

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 01, 2014

F.J. Baker
Cisco Systems
August 28, 2013

Using OSPFv3 with Token-based Access Control
draft-baker-ipv6-ospf-dst-flowlabel-routing-03

Abstract

This note describes the changes necessary for OSPF to route IPv6 traffic specified prefix if and only if the packet contains an authorization token.

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 March 01, 2014.

Copyright Notice

Copyright (c) 2013 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	2
1.1. Requirements Language	2
2. Theory of Routing	2
2.1. Dealing with ambiguity	3
2.2. Interactions with other constraints	3
3. Extensions necessary for IPv6 Authenticated Routing in OSPF .	4
3.1. Authorization Token TLV	4
4. IANA Considerations	4
5. Security Considerations	4
6. Acknowledgements	5
7. References	5
7.1. Normative References	5
7.2. Informative References	5
Appendix A. Change Log	5
Author's Address	5

1. Introduction

This specification builds on OSPF for IPv6 [RFC5340] and its extensible LSA, defined in OSPFv3 LSA Extendibility [I-D.acee-ospfv3-lsa-extend]. This note defines the TLV for an IPv6 [RFC2460] Flow Label, to define routes from to a destination prefix qualified by an authorization token.

The approach may be combined with other qualifying attributes, such as routing "to that destination AND from a specified source". The obvious application is data center inter-tenant routing using a form of token-based access control. If the sender doesn't know the value to insert in the flow label or hop-by-hop option (the receiver's tenant ID), he in effect has no route to that destination.

1.1. 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 [RFC2119].

2. Theory of Routing

Both IS-IS and OSPF perform their calculations by building a lattice of routers and links from the router performing the calculation to each router, and then use routes (sequences in the lattice) to get to destinations that those routes advertise connectivity to. Following the SPF algorithm, calculation starts by selecting a starting point (typically the router doing the calculation), and successively adding {link, router} pairs until one has calculated a route to every router in the network. As each router is added, including the original

router, destinations that it is directly connected to are turned into routes in the route table: "to get to 2001:db8::/32, route traffic to {interface, list of next hop routers}". For immediate neighbors to the originating router, of course, there is no next hop router; traffic is handled locally.

In this context, the route is qualified by an authorization token, carried in the flow label or a hop-by-hop option; It is installed into the FIB with the destination prefix, and the FIB applies the route if and only if the token in the packet matches the token in the route. Of course, there may be multiple LSPs in the RIB with the same destination and differing authorization tokens; these may also have the same or differing next hop lists. The intended forwarding action is to forward matching traffic to one of the next hop routers associated with this destination and authorization tokens, or to discard non-matching traffic as "destination unreachable".

LSAs that lack an authorization token TLV match any token that may be present, by definition.

2.1. Dealing with ambiguity

In any routing protocol, there is the possibility of ambiguity. For example, one router might advertise a fairly general prefix - a default route, a discard prefix (which consumes all traffic that is not directed to an instantiated subnet), or simply an aggregated prefix while another router advertises a more specific one. In source/destination routing, potentially ambiguous cases include cases in which the link state database contains two routes A->B' and A'->B, in which A' is a more specific prefix within the prefix A and B' is a more specific prefix within the prefix B. Traditionally, we have dealt with ambiguous destination routes using a "longest match first" rule. If the same datagram matches more than one destination prefix advertised within an area, we follow the route with the longest matching prefix.

In this case, we follow a similar but slightly different rule; the FIB lookup MUST yield the route with the longest matching destination prefix that also matches the authorization token. A FIB route with no such token matches any authorization token.

2.2. Interactions with other constraints

In the event that there are other constraints on routing, such as proposed in [I-D.baker-ipv6-ospf-dst-src-routing], the effect is a logical AND. The FIB lookup must yield the route with the longest matching destination prefix that also matches each of the constraints.

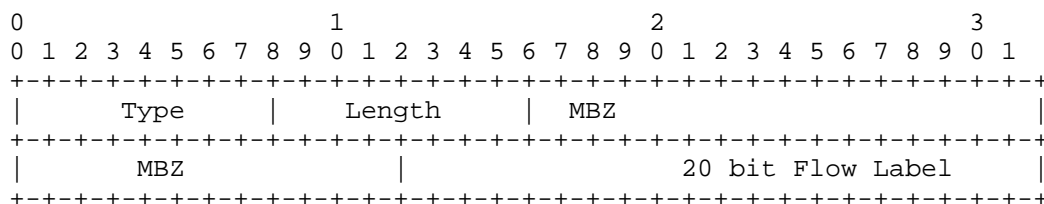
3. Extensions necessary for IPv6 Authenticated Routing in OSPF

Section 2 of [RFC5340] defines the "IPv6 Reachability TLV", and carries in it destination prefix advertisements. It has the capability of extension, using TLVs.

In this model, the flow label is used to prove that the datagram's sender has specific knowledge of its intended receiver. No proof is requested; this is left for higher layer exchanges such as IPSec or TLS. However, if the information is distributed privately, such as through DHCP/DHCPv6, the network can presume that a system that marks traffic with the right flow label has a good chance of being authorized to communicate with its peer.

The key consideration, in this context, is that the flow label is a 20 bit number. As such, an advertised route requiring a given flow label value is calling for an exact match of all 20 bits of the label value.

3.1. Authorization Token TLV



Source Prefix Sub-TLV

Source Prefix Type: assigned by IANA

TLV Length: Length of the TLV in octets

Flow Label: Flow Label value (20 bits)

4. IANA Considerations

The source prefix type mentioned in Section 3 must be defined.

5. Security Considerations

Network layer Token-based Access Control is part of a security solution. It is not, in itself, a complete solution. It acts as a pervasive network layer firewall, preventing unauthorized traffic from arriving at a destination. However, as in any network, a host is its own last bastion of defense; it needs IPsec or TLS-style

authorization and authorization of its peers, and must refuse traffic that contains the authorization token but is in fact malicious.

6. Acknowledgements

7. References

7.1. Normative References

[I-D.acee-ospfv3-lsa-extend]

Lindem, A., Mirtorabi, S., Roy, A., and F. Baker, "OSPFv3 LSA Extendibility", draft-acee-ospfv3-lsa-extend-00 (work in progress), May 2013.

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

[RFC2460] Deering, S.E. and R.M. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.

7.2. Informative References

[I-D.baker-ipv6-ospf-dst-src-routing]

Baker, F., "IPv6 Source/Destination Routing using OSPFv3", draft-baker-ipv6-ospf-dst-src-routing-02 (work in progress), May 2013.

Appendix A. Change Log

Initial Version: February 2013

updated Version: August 2013

Author's Address

Fred Baker
Cisco Systems
Santa Barbara, California 93117
USA

Email: fred@cisco.com

OSPF
Internet-Draft
Intended status: Standards Track
Expires: March 01, 2014

F.J. Baker
Cisco Systems
August 28, 2013

IPv6 Source/Destination Routing using OSPFv3
draft-baker-ipv6-ospf-dst-src-routing-03

Abstract

This note describes the changes necessary for OSPFv3 to route IPv6 traffic from a specified prefix to a specified prefix.

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 March 01, 2014.

Copyright Notice

Copyright (c) 2013 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	2
1.1.	Requirements Language	2
2.	Theory of Routing	2
2.1.	Notation	3
2.2.	Dealing with ambiguity	3
2.3.	Interactions with other constraints	4
3.	Extensions necessary for IPv6 Source/Destination Routing in OSPFv3	4
3.1.	IPv6 Source Prefix TLV	4
4.	IANA Considerations	5
5.	Security Considerations	5
6.	Acknowledgements	5
7.	References	5
7.1.	Normative References	5
7.2.	Informative References	6
	Appendix A. Change Log	6
	Author's Address	6

1. Introduction

This specification builds on OSPF for IPv6 [RFC5340] and the extensible LSAs defined in [I-D.acee-ospfv3-lsa-extend]. It defines the TLV for an IPv6 [RFC2460] Source Prefix, to define routes from a source prefix to a destination prefix.

This implies not simply routing "to a destination", but routing "to that destination AND from a specified source". It may be combined with other qualifying attributes, such as "traffic going to that destination AND using a specified flow label AND from a specified source prefix". The obvious application is egress routing, as required for a multihomed entity with a provider-allocated prefix from each of several upstream networks. Traffic within the network could be source/destination routed as well, or could be implicitly or explicitly routed from "any prefix", `::/0`. Other use cases are described in [I-D.baker-rtgwg-src-dst-routing-use-cases]. If a FIB contains a route to a given destination from one or more prefixes not including `::/0`, and a given packet destined there that has a source address that is in none of them, the packet in effect has no route, just as if the destination itself were not in the route table.

1.1. 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 [RFC2119].

2. Theory of Routing

Both IS-IS and OSPF perform their calculations by building a lattice of routers and links from the router performing the calculation to each router, and then use routes (sequences in the lattice) to get to destinations that those routes advertise connectivity to. Following the SPF algorithm, calculation starts by selecting a starting point (typically the router doing the calculation), and successively adding {link, router} pairs until one has calculated a route to every router in the network. As each router is added, including the original router, destinations that it is directly connected to are turned into routes in the route table: "to get to 2001:db8::/32, route traffic to {interface, list of next hop routers}". For immediate neighbors to the originating router, of course, there is no next hop router; traffic is handled locally.

In this context, the route is qualified by a source prefix; It is installed into the FIB with the destination prefix, and the FIB applies the route if and only if the IPv6 source address also matches the advertised prefix. Of course, there may be multiple LSAs in the LSDB with the same destination and differing source prefixes; these may also have the same or differing next hop lists. The intended forwarding action is to forward matching traffic to one of the next hop routers associated with this destination and source prefix, or to discard non-matching traffic as "destination unreachable".

LSAs that lack a source prefix TLV match any source address (i.e., the source prefix TLV defaults to ::/0), by definition.

2.1. Notation

For the purposes of this document, a route from the prefix A to the prefix B (in other words, whose source prefix is A and whose destination prefix is B) is expressed as A->B. A packet with the source address A and the destination address B is similarly described as A->B.

2.2. Dealing with ambiguity

In any routing protocol, there is the possibility of ambiguity. For example, one router might advertise a fairly general prefix - a default route, a discard prefix (which consumes all traffic that is not directed to an instantiated subnet), or simply an aggregated prefix while another router advertises a more specific one. In source/destination routing, potentially ambiguous cases include cases in which the link state database contains two routes A->B' and A'->B, in which A' is a more specific prefix within the prefix A and B' is a more specific prefix within the prefix B. Traditionally, we have dealt with ambiguous destination routes using a "longest match first" rule. If the same datagram matches more than one destination prefix

advertised within an area, we follow the route with the longest matching prefix.

