a PCE. If this LSR is also an ABR, it can compute a two-area LSP effortlessly. The PCC only needs to send the request to one TE-ABR (if there are multiple TE-ABRs sharing the same border), provided it has knowledge of the TE-ABR whereabouts.

### 4. Scope of the proposal

This document describes solutions for inter-area path computation and does not address inter-domain scenarios.

It is also specific to OSPF as an IGP protocol, though the concepts apply to ISIS which are to be defined in a separate document.

### 5. Area ID TLV

[RFC3630] section 2.4 defines two TLVs. This memo adds a third TLV called the Area ID TLV.





## 5.2. TLV encoding

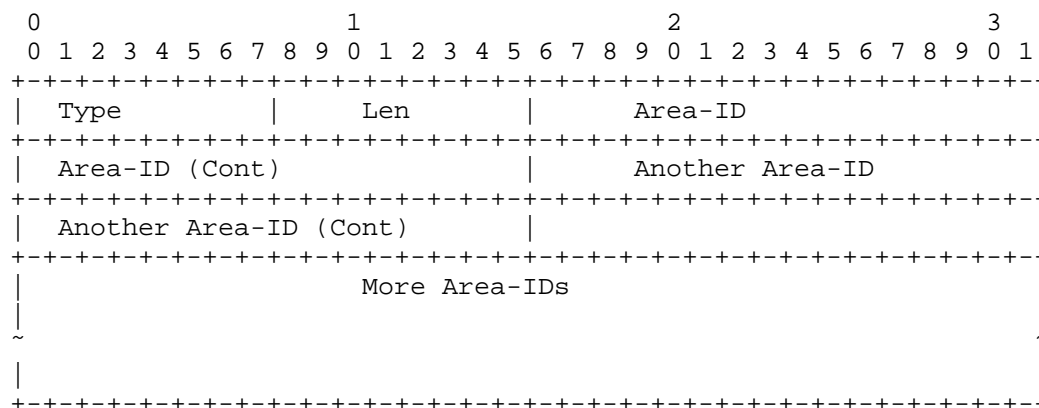


Figure 2: TLV Format

- Type** The Area ID TLV consists of a 1-octet type value which will be allocated by the IANA (suggested value 3).
- Len** 1-octet unsigned integer value 4xN to represent N exit areas, not including the originating area.
- Value** The value fields contains one or more Area IDs. Each Area ID is a four-octet OSPF Area ID associated with an exit area for which the ABR has TE enabled. An ABR may connect to multiple areas. Therefore it may generate 1 TLV with N IDs, where N is the total number of exit areas, not including the originating area. Since the Len field is of 1-octet, this TLV can hold upto 63 IDs. For non-ABR routers this Area-ID TLV SHOULD not occur.

## 5.3. Example

Consider Figure 1 again. Router ABR3 connects to four areas Area0, Area2, Area3, and Area4. Assuming all ABRs and areas are TE enabled except Area4 which is not TE-enabled. ABR3 originates and floods following Area-ID TLVs as shown in Figure 3, assuming type is 3.

To	Type	Len	Area-IDs	...
Area0	3	8	0 0 0 2	0 0 0 3
Area2	3	8	0 0 0 0	0 0 0 3
Area3	3	8	0 0 0 0	0 0 0 2
Area4	None			

Figure 3: Sample TLV

## 6. Applications

### 6.1. Use-Case 1

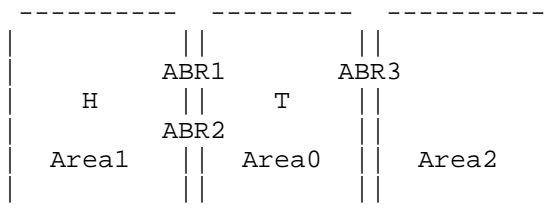


Figure 4: Use-Case 1

The topology is shown in Figure 4. The headend "H" is in Area1. The tailend "T" is in Area0. LSRs are running OSPF-TE and RSVP-TE.

#### 6.1.1. Crankback approach

The crankback method requires a list of ABRs for tryout. The list usually has to be provisioned manually.

With the proposed Area-ID TLVs, the ABR information is made available through the OSPF TE database. Therefore "H" learns the ABR list dynamically from its OSPF TED, which is ABR1 and ABR2. "H" starts the crankback procedure with "ABR1" from which it reaches "T". The LSP thus is established successfully, though not necessarily optimal.

#### 6.1.2. BRPC approach

With the method defined in [RFC5441], the VSPT has to be first built from "T" in Area0. Given that the area sequence is Area0->Area1, the initial VSPT has the candidate exit ABRs ABR1, ABR2, and ABR3.

Now with the proposed OSPF Area ID TLVs the PCE of Area0 knows that ABR3 connects Area2, not Area1, therefore ABR3 is not considered as an exit ABR.

Subsequently the PCE of Area1 takes the VSPT passed by the PCE of Area0 as input and concludes the end-to-end path computation.

## 6.2. Use-Case 2

Very often an LSR is also a PCE. In that case, the Area ID TLVs can make the path computation job more efficient.