With source/destination routes, as noted in [I-D.baker-rtgwg-src-dst-routing-use-cases], we follow a similar but slightly different rule; the FIB lookup MUST yield the route with the longest matching destination prefix that also matches the source prefix constraint. In the event of a tie on the destination prefix, it MUST also match the longest matching source prefix among those options.

An example of the issue is this. Suppose we have two routes:

1. 2001:db8:1::/48 -> 2001:db8:3:3::/64
2. 2001:db8:2::/48 -> 2001:db8:3::/48

and a packet

2001:db8:2::1 -> 2001:db8:3:3::1

If we require the algorithm to follow the longest destination match without regard to the source, the destination address matches 2001:db8:3:3::/64 (the first route), and the source address doesn't match the constraint of the first route; we therefore have no route. The FIB algorithm, in this example, must therefore match the second route, even though it is not the longest destination match, because it also matches the source address.

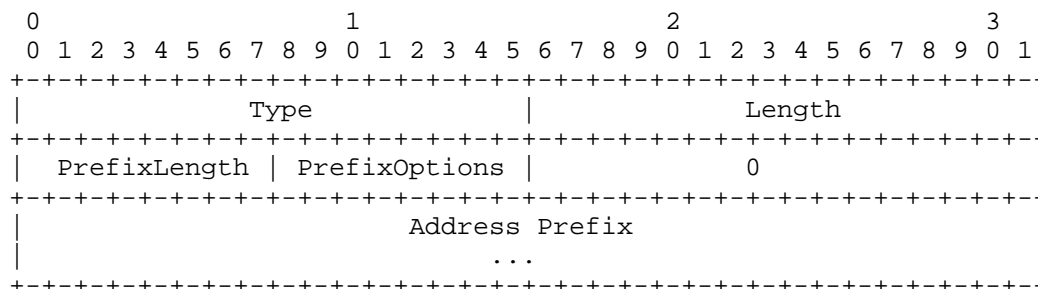
2.3. Interactions with other constraints

In the event that there are other constraints on routing, such as proposed in [I-D.baker-ipv6-ospf-dst-flowlabel-routing], the effect is a logical AND. The FIB lookup must yield the route with the longest matching destination prefix that also matches each of the constraints.

3. Extensions necessary for IPv6 Source/Destination Routing in OSPFv3

The extensible LSA format defined in [I-D.acee-ospfv3-lsa-extend] requires one additional option to accomplish source/destination routing: the source prefix. This is defined here. As noted in Section 2, any prefix LSA that does not specify a source prefix is understood to as specifying ::/0 as the source prefix.

3.1. IPv6 Source Prefix TLV



Source Prefix TLV

Type: assigned by IANA

TLV Length: Length of the value portion of the TLV in octets. This is by definition 20.

PrefixLength, PrefixOptions, and Address Prefix: Representation of an IPv6 address prefix, as described in [RFC5340] Appendix A.4.1

4. IANA Considerations

As discussed in [I-D.acee-ospfv3-lsa-extend], the IPv6 Source Prefix TLV code should be allocated from the OSPFv3 IANA registry.

5. Security Considerations

While source/destination routing could be used as part of a security solution, it is not really intended for the purpose. The approach limits routing, in the sense that it routes traffic to an appropriate egress, or gives a way to prevent communication between systems not included in a source/destination route, and in that sense could be considered similar to an access list that is managed by and scales with routing.

6. Acknowledgements

Acee Lindem contributed to the concepts in this draft.

7. References

7.1. Normative References

[I-D.acee-ospfv3-lsa-extend]
 Lindem, A., Mirtorabi, S., Roy, A., and F. Baker, "OSPFv3 LSA Extendibility", draft-acee-ospfv3-lsa-extend-00 (work in progress), May 2013.

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2460] Deering, S.E. and R.M. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.
- [RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.

7.2. Informative References

- [I-D.baker-ipv6-ospf-dst-flowlabel-routing]
Baker, F., "Using OSPFv3 with Role-Based Access Control", draft-baker-ipv6-ospf-dst-flowlabel-routing-02 (work in progress), May 2013.
- [I-D.baker-rtgwg-src-dst-routing-use-cases]
Baker, F., "Requirements and Use Cases for Source/Destination Routing", draft-baker-rtgwg-src-dst-routing-use-cases-00 (work in progress), August 2013.

Appendix A. Change Log

Initial Version: February 2013

First revision: April 2013

Correction: Corrected the reference to [I-D.acee-ospfv3-lsa-extend]

Use Case: Remove appendices, clarify the discussion of ambiguity and refer to [I-D.baker-rtgwg-src-dst-routing-use-cases]

Author's Address

Fred Baker
Cisco Systems
Santa Barbara, California 93117
USA

Email: fred@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 21, 2013

F.J. Baker
Cisco Systems
February 17, 2013

Extensible OSPF LSAs
draft-baker-ipv6-ospf-extensible-00

Abstract

This note describes the changes necessary for OSPFv3 to route extensible classes of traffic. This implies not routing "to a destination", but "traffic matching a classification tuple" which includes a destination but may also include other attributes such as the source address, DSCP, or Flow Label.

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 August 21, 2013.

Copyright Notice

Copyright (c) 2013 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	2
1.1. Requirements Language	3
2. Theory of Routing	3
2.1. Dealing with ambiguity	3
3. Extensions necessary for OSPFv3	4
3.1. OSPF optional data extensions	4
3.1.1. IPv6 Destination Prefix TLV	4
3.1.2. IPv6 Forwarding Address TLV	5
3.1.3. Referenced Advertising Router TLV	5
3.1.4. Metric TLV	6
3.1.5. External Route Tag TLV	6
3.1.6. Referenced Link State ID TLV	7
3.2. OSPF extensible LSAs	7
3.2.1. Extensible-Inter-area-prefix-LSA	8
3.2.2. Extensible-AS-external-LSA	9
3.2.3. Extensible-Intra-Area-Prefix-LSA	9
4. IANA Considerations	10
5. Security Considerations	10
6. Privacy Considerations	11
7. Acknowledgements	11
8. Change Log	11
9. References	11
9.1. Normative References	11
9.2. Informative References	11
Appendix A. FIB Design	11
A.1. Linux Source-Address Forwarding	12
A.1.1. One FIB per source prefix	12
A.1.2. One FIB per source prefix plus a general FIB	13
A.2. PATRICIA	13
A.2.1. Virtual Bit String	13
A.2.2. Tree Construction	14
A.2.3. Tree Lookup	15
Author's Address	15

1. Introduction

In related documents, the author proposes extensions to OSPF and IS-IS for the routing of IPv6 traffic using more than the destination address as the definition of a class of traffic to be routed. These include the possibility of source/destination routing, and especially egress routing, routing based on the destination plus the DSCP value such as is discussed in [RFC4915], and routing using the destination plus the IPv6 Flow Label for a form of Role Based Access Control - if the sender doesn't know the flow label value that the receiver is using, which it would learn from the network administrator through

configuration, DHCP, or some other means, it in effect has no route to the destination.

These capabilities, in OSPFv3, are have as a premise an extensible LSA; an LSA that contains the necessary elements of any LSA as discussed in section 4.4.1 of [RFC5340], a destination address, and a set of options. This document describes extensible inter-area-prefix-LSAs, intra-area-prefix-LSAs, and AS-external-LSAs. Additional options are defined in other documents.

Existing OSPF LSAs that specify only a destination prefix may be understood as identifying a destination prefix and "any" other option, whether it be source address, flow label, or something else. This is also a useful class of traffic to compactly represent, so existing LSA types are not deprecated, merely added to.

1.1. 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 [RFC2119].

2. Theory of Routing

Both IS-IS and OSPF perform their calculations by building a routes from the router performing the calculation to each router, and then use those routes to get to destinations that those routes advertise connectivity to. Following the SPF algorithm, calculation starts by selecting a starting point (typically the router doing the calculation), and successively adding {link, router} pairs until one has calculated a route to every router in the network. As each router is added, including the original router, destinations that it is directly connected to are turned into routes in the route table: "to get to 2001:db8::/32, route traffic to {interface, list of next hop routers}". For immediate neighbors to the originating router, of course, there is no next hop router; traffic is handled locally.

2.1. Dealing with ambiguity

In any routing protocol, there is the possibility of ambiguity. An area border router might, for example, summarize the routes to other areas into a small set of relatively short prefixes, which have more specific routes within the area. Traditionally, we have dealt with that using a "longest match first" rule. If the same datagram matches more than one destination prefix advertised within the area, we follow the route to the longest matching prefix.

When routing a class of traffic, we follow an analogous "most specific match" rule; we follow the route for the most specific matching tuple. In cases of simple overlap, such as routing to 2001:db8::/32 or 2001:db8:1::/48, that is exactly analogous; we choose one of the two routes.

It is possible, however, to construct an ambiguous case in which neither class subsumes the other. For example, presume that

- o A is a prefix,
- o B is a more-specific prefix within A,
- o C is a different prefix, and
- o D is a more-specific prefix of C.

The two classes {A, D, *, *} and {B, C, *, *} are ambiguous: a datagram within {B, D, *, *} matches both classes, and it is not clear in the data plane what decision to make. Solving this requires the addition of a third route in the FIB corresponding to the class {B, D, *, *}, which is more-specific than either of the first two, and can be given routing guidance based on metrics or other policy in the usual way.

3. Extensions necessary for OSPFv3

Changing OSPF to provide for this type of change requires cloning many of the existing LSAs: the inter-area-prefix-LSAs, the AS-external-LSAs, and the intra-area-prefix LSA. This can be done specifically with the information we have thought about, or designed for extensibility. We choose extensibility.

3.1. OSPF optional data extensions

This section defines a number of optional type-length-value (TLV) information elements that may be included in an extensible LSA. In an extensible LSA, elements not included are not considered in classification and as a result are in effect wild-carded.

3.1.1. IPv6 Destination Prefix TLV

The IPv6 Destination Prefix TLV MAY be used with the IPv6 Source Prefix TLV, but MUST NOT be used with the IPv4 Source Prefix TLV or the IPv4 Destination Prefix TLV.

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

```

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|      Type      |      Length      | Prefix Length  |      Prefix
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Destination Prefix TLV

Destination Prefix Type: assigned by IANA

TLV Length: Length of the TLV in octets

Prefix Length: Length of the prefix in bits, in the range 0..128

Prefix: (Destination prefix length + 7)/8 octets of prefix

3.1.2. IPv6 Forwarding Address TLV

The IPv6 Forwarding Address TLV is only used in the Extensible-AS-external-LSA, and is optional.

```

0                                     1                                     2                                     3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|      Type      |      Length      | 128 bit IPv6 Address
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

IPv6 Forwarding Address TLV

Flow Label Type: assigned by IANA

TLV Length: Length of the TLV in octets

IPv6 Address: A fully qualified IPv6 address (128 bits). If included, data traffic for the advertised destination will be forwarded to this address. It MUST NOT be set to the IPv6 Unspecified Address (0:0:0:0:0:0:0:0) or an IPv6 Link-Local Address (Prefix FE80/10). While OSPFv3 routes are normally installed with link-local addresses, an OSPFv3 implementation advertising a forwarding address MUST advertise a global IPv6 address. This global IPv6 address may be the next-hop gateway for an external prefix or may be obtained through some other method (e.g., configuration).

3.1.3. Referenced Advertising Router TLV

```

0                                     1                                     2                                     3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|      Type      |      Length      | Referenced Advertising Router
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

```

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Referenced Advertising Router TLV

Flow Label Type: assigned by IANA

TLV Length: Length of the TLV in octets

Referenced Link State ID: With the Referenced Link State ID TLV (Referenced LS Type and Referenced Link State ID), Identifies the router-LSA or network-LSA with which the IPv6 traffic classes should be associated. If Referenced LS Type is 0x2001, the prefixes are associated with a router-LSA, Referenced Link State ID should be 0, and Referenced Advertising Router should be the originating router's Router ID. If Referenced LS Type is 0x2002, the prefixes are associated with a network-LSA, Referenced Link State ID should be the Interface ID of the link's Designated Router, and Referenced Advertising Router should be the Designated Router's Router ID.

3.1.4. Metric TLV

```

0                                     1                                     2                                     3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|      Type      |      Length      |      Metric      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| PrefixOptions | Information elements for the traffic class
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Metric TLV

Flow Label Type: assigned by IANA

TLV Length: Length of the TLV in octets

Metric: The cost of this traffic class. Expressed in the same units as the interface costs in router-LSAs.

Information Elements This information element will be followed by zero or more information elements that describe the traffic class. the traffic class will have been fully described when parsing reaches the end of the LSA or finds a new Metric TLV.

3.1.5. External Route Tag TLV

The External Route Tag TLV is only used in the Extensible-AS-external-LSA, and is optional.

```

0                               1                               2                               3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|      Type      |      Length      |      External Route Tag      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

External Route Tag TLV

Flow Label Type: assigned by IANA

TLV Length: Length of the TLV in octets

Route Tag: A 32-bit field that MAY be used to communicate additional information between AS boundary routers.

3.1.6. Referenced Link State ID TLV

```

0                               1                               2                               3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|      Type      |      Length      |      Referenced LS Type      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Referenced Link State ID                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Referenced Link State ID TLV

Flow Label Type: assigned by IANA

TLV Length: Length of the TLV in octets

Referenced LS Type: The LSType of the associate LSA.

Referenced Link State ID: If included, additional information concerning the advertised external route can be found in the LSA having LS type equal to "Referenced LS Type", Link State ID equal to "Referenced Link State ID", and Advertising Router the same as that specified in the Extensible-AS-external-LSA's link-state header. This additional information is not used by the OSPF protocol itself. It may be used to communicate information between AS boundary routers. The precise nature of such information is outside the scope of this specification.

3.2. OSPF extensible LSAs

This section defines the extensible Extensible-Inter-Area-Prefix-LSA, Extensible-AS-external-LSA, and Extensible-Intra-Area-Prefix LSA.

3.2.1. Extensible-Inter-area-prefix-LSA

Extensible-Inter-area-prefix-LSAs have LS type equal to [IANA?]. These LSAs are equivalent to OSPFv2's type 3 summary-LSAs (see Section 12.4.3 of [RFC2328]). Originated by area border routers, they describe IPv4 or IPv6 traffic classes that belong to other areas, and are encoded using the TLVs defined in Section 3.1. A separate inter-area-prefix-LSA is originated for each such traffic class. For details concerning the construction of inter-area-prefix-LSAs, see [RFC5340] Section 4.4.3.4.

For stub areas, inter-area-prefix-LSAs can also be used to describe a (per-area) default route. Default summary routes are used in stub areas instead of flooding a complete set of external routes. When describing a default summary route, the Extensible-inter-area-prefix-LSA omits the Destination Prefix information element, which has the same effect as matching 0.0.0.0/0 or ::/0.

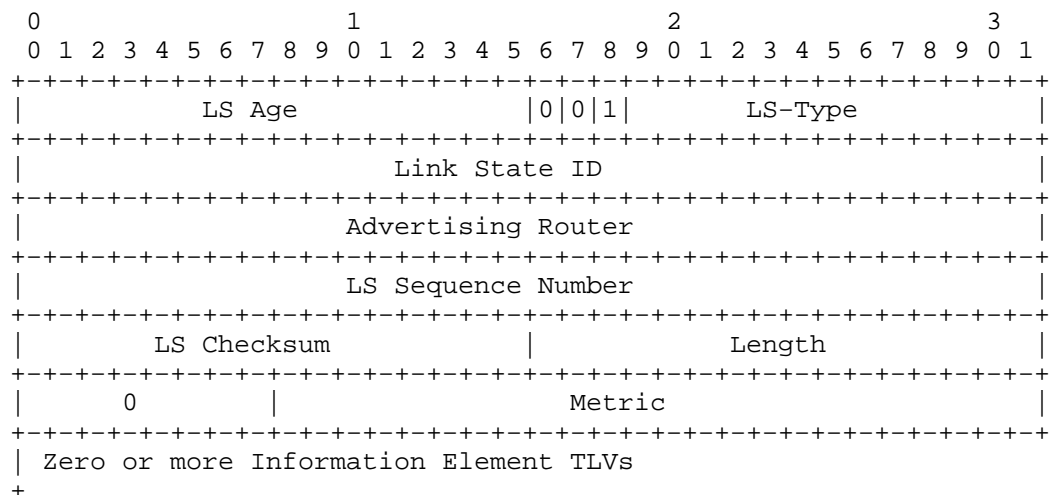


Figure 1: Extensible-Inter-area-prefix-LSA

LS-Type: To be assigned by IANA

Metric: The cost of this route. Expressed in the same units as the interface costs in router-LSAs. When the Extensible-inter-area-prefix-LSA is describing a route to a range of addresses (see [RFC5340] Appendix C.2), the cost is set to the maximum cost to any reachable component of the address range.

3.2.2. Extensible-AS-external-LSA

This is an AS-external-LSAs, but may include other information elements. Unlike the AS-external-LSAs, however, the presence of optional information is determined by the presence of the information elements, not by flags.

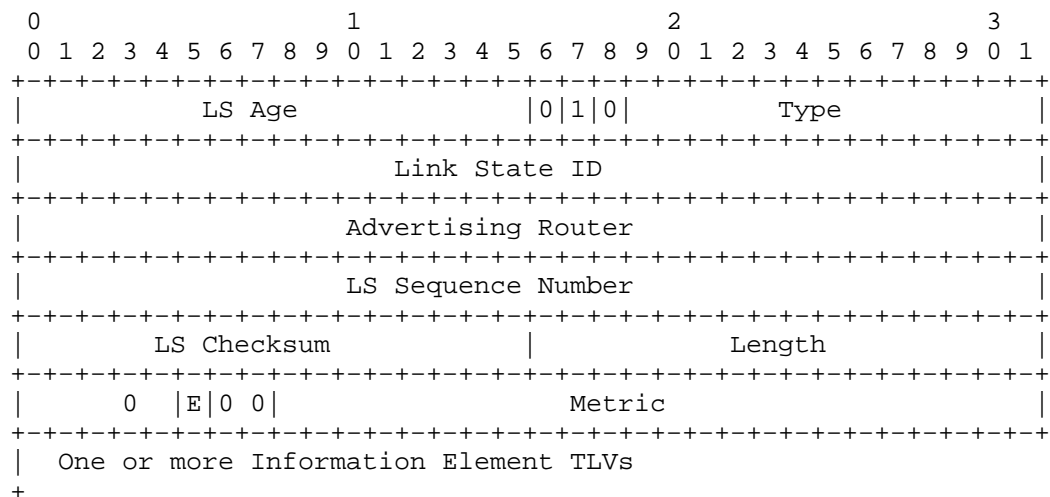


Figure 2: Extensible-AS-external-LSA

E: The type of external metric. If bit E is set, the metric specified is a Type 2 external metric. This means the metric is considered larger than any intra-AS path. If bit E is zero, the specified metric is a Type 1 external metric. This means that it is expressed in the same units as other LSAs (i.e., the same units as the interface costs in router-LSAs).

: The cost of this route. Interpretation depends on the external type indication (bit E above).

3.2.3. Extensible-Intra-Area-Prefix-LSA

This LSA MUST include a Referenced Link State ID TLV and a Referenced Advertising Router TLV immediately following the number of traffic classes. It MUST also include the indicated number of Metric TLVs, each of which is followed by the information elements that define that class of traffic, which will usually include a Destination Prefix TLV and may include a source prefix TLV, Flow Label TLV, or DSCP TLV.

Extensible-Intra-area-prefix-LSAs have LS types assigned by IANA. A router uses Extensible-intra-area-prefix-LSAs to advertise one or more traffic classes that are associated with a local router address, an attached stub network segment, or an attached transit network segment. In IPv4, the first two were accomplished via the router's router-LSA and the last via a network-LSA. In OSPF for IPv6, all addressing information that was advertised in router-LSAs and network-LSAs has been removed and is now advertised in intra-area-prefix-LSAs. For details concerning the construction of intra-area-prefix-LSA, see [RFC5340] Section 4.4.3.9.

A router can originate multiple extensible-intra-area-prefix-LSAs for each router or transit network. Each such LSA is distinguished by its unique Link State ID.