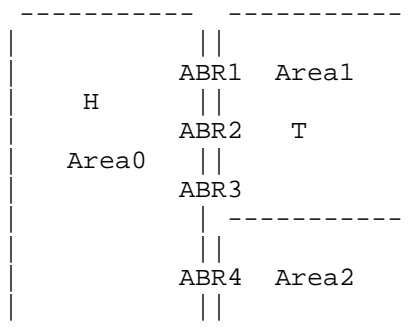


Figure 5: Use-Case 2

As shown in Figure 5, if the headend "H" knows that the tailend "T", or some intermediate node to reach, is in Area1, it sends a PCReq to one of the ABRs that connect to Area1. The draft proposal makes the ABR list conveniently available, which is [ABR1, ABR2, ABR3]. Note that ABR4 is not listed since it is not an exit ABR to Area1.

Assuming by some local policy ABR1 is chosen. Since ABR1 sits across Area0 and Area1, it has visibility to TEDs of both areas. ABR1 which is also a PCE performs the path computation job in the same way as an intra-area path computation.

Note that the resulting path, if existent, does not necessarily go through ABR1. It can be for example H->ABR3->T.

Note also that the ABR selection is a local decision. One can use some criteria, for example the highest te-router-id, to select the ABR. However, the candidate ABRs have to share the common border such as the one between Area0 and Area1. This can be achieved by grouping ABRs according to their exit Area IDs in the proposed OSPF Area ID TLVs.

## 7. Acknowledgements

The author would like to thank Sriganesh Kini, Meral Shirazipour, and Dimitri Papadimitriou for their reviews and comments.

## 8. IANA Considerations

This document defines the following TLV to the OSPF TE Extensions under TE LSA:

Type	Name	Source
TBD (recommend 3)	Area ID TLV	This document

## 9. Security Considerations

There are no specific security considerations within the scope of this document.

## 10. References

### 10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.

### 10.2. Informative References

- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3784] Smit, H. and T. Li, "Intermediate System to Intermediate System (IS-IS) Extensions for Traffic Engineering (TE)", RFC 3784, June 2004.
- [RFC4105] Le Roux, J., Vasseur, J., and J. Boyle, "Requirements for

Inter-Area MPLS Traffic Engineering", RFC 4105, June 2005.

- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4674] Le Roux, J., "Requirements for Path Computation Element (PCE) Discovery", RFC 4674, October 2006.
- [RFC4920] Farrel, A., Satyanarayana, A., Iwata, A., Fujita, N., and G. Ash, "Crankback Signaling Extensions for MPLS and GMPLS RSVP-TE", RFC 4920, July 2007.
- [RFC4927] Le Roux, J., "Path Computation Element Communication Protocol (PCECP) Specific Requirements for Inter-Area MPLS and GMPLS Traffic Engineering", RFC 4927, June 2007.
- [RFC5088] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.
- [RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.

#### Author's Address

Wenhu Lu  
Ericsson  
300 Holger Way  
San Jose, California 95134  
USA

Phone: 408 750-5436  
Email: wenhu.lu@ericsson.com



Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: April 25, 2011

N. Sheth  
L. Wang  
J. Zhang  
Juniper Networks  
October 25, 2010

OSPF Hybrid Broadcast and P2MP Interface Type  
draft-nsheth-ospf-hybrid-bcast-and-p2mp-01.txt

Abstract

This document describes a mechanism to model a broadcast network as a hybrid of broadcast and point-to-multipoint networks for purposes of OSPF operation. Neighbor discovery and maintenance as well as LSA database synchronization are performed using the broadcast model, but the network is represented using the point-to-multipoint model in the router LSAs of the routers connected to it. This allows an accurate representation of the cost of communication between different routers on the network, while maintaining the network efficiency of broadcast operation. This approach is relatively simple and requires minimal changes to OSPF.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

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/lid-abstracts.txt>.

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

This Internet-Draft will expire on April 25, 2011.

Copyright Notice



Copyright (c) 2010 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

# Table of Contents

1. Introduction . . . . .	3
2. Motivation . . . . .	4
3. Operation . . . . .	5
3.1. Interface Parameters . . . . .	5
3.2. Neighbor Data Structure . . . . .	5
3.3. Neighbor Discovery and Maintenance . . . . .	5
3.4. Database Synchronization . . . . .	5
3.5. Generating Network LSAs . . . . .	5
3.6. Generating Router and Intra-Area-Prefix-LSAs . . . . .	5
3.6.1. Stub Links in OSPFv2 Router LSA . . . . .	6
3.6.2. OSPFv3 Intra-Area-Prefix-LSA . . . . .	6
3.7. Next-Hop Calculation . . . . .	6
3.8. Graceful Restart . . . . .	6
4. Compatibility Considerations . . . . .	8
5. Scalability and Deployment Considerations . . . . .	9
6. Security Considerations . . . . .	10
7. IANA Considerations . . . . .	11
8. Normative References . . . . .	12
Authors' Addresses . . . . .	13

## 1. Introduction

OSPF [RFC2328] operation on broadcast interfaces takes advantage of the broadcast capabilities of the underlying medium for doing neighbor discovery and maintenance. Further, it uses a Designated Router and Backup Designated Router to keep the LSA databases of the routers on the network synchronized in an efficient manner. However, it has the limitation that a router cannot advertise different costs to each of the neighboring routers on the network in its router LSA.