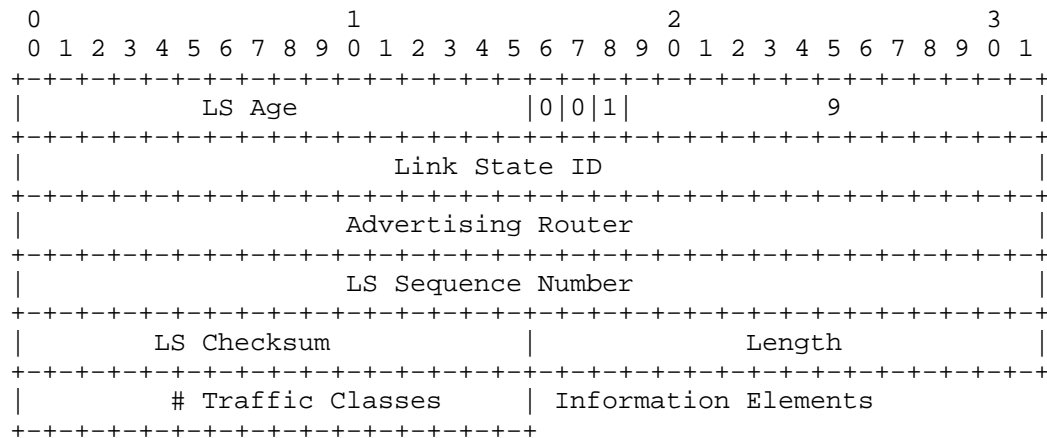


Figure 3: Extensible-Intra-Area-Prefix-LSA

Traffic Classes: The number of traffic classes that will be specified. Each traffic class has, first, a metric TLV, and then one or more other TLVs, normally including a Destination Prefix TLV.

4. IANA Considerations

This section will request LSID values for the LSAs defined, plus define a registry for optional fields. This is deferred to the -01 version of the draft.

5. Security Considerations

To be considered.

6. Privacy Considerations

To be considered.

7. Acknowledgements

8. Change Log

Initial Version: February 2013

9. References

9.1. Normative References

[ISO.10589.1992]
International Organization for Standardization,
"Intermediate system to intermediate system intra-domain-
routing routine information exchange protocol for use in
conjunction with the protocol for providing the
connectionless-mode Network Service (ISO 8473)", ISO
Standard 10589, 1992.

[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.

9.2. Informative References

[PATRICIA]
Morrison, D.R., "Practical Algorithm to Retrieve
Information Coded in Alphanumeric", Journal of the ACM
15(4) pp514-534, October 1968.

[RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P.
Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC
4915, June 2007.

Appendix A. FIB Design

While the design of the Forwarding Information Base is not a matter for standardization, as it only has to work correctly, not interoperate with something else, the design of a FIB for this type of lookup may differ from approaches used in destination routing. We describe one possible approach that is known to work, from the perspective of a proof of concept.

A.1. Linux Source-Address Forwarding

The University of Waikato has added to the Linux Advanced Routing & Traffic Control facility the ability to maintain multiple FIBs, one for each of a set of prefixes. Implementing source/destination routing using this mechanism is not difficult.

The router must know what source prefixes might be used in its domain. This may be by configuration or, at least in concept, learned from the routing protocols themselves. In whichever way that is done, one can imagine two fundamental FIB structures to serve N source prefixes; N FIBs, one per prefix, or N+1 FIBs, one per prefix plus one for destinations for which the source prefix is unspecified.

A.1.1. One FIB per source prefix

In an implementation with one FIB per source prefix, the routing algorithm has two possibilities.

- o If it calculates a route to a prefix (such as a default route) associated with a given source prefix, it stores the route in the FIB for the relevant source prefix.
- o If it calculates a route for which the source prefix is unspecified, it stores that route in all N FIBs.

When forwarding a datagram, the IP forwarder looks at the source address of the datagram to determine which FIB it should use. If it is from an address for which there is no FIB, the forwarder discards the datagram as containing a forged source address. If it is from an address within one of the relevant prefixes, it looks up the destination in the indicated FIB and forwards it in the usual way.

The argument for this approach is simplicity: there is one place to look in making a forwarding decision for any given datagram. The argument against it is memory space; it is likely that the FIBs will be similar, but every destination route not associated with a source prefix is duplicated in each FIB. In addition, since it automatically removes traffic whose source address is not among the configured list, it limits the possibility of user software using improper addresses.

A.1.2. One FIB per source prefix plus a general FIB

In an implementation with N+1 FIBs, the algorithm is slightly more complex.

- o If it calculates a route to a prefix (such as a default route) associated with a given source prefix, it stores the route in the FIB for the relevant source prefix.
- o If it calculates a route for which the source prefix is unspecified, it stores that route in the FIB that is not associated with a source prefix.

When forwarding a datagram, the IP forwarder looks at the source address of the datagram to determine which FIB it should use. If it is from one of the configured prefixes, it looks the destination up in the indicated FIB. In any event it also looks the destination up in the "unspecified source address" FIB. If the destination is found in only one of the two, the indicated route is followed. If the destination is found in both, the more specific route is followed.

The argument for this approach is memory space; if a large percentage of routes are only in the general FIB, such as when egress routing is used for the default route and all other routes are internal, the other FIBs are likely to be very small - perhaps only a single default route. The argument against this approach is complexity: most lookups if not all will be done in a prefix-specific FIB and in the general FIB.

A.2. PATRICIA

One approach is a [PATRICIA] Tree. This is a relative of a Trie, but unlike a Trie, need not use every bit in classification, and does not need the bits used to be contiguous. It depends on treating the bit string as a set of slices of some size, potentially of different sizes. Slice width is an implementation detail; since the algorithm is most easily described using a slice of a single bit, that will be presumed in this description.

A.2.1. Virtual Bit String

It is quite possible to view the fields in a datagram header incorporated into the classification tuple as a virtual bit string such as is shown in Figure 4. This bit string has various regions within it. Some vary and are therefore useful in a radix tree lookup. Some may be essentially constant - all global IPv6 addresses at this writing are within 2000::/3, for example, so while it must be tested to assure a match, incorporating it into the radix tree may

not be very helpful in classification. Others are ignored; if the destination is a remote /64, we really don't care what the EID is. In addition, due to variation in prefix length and other details, the widths of those fields vary among themselves. The algorithm the FIB implements, therefore, must efficiently deal with the fact of a discontinuous lookup key.

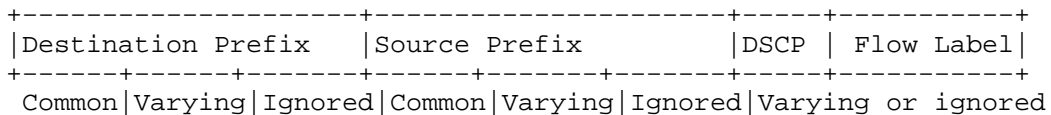


Figure 4: Treating a traffic class as a virtual bit string

A.2.2. Tree Construction

The tree is constructed by recursive slice-wise decomposition. At each stage, the input is a set of classes to be classified. At each stage, the result is the addition of a lookup node in the tree that identifies the location of its slice in the virtual bit string (which might be a bit number), the width of the slice to be inspected, and an enumerated set of results. Each result is a similar set of classes, and is analyzed in a similar manner.

The analysis is performed by enumerating which bits that have not already been considered are best suited to classification. For a slice of N bits, one wants to select a slide that most evenly divides the set of classes into 2^N subsets. If one or more bits in the slice is ignored in some of the classes, those classes must be included in every subset, as the actual classification of them will depend on other bits.

```

Input:{2001:db8::/32, ::/0, *, *}
      {2001:db8:1::/48, ::/0, AF41, *}
      {2001:db8:1::/48, ::/0, AF42, *}
      {2001:db8:1::/48, ::/0, AF43, *}
Common parts: Destination prefix 2001:dba, source prefix, and label
Varying parts: DSCP and the third set of sixteen bits in the
                destination prefix
One possible decomposition:
(1) slice = DSCP
    enumerated cases:
(a) { {2001:db8::/32, ::/0, *, *}, {2001:db8:1::/48, ::/0, AF41, *} }
(b) { {2001:db8::/32, ::/0, *, *}, {2001:db8:1::/48, ::/0, AF42, *} }
(c) { {2001:db8::/32, ::/0, *, *}, {2001:db8:1::/48, ::/0, AF43, *} }
(2) slice = third sixteen bit field in destination
    This divides each enumerated case into those containing 0001 and
    "everything else", which would imply 2001:db8::/32

```

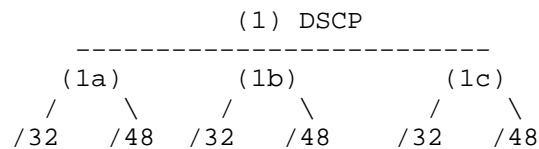


Figure 5: Example PATRICIA Tree

A.2.3. Tree Lookup

To look something up in a PATRICIA Tree, one starts at the root of the tree and performs the indicated comparisons recursively walking down the tree until one reaches a terminal node. When the enumerated subset is empty or contains only a single class, classification stops. Either classification has failed (there was no matching class, or one has presumably found the indicated class. At that point, every bit in the virtual bit string must be compared to the classifier; classification is accepted on a perfect match.

In the example in Figure 5, if a packet {2001:db8:1:2:3:4:5:6, 2001:db8:2:3:4:5:6:7, AF41, 0} arrives, we start at the root. Since it is an AF41 packet, we deduce that case (1a) applies, and since the destination has 0001 in the third sixteen bit field of the destination address, we are comparing to {2001:db8:1::/48, ::/0, AF41, *}. Since the destination address is within 2001:db8:1::/48, classification as that succeeds.

Author's Address

Fred Baker
 Cisco Systems
 Santa Barbara, California 93117
 USA

Email: fred@cisco.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: April 29, 2015

H. Chen
R. Li
A. Kumar S N
Huawei Technologies
G. Cauchie

A. Retana
Cisco Systems, Inc.
N. So
Tata Communications
V. Liu
China Mobile
M. Toy
Comcast
L. Liu
UC Davis
October 26, 2014

OSPF Topology-Transparent Zone
draft-chen-ospf-ttz-09.txt

Abstract

This document presents a topology-transparent zone in a domain. A topology-transparent zone comprises a group of routers and a number of links connecting these routers. Any router outside of the zone is not aware of the zone. The information about the links and routers inside the zone is not distributed to any router outside of the zone. Any link state change such as a link down inside the zone is not seen by any router outside of the zone.

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). 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 April 29, 2015.

Copyright Notice

Copyright (c) 2014 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	4
2. Conventions Used in This Document	5
3. Requirements	5
4. Topology-Transparent Zone	5
4.1. Overview of Topology-Transparent Zone	5
4.2. An Example of TTZ	6
5. Extensions to OSPF Protocols	7
5.1. Opaque LSAs for TTZ	7
5.2. A TTZ Capability TLV in Router Information LSA	10
6. Constructing LSAs for TTZ	11
7. Establishing Adjacencies	12
7.1. Discover TTZ Neighbor over Normal Adjacency	12
7.2. Establishing TTZ Adjacencies	12
7.3. Adjacency between TTZ Edge and Router outside	13
8. Distribution of LSAs	13
8.1. Distribution of LSAs within TTZ	14
8.2. Distribution of LSAs through TTZ	14
9. Computation of Routing Table	14
10. Operations	14
10.1. Configuring TTZ	14
10.2. Smooth Migration to TTZ	15
10.3. Adding a Router into TTZ	16
11. Prototype Implementation	16
11.1. What are Implemented and Tested	16
11.2. Implementation Experience	18
12. Security Considerations	18
13. IANA Considerations	19
14. Contributors	19
15. Acknowledgement	19
16. References	19
16.1. Normative References	19
16.2. Informative References	19
Authors' Addresses	20

1. Introduction

The number of routers in a network becomes larger and larger as the Internet traffic keeps growing. Through splitting the network into multiple areas, we can extend the network further. However, there are a number of issues when a network is split further into more areas.

At first, dividing a network from one area into multiple areas or from a number of existing areas to even more areas is a very challenging and time consuming task since it is involved in significant network architecture changes. Considering the one area case, originally the network has only one area, which is the backbone. This original backbone area will be split into a new backbone and a number of non backbone areas. In general, each of the non backbone areas is connected to the new backbone area through the area border routers between the non backbone and the backbone area. There is not any direct connection between any two non backbone areas. Each area border router summarizes the topology of its attached non backbone area for transmission on the backbone area, and hence to all other area border routers.

Secondly, the services carried by the network may be interrupted while the network is being split from one area into multiple areas or from a number of existing areas into even more areas.

Furthermore, it is complex for a Multi-Protocol Label Switching (MPLS) Traffic Engineering (TE) Label Switching Path (LSP) crossing multiple areas to be setup. In one option, a TE path crossing multiple areas is computed by using collaborating Path Computation Elements (PCEs) [RFC5441] through the PCE Communication Protocol (PCEP)[RFC5440], which is not easy to configure by operators since the manual configuration of the sequence of domains is required. Although this issue can be addressed by using the Hierarchical PCE, this solution may further increase the complexity of network design. Especially, the current PCE standard method may not guarantee that the path found is optimal.

This document presents a topology-transparent zone in an area and describes extensions to OSPF for supporting the topology-transparent zone, which is scalable and resolves the issues above.

A topology-transparent zone comprises a group of routers and a number of links connecting these routers. Any router outside of the zone is not aware of the zone. The information about the links and routers inside the zone is not distributed to any router outside of the zone. Any link state change such as a link down inside the zone is not seen by any router outside of the zone.

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.

3. Requirements

Topology-Transparent Zone (TTZ) may be deployed for resolving some critical issues in existing networks and future networks. The requirements for TTZ are listed as follows:

- o TTZ MUST be backward compatible. When a TTZ is deployed on a set of routers in a network, the routers outside of the TTZ in the network do not need to know or support TTZ.
- o TTZ MUST support at least one more levels of network hierarchies, in addition to the hierarchies supported by existing routing protocols.
- o Users SHOULD be able to easily set up an end to end service crossing TTZs.
- o The configuration for a TTZ in a network SHOULD be minimum.
- o The changes on the existing protocols for supporting TTZ SHOULD be minimum.

4. Topology-Transparent Zone

4.1. Overview of Topology-Transparent Zone

A Topology-Transparent Zone (TTZ) is identified by an Identifier (ID), and it includes a group of routers and a number of links connecting the routers. A TTZ is in an OSPF area.

The ID of a TTZ or TTZ ID is a number that is unique for identifying an entity such as a node in an OSPF domain. It is not zero in general.

In addition to having the functions of an OSPF area, an OSPF TTZ makes some improvements on an OSPF area, which include:

- o An OSPF TTZ is virtualized as the TTZ edge routers connected.

- o An OSPF TTZ receives the link state information about the topology outside of the TTZ, stores the information in the TTZ and floods the information through the TTZ to the routers outside of the TTZ.