Operation on point-to-multipoint interfaces could require explicit configuration of the identity of its neighboring routers. It also requires the router to send separate hellos to each neighbor on the network. Further, it mandates establishment of adjacencies to all all configured or discovered neighbors on the network. However, it gives the routers the flexibility to advertise different costs to each of the neighboring routers in their router LSAs.

This document proposes a new interface type that can be used on layer 2 networks that have broadcast capability. In this mode, neighbor discovery and maintenance, as well as database synchronization are performed using existing procedures for broadcast mode. The network is modeled as a collection of point-to-point links in the router LSA, just as it would be in point-to-multipoint mode. This new interface type is referred to as hybrid-broadcast-and-p2mp in the rest of this document.

## 2. Motivation

There are some layer 2 networks that are broadcast capable but have a potentially different cost associated with communication between any given pair of nodes. The cost could be based on the underlying layer 2 topology as well as various link quality metrics such as bandwidth, delay and jitter among others.

It is not accurate to treat such networks as OSPF broadcast networks since that does not allow a router to advertise a different cost to each of the other routers. Using OSPF point-to-multipoint mode would satisfy the requirement to correctly describe the cost to reach each router. However, it would be inefficient in the sense that it would require forming  $O(N^2)$  adjacencies when there are  $N$  routers on the network.

It is advantageous to use the hybrid-broadcast-and-p2mp type for such networks. This combines the flexibility of point-to-multipoint type with the advantages and efficiencies of broadcast interface type.

### 3. Operation

OSPF routers supporting the capabilities described herein should have support for an additional hybrid-broadcast-and-p2mp type for the Type data item described in section 9 of [RFC2328].

The following sub-sections describe salient aspects of OSPF operation on routers configured with a hybrid-broadcast-and-p2mp interface.

#### 3.1. Interface Parameters

Routers MUST support configuration of the Router Priority for the interface.

The default value of the LinkLSASuppression is "disabled". It MAY be set to "enabled" via configuration.

#### 3.2. Neighbor Data Structure

Routers MUST support an additional field called the Neighbor Output Cost. This is the cost of sending a data packet to the neighbor, expressed in the link state metric. The default value of this field is the Interface output cost. It MAY be set to a different value using mechanisms which are outside the scope of this document, like static per-neighbor configuration, or any dynamic discovery mechanism that is supported by the underlying network.

#### 3.3. Neighbor Discovery and Maintenance

Routers send and receive Hellos so as to perform neighbor discovery and maintenance on the interface using the procedures specified for broadcast interfaces in [RFC2328] and [RFC5340].

#### 3.4. Database Synchronization

Routers elect a DR and BDR for the interface and use them for initial and ongoing database synchronization using the procedures specified for broadcast interfaces in [RFC2328] and [RFC5340].

#### 3.5. Generating Network LSAs

Since a hybrid-broadcast-and-p2mp interface is described in router LSAs using a collection of point-to-point links, the DR SHOULD NOT generate a network LSA for the interface.

#### 3.6. Generating Router and Intra-Area-Prefix-LSAs

Routers describe the interface in their router LSA as specified for a point-to-multipoint interface in section 12.4.1.4 of [RFC2328] and section 4.4.3.2 of [RFC5340], with the following modifications for

Type 1 links:

- o If a router is not the DR, it MUST NOT add any Type 1 links if it does not have a full adjacency to the DR.
- o If a router is not the DR and has a full adjacency to the DR, it MUST add a Type 1 link corresponding to each neighbor that is in state 2-Way or higher.
- o The cost for a Type 1 link corresponding to a neighbor SHOULD be set to the value of the Neighbor Output Cost field as defined in Section 3.2

#### 3.6.1. Stub Links in OSPFv2 Router LSA

Routers MUST add a Type 3 link for their own IP address to the router LSA as described in section 12.4.1.4 of [RFC2328]. Further, they MUST also add a Type 3 link with the Link ID set to the IP subnet address, Link Data set to the IP subnet mask, and cost equal to the configured output cost of the interface.

#### 3.6.2. OSPFv3 Intra-Area-Prefix-LSA

Routers MUST add global scoped IPv6 addresses on the interface to the intra-area-prefix-LSA as described for point-to-multipoint interfaces in section 4.4.3.9 of [RFC5340]. In addition, they MUST also add all global scoped IPv6 prefixes on the interface to the LSA by specifying the PrefixLength, PrefixOptions, and Address Prefix fields. The Metric field for each of these prefixes is set to the configured output cost of the interface.

The DR SHOULD NOT generate an intra-area-prefix-LSA for the transit network for this interface since it does not generate a network LSA for the interface. Note that the global prefixes associated with the interface are advertised in the intra-area-prefix-LSA for the router as described above.

#### 3.7. Next-Hop Calculation

Next-Hops to destinations that are directly connected to a router via the interface are calculated as specified for a point-to-multipoint interface in section 16.1.1 of [RFC2328].