4.2. An Example of TTZ

The figure below shows an area containing a TTZ: TTZ 600.

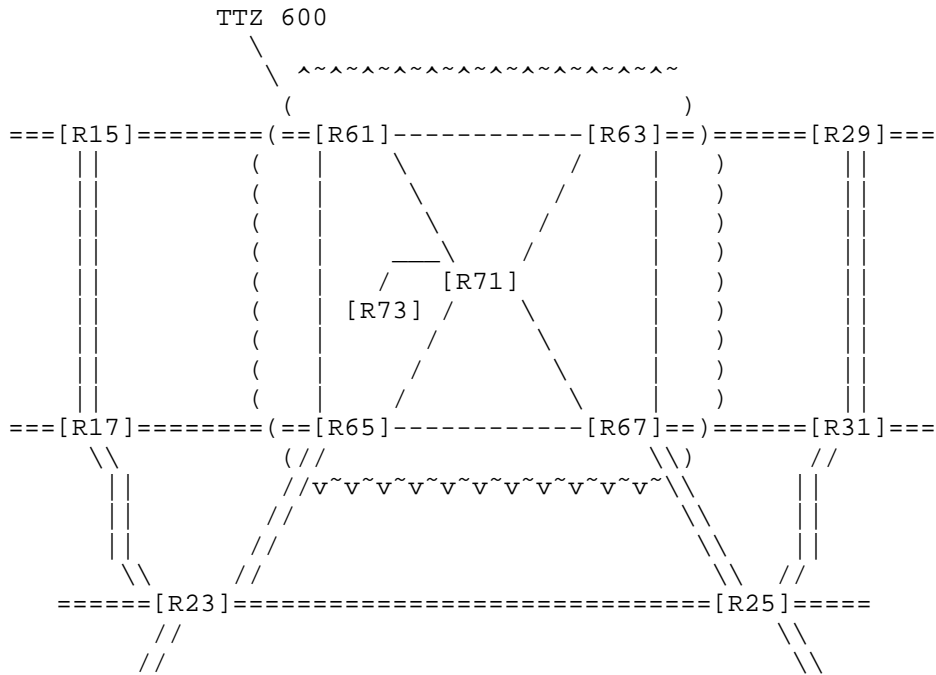


Figure 1: An Example of TTZ

The area comprises routers R15, R17, R23, R25, R29 and R31. It also contains TTZ 600, which comprises routers R61, R63, R65, R67, R71 and R73, and the links connecting them.

There are two types of routers in a TTZ: TTZ internal routers and TTZ edge routers. A TTZ internal router is a router inside the TTZ and its adjacent routers are in the TTZ. A TTZ edge router is a router inside the TTZ and has at least one adjacent router that is outside of the TTZ.

The TTZ in the figure above comprises four TTZ edge routers R61, R63, R65 and R67. Each TTZ edge router is connected to at least one router outside of the TTZ. For instance, router R61 is a TTZ edge

router since it is connected to router R15, which is outside of the TTZ.

In addition, the TTZ comprises two TTZ internal routers R71 and R73. A TTZ internal router is not connected to any router outside of the TTZ. For instance, router R71 is a TTZ internal router since it is not connected to any router outside of the TTZ. It is just connected to routers R61, R63, R65, R67 and R73 in the TTZ.

A TTZ MUST hide the information inside the TTZ from the outside. It MUST NOT directly distribute any internal information about the TTZ to a router outside of the TTZ.

For instance, the TTZ in the figure above MUST NOT send the information about TTZ internal router R71 to any router outside of the TTZ in the routing domain; it MUST NOT send the information about the link between TTZ router R61 and R65 to any router outside of the TTZ.

In order to create a TTZ, we MUST configure the same TTZ ID on the edge routers and identify the TTZ internal links on them. In addition, we SHOULD configure the TTZ ID on every TTZ internal router which indicates that every link of the router is a TTZ internal link.

From a router outside of the TTZ, a TTZ is seen as a group of routers fully connected. For instance, router R15 in the figure above, which is outside of TTZ 600, sees TTZ 600 as a group of TTZ edge routers: R61, R63, R65 and R67. These four TTZ edge routers are fully connected.

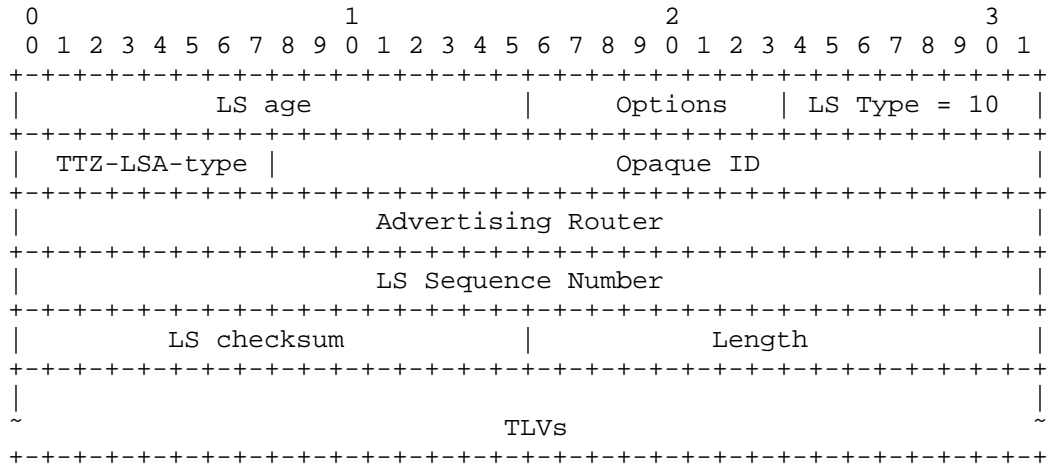
In addition, a router outside of the TTZ sees TTZ edge routers having normal connections to the routers outside of the TTZ. For example, router R15 sees four TTZ edge routers R61, R63, R65 and R67, which have the normal connections to R15, R29, R17 and R23, R25 and R31 respectively.

5. Extensions to OSPF Protocols

5.1. Opaque LSAs for TTZ

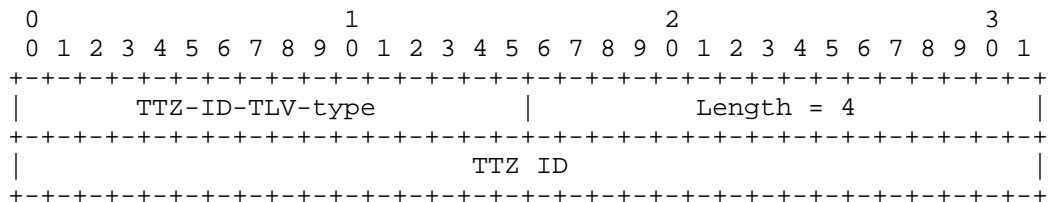
The link state information about a TTZ includes router LSAs and network LSAs describing the TTZ topology. These LSAs can be contained and distributed in opaque LSAs within the TTZ. Some control information on a TTZ can also be contained and distributed in opaque LSAs within the TTZ. These opaque LSAs are called TTZ opaque LSAs or TTZ LSAs for short.

The following is a general form of a TTZ LSA. It has an LS type = 10 and TTZ-LSA-Type, and contains a number of TLVs.

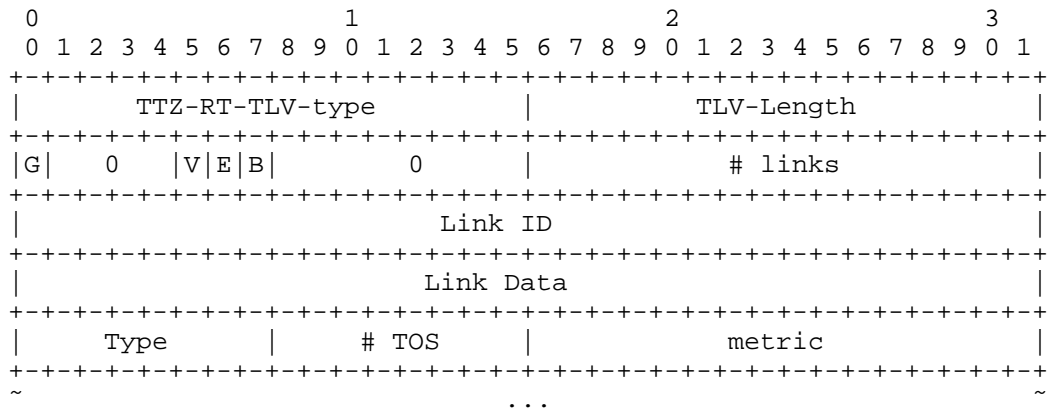


Where TTZ-LSA-type may be TBD1 (TTZ-RT-LSA-type) for TTZ Router LSA, TBD2 (TTZ-NW-LSA-type) for TTZ Network LSA, and TBD3 (TTZ-CT-LSA-type) for TTZ Control LSA.

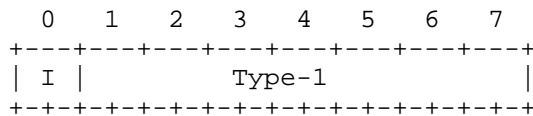
There are four types of TLVs: TTZ ID TLV, TTZ Router TLV, TTZ network TLV and TTZ Options TLV. A TTZ ID TLV has the following format. It contains a TTZ ID.



The format of a TTZ Router TLV is as follows. It contains the contents of a normal router LSA. A TTZ router LSA includes a TTZ ID TLV and a TTZ Router TLV.



Where G = 1/0 indicates that the router is an edge/internal router of TTZ. For a router link, the existing eight bit Type field for a router link may be split into two fields as follows:



I bit flag:

1: Router link is an internal link to a router inside TTZ.

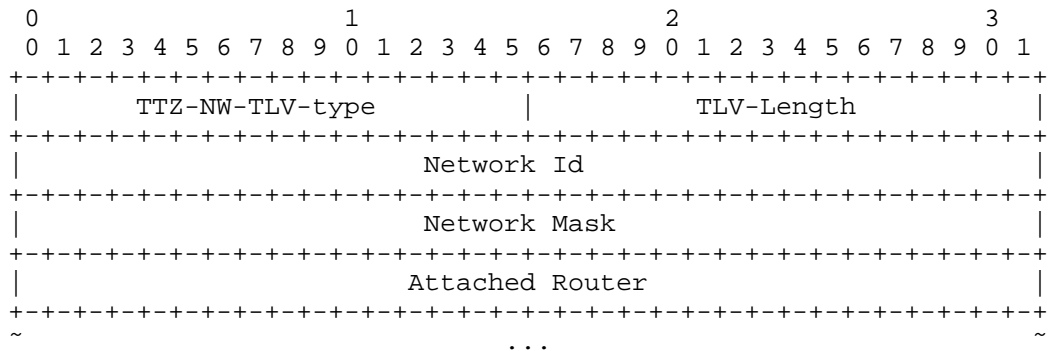
0: This indicates that the router link is an external link.

Type-1: The kind of the link.

For a link inside a TTZ, I bit flag is set to one, indicating that this link is an internal TTZ link. For a link connecting to a router outside of a TTZ from a TTZ edge router, I bit flag is set to zero, indicating that this link is an external TTZ link.

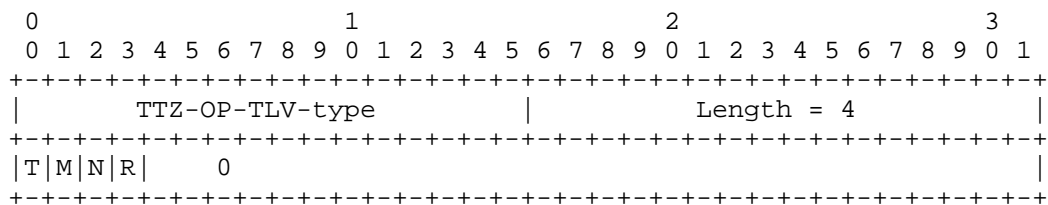
The value of Type-1 may be 1, 2, 3, or 4, which indicates that the kind of a link being described is a point-to-point connection to another router, a connection to a transit network, a connection to a stub network, or a virtual link respectively.

A TTZ Network TLV has the following format. It contains the contents of a normal network LSA. A TTZ network LSA includes a TTZ ID TLV and a TTZ network TLV.



Where Network ID is the interface address of the DR, which is followed by the contents of a network LSA.

The format of TTZ Options TLV is as follows. A TTZ control LSA contains a TTZ ID TLV and a TTZ Options TLV.



T = 1: Distributing TTZ Topology Information for Migration

M = 1: Migrating to TTZ

N = 1: Distributing Normal Topology Information for Rollback

R = 1: Rolling back from TTZ

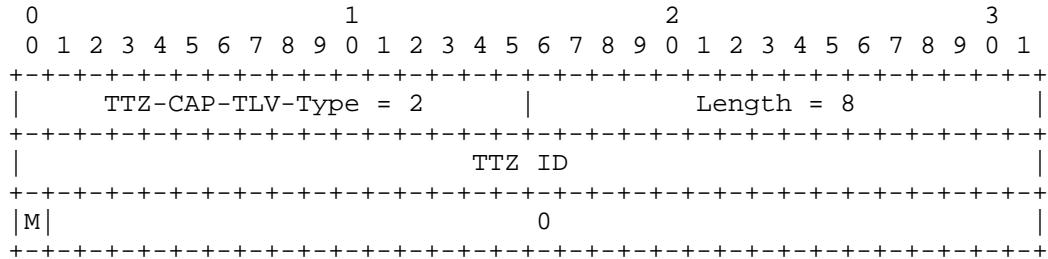
5.2. A TTZ Capability TLV in Router Information LSA

A new bit such as bit 6 for TTZ capability may be defined in the Router Informational Capabilities TLV as follows:

Bit	Capabilities
0	OSPF graceful restart capable [GRACE]
:	...
5	OSPF Experimental TE [EXP-TE]
6	OSPF TTZ capable [OSPF-TTZ]
7-31	Unassigned (Standards Action)

When the OSPF TTZ capable bit is set to one, a TTZ capability TLV must follow the Router Informational Capabilities TLV to indicate a link/router's TTZ capability and the TTZ to which the link/router

belongs. It has the following format.



It contains a TTZ ID and a number of TTZ bits. The following bits in the TLV are assigned:

Bit	Meaning
M	Have Migrated to TTZ (i.e., works as TTZ)
1-31	Unassigned (Standards Action)

A link scope RI LSA with a OSPF TTZ capable bit set to one and a TTZ Capability TLV will be used to discover a TTZ neighbor.

6. Constructing LSAs for TTZ

There are three types of LSAs for representing a TTZ: TTZ router LSA, TTZ network LSA and Router LSA for virtualizing TTZ. The first two may be generated by a TTZ router, and the third by a TTZ edge router.

A TTZ router LSA generated by a TTZ router has a TTZ ID TLV and a TTZ Router TLV. The former includes the ID of the TTZ to which the router belongs. The latter contains the links to the router.

A TTZ network LSA for a broadcast link is generated by the DR for the link. It contains a TTZ ID TLV and a TTZ network TLV. The former has the ID of the TTZ to which the link belongs. The latter includes the DR's address, the network mask, and the routers attached.

A router LSA for virtualizing a TTZ generated by an edge router of the TTZ comprises three groups of links in general.

The first group are the router links connecting the routers outside of the TTZ. These router links are normal router links. There is a router link for every adjacency between this TTZ edge router and a router outside of the TTZ.

The second group are the "virtual" router links. For each of the other TTZ edge routers, there is a point-to-point router link to it.

The cost of the link may be the cost of the shortest path from this TTZ edge router to it within the TTZ.

In addition, the LSA may contain a third group of links, which are stub links for other destinations inside the TTZ. They may be the loopback addresses to be accessed by a node outside of the TTZ.

7. Establishing Adjacencies

This section describes the adjacencies in some different cases.

7.1. Discover TTZ Neighbor over Normal Adjacency

For two routers A and B connected by a P2P link and having a normal adjacency, they discover TTZ each other through a link scope RI LSA with an OSPF TTZ capable bit and a TTZ ID. We call this LSA D-LSA for short. If two ends of the link have the same TTZ ID, A and B are TTZ neighbors. The following is a sequence of events related to TTZ.