#### 3.8. Graceful Restart

The following modifications to the procedures defined in section 2.2, item 1 of [RFC3623] are required in order to ensure that the router correctly exits graceful restart.

- o If a router is the DR on the interface, it MUST NOT examine the pre-restart network LSA for the interface in order to determine the previous set of adjacencies.
- o If a router is in state DROther on the interface, it MUST consider an adjacency to non-DR and non-BDR neighbors as reestablished when the neighbor state reaches 2-Way.

#### 4. Compatibility Considerations

All routers on the network must support the hybrid-broadcast-and-p2mp interface type for successful operation. Otherwise, the interface should be configured as a standard broadcast interface.

If some routers on the network treat the interface as broadcast and others as hybrid-broadcast-and-p2mp, neighbors and adjacencies will still get formed as for a broadcast interface. However, due to the differences in how router and network LSAs are built for these two interface types, there will be no traffic traversing certain pairs of routers. Note that this will not cause any persistent loops or black holing of traffic.



## 5. Scalability and Deployment Considerations

Treating a broadcast interface as hybrid-broadcast-and-p2mp results in  $O(N^2)$  links to represent the network instead of  $O(N)$ , when there are  $N$  routers on the network. This will increase memory usage and have a negative impact on route calculation performance on all the routers in the area. Network designers should carefully weigh the benefits of using the new interface type against the disadvantages mentioned here.

## 6. Security Considerations

This document raises no new security issues for OSPF. Security considerations for the base OSPF protocol are covered in [RFC2328] and [RFC5340].

## 7. IANA Considerations

This document has no IANA considerations.

This section should be removed by the RFC Editor to final publication.

## 8. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.
- [RFC3623] Moy, J., Pillay-Esnault, P., and A. Lindem, "Graceful OSPF Restart", RFC 3623, November 2003.

Authors' Addresses

Nischal Sheth  
Juniper Networks  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089  
US

Email: nsheth@juniper.net

Lili Wang  
Juniper Networks  
10 Technology Park Dr.  
Westford, MA 01886  
US

Email: lilw@juniper.net

Jeffrey Zhang  
Juniper Networks  
10 Technology Park Dr.  
Westford, MA 01886  
US

Email: zzhang@juniper.net



OSPF Working Group  
Updates: 5614  
Internet-Draft  
Intended status: Experimental  
Expires: November 18, 2011

R. Ogier  
May 17, 2011

Use of OSPF-MDR in Single-Hop Broadcast Networks  
draft-ogier-ospf-manet-single-hop-00.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

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/lid-abstracts.html>

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

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Abstract

RFC 5614 (OSPF-MDR) extends OSPF to support mobile ad hoc networks (MANETs) by specifying its operation on the new OSPF interface of type MANET. This document describes the use of OSPF-MDR in a single-hop broadcast network, which is a special case of a MANET in which each router is a (one-hop) neighbor of each other router. Unlike an OSPF broadcast interface, such an interface can have a different cost associated with each neighbor. The document includes configuration recommendations and simplified mechanisms that can be used in single-hop networks.

## 1. Introduction

OSPF-MDR [RFC5614] specifies an extension of OSPF [RFC2328, RFC5340] to support mobile ad-hoc networks (MANETs) by specifying its operation on the new OSPF interface of type MANET. OSPF-MDR generalizes the Designated Router (DR) to a connected dominating set (CDS) consisting of a typically small subset of routers called MANET Designated Routers (MDRs). Similarly, the Backup Designated Router (BDR) is generalized to a subset of routers called Backup MDRs (BMDRs). MDRs achieve scalability in MANETs similar to the way DRs achieve scalability in broadcast networks:

- o MDRs have primary responsibility for flooding LSAs. Backup MDRs provide backup flooding when MDRs temporarily fail.
- o MDRs allow the number of adjacencies to be dramatically reduced, by requiring adjacencies to be formed only between MDR/BMDR routers and their neighbors.

In addition, OSPF-MDR has the following features:

- o MDRs and BMDRs are elected based on information obtained from modified Hello packets received from neighbors.
- o If adjacency reduction is used (the default), adjacencies are formed between MDRs so as to form a connected subgraph. An option (AdjConnectivity = 2) allows for additional adjacencies to be formed between MDRs/BMDRs to form a biconnected subgraph.
- o Each non-MDR router becomes adjacent with an MDR called its Parent, and optionally (if AdjConnectivity = 2) becomes adjacent with another MDR or BMDR called its Backup Parent.
- o Each router advertises connections to its neighbor routers as point-to-point links in its router-LSA. Network-LSAs are not used.
- o In addition to full-topology LSAs, partial-topology LSAs may be used to reduce the size of router-LSAs. Such LSAs are formatted as standard LSAs, but advertise links to only a subset of neighbors.
- o Optionally, differential Hellos can be used, which reduce overhead by reporting only changes in neighbor states.

This document describes the use of OSPF-MDR in a single-hop broadcast network, which is a special case of a MANET in which each router is a (one-hop) neighbor of each other router. Unlike an OSPF broadcast interface, such an interface can have a different cost associated with each neighbor. An example use case is when the underlying radio system performs layer-2 routing, but has a different number of (layer-2) hops to (layer-3) neighbors.