```

      A                                     B
  Configure TTZ                         Configure TTZ
      D-LSA (TTZ-ID=100)
      -----> Same TTZ ID
                                     A is B's TTZ Neighbor
      D-LSA (TTZ-ID=100)
  Same TTZ ID <-----
  B is A's TTZ Neighbor

```

A sends B a D-LSA with TTZ-ID after the TTZ is configured on it. B sends A a D-LSA with TTZ-ID after the TTZ is configured on it. When A receives the D-LSA from B and determines they have the same TTZ ID, B is A's TTZ neighbor. When B receives the D-LSA from A and determines they have the same TTZ ID, A is B's TTZ neighbor.

For a number of routers connected through a broadcast link and having normal adjacencies among them, they also discover TTZ each other through D-LSAs. The DR for the link "forms" TTZ adjacency with each of the other routers if all the routers attached to the link have the same TTZ ID configured on the connections to the link.

7.2. Establishing TTZ Adjacencies

When a router (say A) is connected via a P2P link to another router (say B) and there is not any adjacency between them over the link, a user configures TTZ on two ends of the link to form a TTZ adjacency.

While A and B are forming an adjacency, they start to discover TTZ each other through D-LSAs in the same way as described above after the normal adjacency is greater than ExStart. When the normal adjacency is full and B becomes A's TTZ neighbor, A forms a TTZ adjacency with B. Similarly, B forms a TTZ adjacency with A.

For a number of routers connected through a broadcast link and having no adjacency among them, they start to form TTZ adjacencies after TTZ is configured on the link. While forming adjacencies, they discover TTZ each other through D-LSAs in the same way as described above after the normal adjacency is greater than ExStart. The DR for the link forms TTZ adjacency with each of the other routers if all the routers attached to the link have the same TTZ ID configured on the connections to the link. Otherwise, the DR does not form any adjacency with any router attached to the link.

An alternative way for forming an adjacency between two routers in a TTZ is to extend hello protocol. Hello protocol is extended to include TTZ ID in LLS of a hello packet. The procedure for handling hellos is changed to consider TTZ ID. If two routers have the same TTZ IDs in their hellos, an adjacency between these two routers is to be formed; otherwise, no adjacency is formed.

7.3. Adjacency between TTZ Edge and Router outside

For an edge router in a TTZ, it forms an adjacency with any router outside of the TTZ that has a connection with it.

When the edge router synchronizes its link state database with the router outside of the TTZ, it sends the router outside of the TTZ the information about all the LSAs except for the LSAs belonging to the TTZ that are hidden from any router outside of the TTZ.

At the end of the link state database synchronization, the edge router originates its own router LSA for virtualizing the TTZ and sends this LSA to the router outside of the TTZ.

From the point of view of the router outside of the TTZ, it sees the other end as a normal router and forms the adjacency in the same way as a normal router. It is not aware of anything about its neighboring TTZ. From the LSAs related to the TTZ edge router in the other end, it knows that the TTZ edge router is connected to each of the other TTZ edge routers and some routers outside of the TTZ.

8. Distribution of LSAs

LSAs can be divided into a couple of classes according to their

distributions. The first class of LSAs is distributed within a TTZ. The second is distributed through a TTZ.

8.1. Distribution of LSAs within TTZ

Any LSA about a link state in a TTZ is distributed within the TTZ. It is not distributed to any router outside of the TTZ. For example, a router LSA generated for a router in a TTZ is distributed within the TTZ and not distributed to any router outside of the TTZ.

Any network LSA generated for a broadcast or NBMA network in a TTZ is distributed in the TTZ and not sent to a router outside of the TTZ.

Any opaque LSA generated for a TTZ internal TE link is distributed within the TTZ and not distributed to any router outside of the TTZ.

8.2. Distribution of LSAs through TTZ

Any LSA about a link state outside of a TTZ received by an edge router of the TTZ is distributed through the TTZ. For example, when an edge router of a TTZ receives an LSA from a router outside of the TTZ, it floods it to its neighboring routers both inside the TTZ and outside of the TTZ. This LSA may be any LSA such as a router LSA that is distributed in a domain.

The routers in the TTZ continue to flood the LSA. When another edge router of the TTZ receives the LSA, it floods the LSA to its neighboring routers both outside of the TTZ and inside the TTZ.

9. Computation of Routing Table

The computation of the routing table on a router is the same as that described in RFC 2328, with one exception. A router in a TTZ MUST ignore the router LSAs generated by the edge routers of the TTZ for virtualizing the TTZ. It computes routes through using the TTZ topology represented by TTZ LSAs and the topology outside of the TTZ.

10. Operations

10.1. Configuring TTZ

This section proposes some options for configuring a TTZ.

1. Configuring TTZ on Every Link in TTZ

If every link in a TTZ is configured with a same TTZ ID as a TTZ

link, the TTZ is determined. A router with some TTZ links and some normal links is a TTZ edge router. A router with only TTZ links is a TTZ internal router.

2. Configuring TTZ on Every Router in TTZ

We may configure a same TTZ ID on every router in the TTZ, and on every edge router's links connecting to the routers in the TTZ.

A router configured with the TTZ ID on some of its links is a TTZ edge router. A router configured with the TTZ ID only is a TTZ internal router. All the links on a TTZ internal router are TTZ links. This option is simpler than the above one.

10.2. Smooth Migration to TTZ

For a group of routers and a number of links connecting the routers in an area, making them transfer to work as a TTZ without any service interruption may take a few of steps or stages.

At first, users configure the TTZ feature on every router in the TTZ. In this stage, a router does not originate its TTZ router LSA or TTZ network LSAs. It will discover its TTZ neighbors.

Secondly, after configuring the TTZ, users may issue a CLI command on one router in the TTZ, which triggers every router in the TTZ to generate and distribute TTZ information among the routers in the TTZ. When the router receives the command, it originates its TTZ router LSA and TTZ network LSAs as needed, and distributes them to its TTZ neighbors. It also originates a TTZ control LSA with T=1 (indicating TTZ information generation and distribution for migration). When a router in the TTZ receives the LSA with T=1, it originates its TTZ router LSA and TTZ network LSAs as needed. In this stage, every router in the TTZ has dual roles. One is to function as a normal router. The other is to generate and distribute TTZ information.

Thirdly, users SHOULD check whether every router in the TTZ is ready for transferring to work as a TTZ router. A router in the TTZ is ready after it has received all the necessary information from all the routers in the TTZ. This information may be displayed on a router through a CLI command.

And then users may activate the TTZ through using a CLI command such as migrate to TTZ on one router in the TTZ. The router transfers to work as a TTZ router, generates and distributes a TTZ control LSA with M=1 (indicating Migrating to TTZ) after it receives the command.

After a router in the TTZ receives the TTZ control LSA with M=1, it

also transfers to work as a TTZ router. Thus, activating the TTZ on one TTZ router makes every router in the TTZ transfer to work as a TTZ router, which flushes its normal router LSA and network LSAs, computes routes through using the TTZ topology represented by TTZ LSAs and the topology outside of the TTZ.

For an edge router of the TTZ, transferring to work as a TTZ router comprises generating a router LSA to virtualize the TTZ and flooding this LSA to all its neighboring routers.

10.3. Adding a Router into TTZ

When a non TTZ router (say R1) is connected via a P2P link to a TTZ router (say T1) working as TTZ and there is a normal adjacency between them over the link, a user can configure TTZ on two ends of the link to add R1 into the TTZ to which T1 belongs. They discover TTZ each other in the same way as described in section 7.1.

When a number of non TTZ routers are connected via a broadcast link to a TTZ router (say T1) working as TTZ and there are normal adjacencies among them, a user configures TTZ on the connection to the link on every router to add the non TTZ routers into the TTZ to which T1 belongs. The DR for the link "forms" TTZ adjacency with each of the other routers if all the routers have the same TTZ ID configured on the connections to the link.

When a router (say R1) is connected via a P2P link to a TTZ router (say T1) and there is not any adjacency between them over the link, a user can configure TTZ on two ends of the link to add R1 into the TTZ to which T1 belongs. R1 and T1 will form an adjacency in the same way as described in section 7.2.

When a router (say R1) is connected via a broadcast link to a group of TTZ routers on the link and there is not any adjacency between R1 and any over the link, a user can configure TTZ on the connection to the link on R1 to add R1 into the TTZ to which the TTZ routers belong. R1 starts to form an adjacency with the DR for the link after the configuration.

11. Prototype Implementation

11.1. What are Implemented and Tested

1. CLI Commands for TTZ

The CLIs implemented and tested include:

- o the CLIs of the simpler option for configuring TTZ, and
- o the CLIs for controlling migration to TTZ.

2. Extensions to OSPF Protocols for TTZ

All the extensions defined in section "Extensions to OSPF Protocols" are implemented and tested except for rolling back from TTZ. The testing results illustrate:

- o A TTZ is virtualized to outside as its edge routers fully connected. Any router outside of the TTZ sees the edge routers (as normal routers) connecting each other and to some other routers.
- o The link state information about the routers and links inside the TTZ is contained within the TTZ. It is not distributed to any router outside of the TTZ.
- o TTZ is transparent. From a router inside a TTZ, it sees the topology (link state) outside of the TTZ. From a router outside of the TTZ, it sees the topology beyond the TTZ. The link state information outside of the TTZ is distributed through the TTZ.
- o TTZ is backward compatible. Any router outside of a TTZ does not need to support or know TTZ.

3. Smooth Migration to TTZ

The procedures and related protocol extensions for smooth migration to TTZ are implemented and tested. The testing results show:

- o A part of an area is smoothly migrated to a TTZ without any routing disruptions. The routes on every router are stable while the part of the area is being migrated to the TTZ.
- o Migration to TTZ is very easy to operate.

4. Add a Router to TTZ

Adding a router into TTZ is implemented and tested. The testing results illustrate:

- o A router can be easily added into a TTZ and becomes a TTZ router.

- o The router added into the TTZ is not seen on any router outside of the TTZ, but it is a part of the TTZ.

5. Leak TTZ Loopbacks Outside

Leaking loopback addresses in a TTZ to routers outside of the TTZ is implemented and tested. The testing results illustrate:

- o The loopback addresses inside the TTZ are distributed to the routers outside of the TTZ.
- o The loopback addresses are accessible from a router outside of the TTZ.

11.2. Implementation Experience

The implementation of TTZ is relatively easy compared to other features of OSPF. Re-using the existing OSPF code along with additional simple logic does the work. A couple of engineers started to work on implementing the TTZ from the middle of June, 2014 and finished coding it just before IETF 90. After some testing and bug fixes, it works as expected.

In our implementation, the link state information in a TTZ opaque LSA is stored in the same link state database as the link state information in a normal LSA. For each TTZ link in the TTZ opaque LSA stored, there is an additional flag, which is used to differentiate between a TTZ link and a Normal link.