Section 2 describes the operation of OSPF-MDR in a single-hop broadcast network with recommended parameter settings. Section 3 describes an alternative procedure which may be used to decide which neighbors on a single-hop broadcast network to advertise in the router-LSA. Section 4 describes a simplified version of the MDR selection algorithm for single-hop networks.

The alternative procedure of Section 3 and the simplified algorithm of Section 4 are optional and MUST NOT be used if it is possible for two routers in the network to be more than one hop from each other.

### 1.1. Terminology

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 [RFC2119].

## 2. Operation in a Single-Hop Broadcast Network

When OSPF-MDR is used in a single-hop broadcast network, the following parameter settings and options (defined in [RFC5614]) should be used:

- o AdjConnectivity SHOULD be equal to 2 (biconnected), MAY be equal to 1 (uniconnected), and SHOULD NOT be equal to 0 (full topology).
- o An adjacency SHOULD be eliminated if neither the router nor the neighbor is an MDR or BMDR (see Section 7.3 of [RFC5614]).
- o LSAPFullness MUST be equal to 4 or 5 if full-topology LSAs are required. (The value 5 is defined in Section 3 of this document.)
- o LSAPFullness MAY be equal to 1 (min-cost LSAs) if full-topology LSAs are not required. This option reduces the number of advertised links while still providing shortest paths.

If AdjConnectivity equals 1 or 2 and full-topology LSAs are used, OSPF-MDR running on a single-hop broadcast network has the following properties:

- o A single MDR is selected, which becomes adjacent with every other router, as in an OSPF broadcast network.
- o Two BMDRs are selected. This occurs because the MDR selection algorithm ensures that the MDR/BMDR backbone is biconnected. If AdjConnectivity = 2, every non-MDR/BMDR router becomes adjacent with one of the BMDRs in addition to the MDR.
- o When all adjacencies are fully adjacent, the router-LSA for each router includes point-to-point (type 1) links to all bidirectional

neighbors (in state 2-Way or greater).

### 3. Originating Router-LSAs

A router running OSPF-MDR with LSAPFullness = 4 includes in its router-LSA point-to-point (type 1) links for all fully adjacent neighbors, and for all bidirectional neighbors that are routable. A neighbor is routable if the SPF calculation has produced a route to the neighbor and a flexible quality condition is satisfied.

This section describes an alternative procedure which MAY be used instead of the procedure described in Section 6 of [RFC5614], to decide which neighbors on a single-hop broadcast network to advertise in the router-LSA. The alternative procedure will correspond to LSAPFullness = 5, and is interoperable with the other choices for LSAPFullness. This procedure avoids the need to check whether a neighbor is routable, and thus avoids having to update the set of routable neighbors.

If LSAPFullness = 5, then the Selected Advertised Neighbor Set (SANS) is the same as specified for LSAPFullness = 4, and the following steps are performed instead of the first paragraph of Section 9.4 in [RFC5614].

- (1) The MDR includes in its router-LSA a point-to-point (type 1) link for each fully adjacent neighbor. (Note that the MDR becomes adjacent with all of its neighbors.)
- (2) Each non-MDR router includes in its router-LSA a point-to-point link for each fully adjacent neighbor, and, if the router is fully adjacent with the MDR, for each bidirectional neighbor j such that the MDR's router-LSA includes a link to j.

#### 3.1. Discussion

To discuss the above procedure, let i and j be two non-MDR routers. Since the SPF calculation (Section 16.1 of [RFC2328]) allows router i to use router j as a next hop only if router j advertises a link back to router i, routers i and j must both advertise a link to each other in their router-LSAs before either can use the other as a next hop. Therefore, the above procedure for non-MDR routers (Step 2) implies there must exist a path of fully adjacent links between i and j (via the MDR) in both directions before this can happen.

The above procedure for non-MDR routers (Step 2) is similar to one described in Section 3.6 of [HYBRID] for non-DR routers. However, the latter procedure only requires that the router be fully adjacent with the DR, and does not require that the DR's router-LSA include a link to the neighbor j. Thus (based on the previous paragraph) it ensures only that routers i and j are fully adjacent with the DR

before either can use the other as a next hop. As a result, the DR might not be fully adjacent with router i or j, and thus i and j may not be fully synchronized. Note that full adjacency with a neighbor does not imply that the link state databases are synchronized (see footnote 23 in [RFC2328]). It only means that the router is at least as up-to-date as the neighbor, since it only means that all Link State Requests have been satisfied.

#### 4. MDR Selection Algorithm

The MDR selection algorithm of [RFC5614] simplifies as follows in single-hop networks. The resulting algorithm is similar to the DR election algorithm of OSPF, but is slightly different (e.g., two Backup MDRs are selected). The following simplified algorithm is interoperable with the full MDR selection algorithm.

Note that lexicographic order is used when comparing tuples of the form (RtrPri, MDR Level, RID). Also note that each router will form adjacencies with its parents and dependent neighbors. In the following, the term "neighbor" refers to a bidirectional neighbor (in state 2-Way or greater).

Phase 1 (creating the neighbor connectivity matrix) is not required.

Phase 2: MDR Selection

- (2.1) The set of Dependent Neighbors is initialized to be empty.
- (2.2) If the router has a larger value of (RtrPri, MDR Level, RID) than all of its (bidirectional) neighbors: the router selects itself as an MDR, selects its BMDR neighbors as Dependent Neighbors if AdjConnectivity = 2, then proceeds to Phase 4.
- (2.3) Otherwise, if the router's MDR Level is currently MDR, then it is changed to BMDR before executing Phase 3.

Phase 3: Backup MDR Selection

- (3.1) Let Rmax be the neighbor with the largest value of (RtrPri, MDR Level, RID).
- (3.2) Determine whether or not there exist two neighbors, other than Rmax, with a larger value of (RtrPri, MDR Level, RID) than the router itself.
- (3.3) If there exist two such neighbors, then the router sets its MDR Level to MDR Other.
- (3.4) Else, the router sets its MDR Level to BMDR, and if AdjConnectivity = 2, adds Rmax and its MDR/BMDR neighbors as

Dependent Neighbors.

- (3.5) If steps 3.1 through 3.4 resulted in the MDR Level changing from MDR Other to BMDR, then execute Step 2.2 again before proceeding to Phase 4. (This is necessary because running Step 2.2 again can cause the MDR Level to change to MDR.)

#### Phase 4: Parent Selection

Each router selects a Parent and (if AdjConnectivity = 2) a Backup Parent for the single-hop broadcast network. The Parent for a non-MDR router will be the MDR. The Backup Parent for an MDR Other, if it exists, will be a BMDR. Each non-MDR router becomes adjacent with its Parent and its Backup Parent, if it exists. The parent selection algorithm is already simple, so a simplified version is not given here.

The Parent and Backup Parent are analogous to the Designated Router and Backup Designated Router interface data items in OSPF. As in OSPF, these are advertised in the DR and Backup DR fields of each Hello sent on the interface.

#### 5. Security Considerations

This document describes the use of OSPF-MDR in a single-hop broadcast network, and raises no security issues in addition to those already covered in [RFC5614].

#### 6. IANA Considerations

This document has no IANA considerations.

#### 7. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.
- [RFC5614] Ogier, R. and P. Spagnolo, "Mobile Ad Hoc Network (MANET) Extension of OSPF Using Connected Dominating Set (CDS) Flooding", RFC 5614, August 2009.

## 8. Informative References

[HYBRID] Sheth, N., L. Wang, and J. Zhang, "OSPF Hybrid Broadcast and P2MP Interface Type", draft-nsheth-ospf-hybrid-bcast-and-p2mp-01.txt, October 2010, work in progress.

## Author's Address

Richard G. Ogier  
Email: rich.ogier@earthlink.net



Network Working Group  
Internet-Draft  
Expires: January 12, 2012

Y. Ohara  
JAIST  
A. Kato  
Keio Univ.  
Jul 11, 2011

Consideration on OSPF LSDB Monitoring  
draft-ohara-ospf-lsdb-monitoring-consideration-01

Abstract

Many people believe that any LSA once flooded throughout the OSPF area can be monitored on all OSPF routers in the area. This is not always true, and a malicious OSPF router that pretends to be legal may want to, and be able to, hide malicious LSAs. This document proposes the modifications to OSPF specification to prevent hiding the malicious LSAs, and to make LSDB monitoring more successful (hence, secure).

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on January 12, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must

include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. The Insecure LSDB Monitoring: a failure to capture the malicious LSA . . . . .	5
3. The Secure LSDB Monitoring . . . . .	7
4. Suggested Modifications to the OSPF Specification . . . . .	8
5. Conservative Deployment . . . . .	9
6. References . . . . .	10
Appendix A. Acknowledgements . . . . .	11
Appendix B. Changes . . . . .	12
B.1. Changes from -00 . . . . .	12
Authors' Addresses . . . . .	13



## 1. Introduction

All systems should be monitored to confirm that it is not executing any undesirable activities. OSPF [RFC2328], a core protocol for IP network, is also such a system.

In OSPF, it is believed that one can monitor the whole area by monitoring LSDB change in any router in the area. This is due to the OSPF's link state principle; all information is flooded to all routers in the area, so that all routers can compute their routes autonomously and independently (distributed computation). If a network operator is watching LSA changes in the area at one point, it is almost equivalent to monitoring OSPF activities at all routers in the area, due to the nature of the link state protocol.

However, the LSDB monitoring technique (watching LSDB exchanges of a router at one point in the area) has a few problems.

1. The origin of an LSA can easily be spoofed. There is no way to make it sure where an LSA is originated, other than to monitor OSPF traffic of all the link in the OSPF area. (Monitoring OSPF traffic of all the link is very expensive in terms of operational cost, and usually is very difficult.) Hence, in general, an LSA can be easily overridden or removed by another router when the source origin ID of the LSA (i.e., Advertising Router) is spoofed.
2. Monitoring all the contents of many LSAs and detecting incorrect contents (either erroneous or malicious) are sometimes tough task. The large number of types of LSAs and their data fields are making it hard to detect an unusual information, unless the LSDB monitor parses all LSA formats and their contents (usually not done). An example of this consequence is the problem of Covert Channel using OSPF, where malicious communication channel using OSPF is open, while the existence of the channel is not obvious.
3. It is possible that some of the OSPF activities in the area cannot be monitored by a single monitoring point. In a very rare case depending on the timing of the LSA flooding and the topology, some LSA instances may not reach to all OSPF routers in the area, and hence not to the monitoring point. A detailed example is described in Section 2. "The Insecure LSDB Monitoring: a failure to capture the malicious LSA."

As a first step to make the OSPF LSDB monitoring more secure, this document focuses on the last bullet (bullet 3), assuming that if all OSPF activities are guaranteed to be monitored, bullets 1 and 2 (the

problems of finding the true LSA origin and unusual LSA contents) both become easier to solve.

Many OSPF operators tend to think that "all OSPF routers in the area see all the LSAs flooded in the area". This is not correct, to be precise. The fact that "an OSPF LSA might not entirely be flooded throughout the OSPF area" and "Some contents in the LSAs might not be captured by some of the other OSPF routers in the area" are not so commonly understood.

We denote that the LSDB monitoring at one point in the area is "insecure" because it may fail to capture all LSAs. The goal of this document is to provide a way to achieve "secure" LSDB monitoring without excessive operational burden, where LSDB monitoring always capture all LSA instances. The secure LSDB monitoring surely contributes to the problem of finding the unusual LSA contents such as malicious activities.

This document proposes some modifications to the OSPF specification, to guarantee that "every LSA content flooded in the area is always delivered to all OSPF routers" (i.e. the provision of the secure LSDB monitoring). To achieve this, this document proposes two modifications.

First, this document proposes to add a check on reception of a premature-aging (MaxAged) LSA. The check is to see if the contents is the same between the local LSA copy (being removed) and the received LSA (being removing). This additional check prevents all LSA contents from being removed without being flooded in the area at least once.

Second, this document proposes to add two additional conditions on updating LSA instances. They are 1) that the LSA instance is not listed on any retransmission-list, and 2) that the increment of LS Sequence Number is just 1. Those are to guarantee that all LSA contents are flooded in the area before being overridden with newer contents. The former condition strictly mandates that the LSDB is completely synchronized on all routers in the area in each step. The latter prevents the skipping (hence, dropping) of any LSA instance on any router.

With these modifications, all LSA contents are guaranteed to be flooded throughout the area. Hence, the secure LSDB monitoring works on any router as expected. Details of the proposal is explained in Section 3. "The Secure LSDB Monitoring."

## 2. The Insecure LSDB Monitoring: a failure to capture the malicious LSA

The malicious OSPF routers might try to withdraw the malicious LSA (say, immediately after the malicious contents accomplished its purpose), trying best to hide its malicious activity, from all other routers in the area. As a consequence of the withdraw, some of OSPF routers in the OSPF area may not see the malicious LSA, depending on the timing of LSA flooding and the topology. An example is explained in this section.

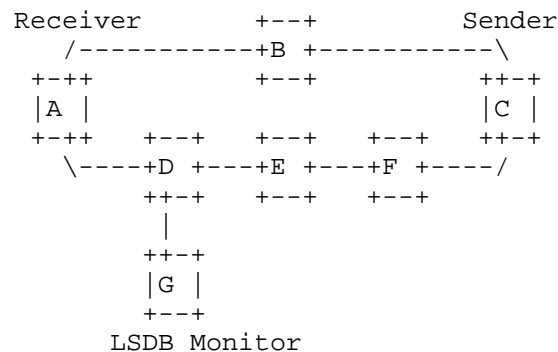


Figure 1: Example Topology

Suppose that the two malicious OSPF routers (the router A and C in Figure 1) are trying to communicate with each other. The router C floods a malicious LSA to convey a malicious contents to be received by the router A. Suppose that when the malicious LSA reached to the router A, the malicious LSA instances are held in the LSDBs of the routers marked with "x" (Figure 2).

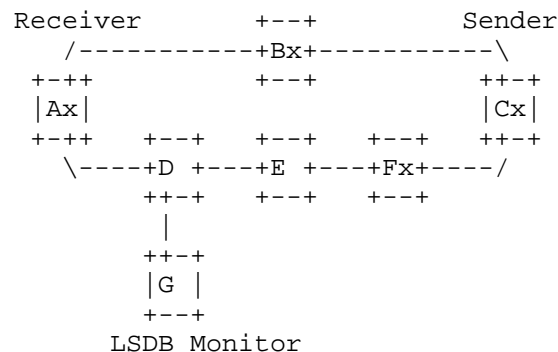


Figure 2: Flooding of the Malicious LSA

When the router A receives the malicious LSA, it immediately execute the procedure of premature aging for the malicious LSA, trying to hide the malicious LSA from all other routers. In doing so, the router A advertises the Max-Aged, empty contents, instance of the malicious LSA, spoofing the Advertising Router field. Let us illustrate the Max-Aged LSA "p". The LSA "p" is flooded from the router A, while the original malicious LSA instance "x" is also kept flooding (from the router F to E). The status becomes now as in Figure 3.

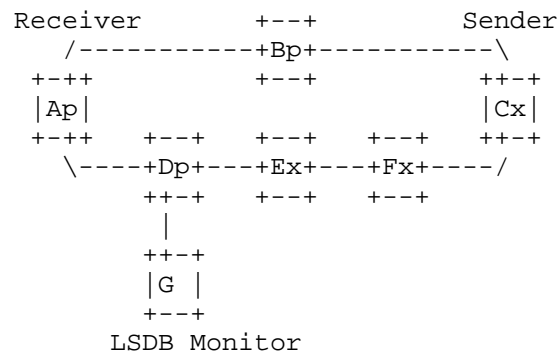


Figure 3: Premature Aging

Note that the LSDB Monitoring router G does not receive the malicious LSA "x", because "p" is recognized more recent than "x" and the LSA "x" is rejected in the router D. Hence, the LSDB monitor fails to capture the malicious contents of "x" because the only LSA flooded to the router G, "p", does not contain the malicious contents.

### 3. The Secure LSDB Monitoring

The source of the problem in Section 2 lies in the acceptance check in premature aging procedure. The current OSPF specification allows to receive the Max-Aged instance of the previous LSA, with the different contents (possibly empty). We propose to modify the OSPF specification to mandate additional identity check of the contents in premature aging procedure. By doing so, the malicious LSA (the router A in Section 2) cannot hide the malicious contents from the LSDB monitor (the router G) using premature aging procedure. The LSA "p" (i.e., a premature-aged version of "x" with empty contents) is rejected by the routers B and D, and the LSA "x" is flooded further to the router G.

However, the router A still be able to hide the malicious contents, using ordinal LSA update procedure. Suppose if the router A overrides the LSA "x" with the newer LSA with contents that seems to be valid. This case does not fall into the premature aging procedure, so the above change does not take effect.

To prevent hiding using ordinal LSA update, we propose to modify OSPF specification further, to mandate that every LSAs are updated only when all the retrans-lists do not contain the LSA instance and the newer LSA is advanced just one LS Sequence Number. This modification mandates that all the LSA instance is flooded throughout the area, and restricts that the reject of the stale update does not happen.

For the case of updating using the same LS Sequence Number, the current OSPF specification does not allow overwrite. When the router A tries to override the LSA "x" with the same LS Sequence Number, no routers will receive the LSA because it is recognized that the LSA is the same instance with "x", and that receiving it again is not needed. Hence, the attempt to hide by overriding the LSA does not succeed.

The most important disadvantages of this modification is that the LSA update is totally sequentialized among the entire network. The LSA update goes as LS SeqNum 1, 2, 3, and so on, with advance of just 1. This may make the efficacy and the speed of LSA flooding to degrade. This is the cost of achieving the secure LSDB monitoring.

#### 4. Suggested Modifications to the OSPF Specification

We propose the following modifications to the OSPF specification.

1. The premature aging can happen only when the LSA contents are identical between old and new (i.e., removed and removing) LSAs.
2. All LSA are updated only when 1) none of its instance is on any retrans-list, and 2) the LS Sequence Number is incremented by 1.

These conditions are added either in 13-(5)-(a) or in between 13-(5)-(a) and 13-(5)-(b), of [RFC2328]. If these conditions fail on a LSA, the LSA is silently discarded (i.e. without acknowledging it), just the same as the LSA arrived within MinLSArrival (13-(5)-(a)). This modifications make 13-(5)-(c) (replacing the instances on retrans-lists) never occur.

## 5. Conservative Deployment

Even just logging the occurrences of the failures of these new conditions, and acting the unmodified protocol behavior, will help network operators to find erroneous or illegal incidents on their OSPF network.

## 6. References

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.



## Appendix A. Acknowledgements

This document is supported by a commissioned research named "Research and Development on Evaluating Security of Communication Protocols and its Implementations" (2011) of National Institute of Information and Communications Technology (NICT), Japan.

## Appendix B. Changes

### B.1. Changes from -00

Described the modifications specifically in 1. "Introduction."

Added the specific part and the consequences of the modifications in 4. "Suggested Modifications to the OSPF Specification."

Added 5. "Conservative Deployment" as a moderate proposal.

Minor fixes of texts.

Added References, Acknowledgements, and Changes sections.

Authors' Addresses

Yasuhiro Ohara  
Japan Advanced Institute of Science and Technology  
Asahidai 1-1  
Nomi, Ishikawa 923-1292  
Japan

Email: [yasu@jaist.ac.jp](mailto:yasu@jaist.ac.jp)  
URI: <http://www.jaist.ac.jp/~yasu/>

Akira Kato  
Keio University, Graduate School of Media Design  
Hiyoshi 4-1-1, Kohoku  
Yokohama, Kanagawa 223-8526  
Japan

Email: [kato@wide.ad.jp](mailto:kato@wide.ad.jp)