Before migration to TTZ, every router in the TTZ computes its routing table using the normal links. After migration to TTZ, every router in the TTZ computes its routing table using the TTZ links and normal links. In the case that there are one TTZ link and one normal link to select, the TTZ link is used. In SPF calculation, the back-link check passes if and only if the corresponding new additional bit matches. If link type bit is TTZ link, then the lookup is for corresponding TTZ LSA. In case of normal link, the lookup is based on normal link.

12. Security Considerations

The mechanism described in this document does not raise any new security issues for the OSPF protocols.

13. IANA Considerations

TBD

14. Contributors

Veerendranatha Reddy Vallem
Huawei Technologies
Bangalore
India
Email: veerendranatharv@huawei.com

15. Acknowledgement

The author would like to thank Acee Lindem, Abhay Roy, Dean Cheng, Russ White, William McCall, Tony Przygienda, Lin Han and Yang Yu for their valuable comments on this draft.

16. References

16.1. 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.
- [RFC4970] Lindem, A., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.
- [RFC2370] Coltun, R., "The OSPF Opaque LSA Option", RFC 2370, July 1998.
- [RFC5613] Zinin, A., Roy, A., Nguyen, L., Friedman, B., and D. Yeung, "OSPF Link-Local Signaling", RFC 5613, August 2009.

16.2. Informative References

- [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.

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: huaimo.chen@huawei.com

Renwei Li
Huawei Technologies
2330 Central expressway
Santa Clara, CA
USA

Email: renwei.li@huawei.com

Anil Kumar S N
Huawei Technologies
Bangalore
India

Email: anil.sn@huawei.com

Gregory Cauchie
FRANCE

Email: greg.cauchie@gmail.com

Alvaro Retana
Cisco Systems, Inc.
7025 Kit Creek Rd.
Raleigh, NC 27709
USA

Email: aretana@cisco.com

Ning So
Tata Communications
2613 Fairbourne Cir.
Plano, TX 75082
USA

Email: ningso01@gmail.com

Vic Liu
China Mobile
No.32 Xuanwumen West Street, Xicheng District
Beijing, 100053
China

Email: liuzhiheng@chinamobile.com

Mehmet Toy
Comcast
1800 Bishops Gate Blvd.
Mount Laurel, NJ 08054
USA

Email: mehmet_toy@cable.comcast.com

Lei Liu
UC Davis
CA
USA

Email: liulei.kddi@gmail.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: January 5, 2015

H. Chen
R. Li
Huawei Technologies
G. Cauchie

N. So
Tata Communications
V. Liu
China Mobile
L. Liu
UC Davis
July 4, 2014

Applicability of OSPF Topology-Transparent Zone
draft-chen-ospf-ttz-app-05.txt

Abstract

This document discusses the applicability of "OSPF Topology-Transparent Zone". It briefs the protocol and its operations first, and then illustrates the application scenarios of OSPF Topology-Transparent Zone. This document is intended for accompanying "OSPF Topology-Transparent Zone" to the Internet standards track.

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). 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 5, 2015.

Copyright Notice

Copyright (c) 2014 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. Overview of Topology-Transparent Zone	3
2.1. Definitions of Topology-Transparent Zone	3
2.2. An Example of TTZ	4
3. Applicability of Topology-Transparent Zone	6
3.1. One Area Network	6
3.1.1. Issues on Splitting Network into Areas	6
3.1.2. Use of TTZ in One Area Network	7
3.2. Multi-Area Network	9
3.2.1. Issues on Splitting Network into More Areas	9
3.2.2. Use of TTZ in Multi-Area Network	10
3.3. Use of TTZ on Routers in POP	11
3.4. Use of TTZ on Routers in IPRAN	11
3.5. Use of TTZ on Routers from Same Vendor	12
3.6. Use of TTZ on Routers in a Power Saving Group	12
4. Security Considerations	12
5. Contributors	12
6. Acknowledgement	13
7. References	13
7.1. Normative References	13
7.2. Informative References	13
Authors' Addresses	13

1. Introduction

The number of routers in a network becomes larger and larger as the Internet traffic keeps growing. Through splitting the network into multiple areas, we can extend the network further. However, there are a number of issues when a network is split further into more areas.

At first, dividing an AS or an area into multiple areas is a very challenging task since it is involved in significant network architecture changes.

Secondly, the services carried by the network may be interrupted while the network is being split from one area into multiple areas or from a number of existing areas into even more areas.

Moreover, it is complex for a Multi-Protocol Label Switching (MPLS) Traffic Engineering (TE) Label Switching Path (LSP) crossing multiple areas to be setup. In one option, a TE path crossing multiple areas is computed by using collaborating Path Computation Elements (PCEs) [RFC5441] through the PCE Communication Protocol (PCEP) [RFC5440], which is not easy to configure by operators since the manual configuration of the sequence of domains is required. Although this issue can be addressed by using the Hierarchical PCE, this solution may further increase the complexity of network design. Especially, the current PCE standard method may not guarantee that the path found is optimal.

This document introduces a technology called Topology-Transparent Zone (TTZ), presents a number of application scenarios of TTZ and illustrates that TTZ can resolve the issues above.

2. Overview of Topology-Transparent Zone

This section briefs the concept of Topology-Transparent Zone (TTZ) and explains the TTZ in some details through an example.

2.1. Definitions of Topology-Transparent Zone

A Topology-Transparent Zone (TTZ) comprises an Identifier (ID), a group of routers and a number of links connecting the routers. A Topology-Transparent Zone is in an OSPF area.

The ID of a Topology-Transparent Zone (TTZ) or TTZ ID for short is a number that is unique for identifying a node in an OSPF domain. It is not zero in general.

A TTZ may be virtualized as a different object in a different way. Two typical ways are given below.

In a first way, a TTZ may be virtualized as a group of TTZ edge routers fully connected. From a router outside of the TTZ, a TTZ is seen as a group of TTZ edge routers, which are fully connected.

In a second way, a TTZ may be seen as a single router. From a router outside of the TTZ, a TTZ is seen as a special single router. This router has a router ID, which is the TTZ ID or the maximum ID among all the router IDs of the routers in the TTZ. For every connection between a TTZ edge router and a router outside of TTZ, there is a connection between the special router and the router outside of TTZ.

The virtualization of TTZ in the first way is described and used below.

2.2. An Example of TTZ

The figure below illustrates an example of a routing area containing a topology-transparent zone: TTZ 600.

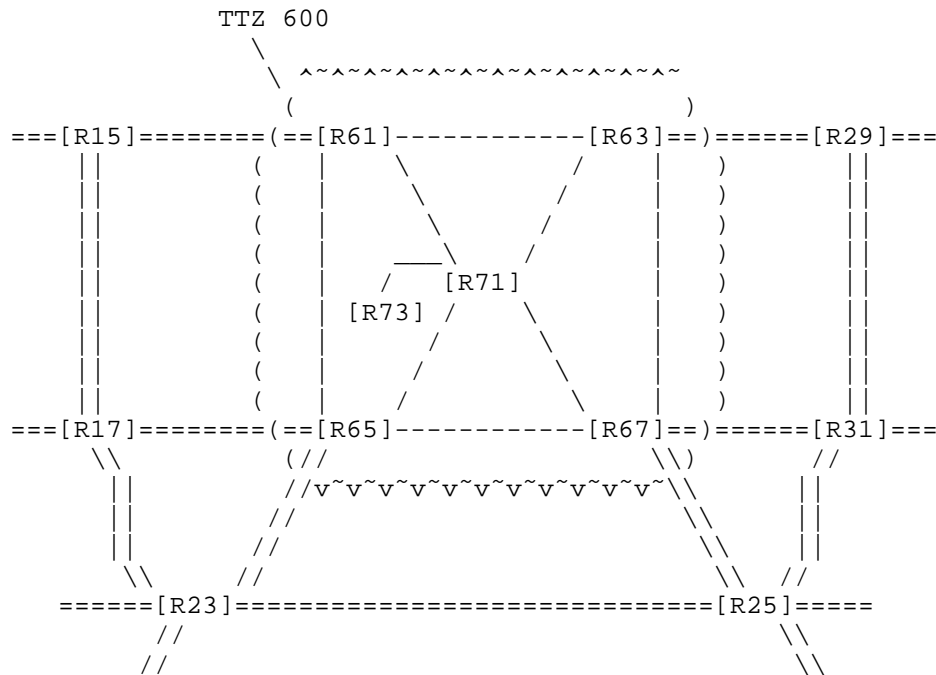


Figure 1: An Example of TTZ

The routing area comprises routers R15, R17, R23, R25, R29 and R31. It also contains a topology-transparent zone TTZ 600, which comprises routers R61, R63, R65, R67, R71 and R73, and the links connecting them.

There are two types of routers in a Topology-Transparent Zone (TTZ): TTZ internal routers and TTZ edge routers. A TTZ internal router is a router inside the TTZ and every adjacent router of the TTZ internal router is a router inside the TTZ. A TTZ edge router is a router inside the TTZ and has at least one adjacent router that is outside of the TTZ and at least one adjacent router that is inside the TTZ.

The TTZ in the figure above comprises four TTZ edge routers R61, R63, R65 and R67. Each TTZ edge router is connected to at least one router outside of the TTZ. For instance, router R61 is a TTZ edge router since it is connected to router R15, which is outside of the TTZ.

In addition, the TTZ comprises two TTZ internal routers R71 and R73. A TTZ internal router is not connected to any router outside of the TTZ. For instance, router R71 is a TTZ internal router since it is not connected to any router outside of the TTZ. It is just connected to routers R61, R63, R65, R67 and R73 inside the TTZ.

A TTZ may hide the information inside the TTZ from the outside. It may not distribute any internal information about the TTZ to a router outside of the TTZ.

For instance, the TTZ in the figure above does not send the information about TTZ internal router R71 to any router outside of the TTZ in the routing domain; it does not send the information about the link between R61 and R65 to any router outside of the TTZ.

In order to create a TTZ, we MUST configure the same TTZ ID on the edge routers and identify the TTZ internal links on them. In addition, we SHOULD configure the TTZ ID on every TTZ internal router which indicates that every link of the router is a TTZ internal link.

From a router outside of the TTZ, a TTZ is seen as a group of routers fully connected. For instance, router R15 in the figure above, which is outside of TTZ 600, sees TTZ 600 as a group of TTZ edge routers: R61, R63, R65 and R67. These four TTZ edge routers are fully connected.

In addition, a router outside of the TTZ sees TTZ edge routers having normal connections to the routers outside of the TTZ. For example, router R15 sees four TTZ edge routers R61, R63, R65 and R67, which have the normal connections to R15, R29, R17 and R23, R25 and R31

respectively.

3. Applicability of Topology-Transparent Zone

Topology-Transparent Zone (TTZ) may be used in different cases. This section presents a number of application scenarios of TTZ and illustrates the benefits that TTZ brings in each scenario.

3.1. One Area Network

Many networks start with one area. A network with only one area is easy to operate and maintain. As a network with one area becomes bigger and bigger because the increasing traffic in the network drives the expansion of the network, it needs to be split into multiple areas in general.

3.1.1. Issues on Splitting Network into Areas

Splitting a network with only one area into multiple areas is a very challenging task and may raise a number of issues.

1. Significant Changes on Network Architecture

There are significant changes on network architecture when splitting a network with one area into multiple areas. Originally the network has only one area, which is backbone area. This original backbone area will be split into a new backbone area and a number of non backbone areas.

In general, each of the non backbone areas is connected to the new backbone area through the area border routers between the non backbone area and the backbone area. There is not any direct connection between any two non backbone areas. Each area border router summarizes the topology of its attached non backbone area for transmission on the backbone area, and hence to all other area border routers.

Before splitting the network into areas, every router in the network has the information about the network topology. However, after splitting the network into areas, each router in an area has the information of the topology of the area, and it does not have the information of the topology of any other area. It has only the summary information about the other areas.

2. Service Interruptions

The services carried by the network may be interrupted while the

network is being split from one area into multiple areas.

3. Complex for MPLS TE Tunnel Setup

Each of the MPLS TE LSP tunnels originally in one area, which has its ingress and egress in different areas after the network splitting, needs to be re-configured and re-established. It is very complex for a MPLS TE LSP tunnel crossing areas to be set up.

In order to reduce the manual configurations for a MPLS TE LSP tunnel crossing multiple areas, we use PCEs to compute the path for the tunnel. Thus we must configure PCEs for the network split into multiple areas.

In addition, we need to provide a sequence of areas for the tunnel through manual configurations. The tunnel will go through the sequence of areas provided.

More critically, there are some issues on using PCEs. One of them is that the path computed by PCEs for the tunnel may not be optimal. If the optimal path for the tunnel is not in the sequence of areas configured by users, the path found by PCEs for the tunnel will not be optimal.

3.1.2. Use of TTZ in One Area Network

The issues mentioned above on splitting network into areas disappear if we do not split network into areas and use OSPF Topology Transparent Zone (TTZ) instead.

TTZ may be applied to a group of routers and links in the network directly. For a group of routers and links connecting the routers in the group in the network, no matter where it resides in the network, we may configure it as an OSPF TTZ as long as each router in the group can reach the other routers in the group through those links.

1. No Significant Changes on Network Architecture

There is not any significant changes on network architecture when an OSPF TTZ is applied to a group of routers and links in the network directly.

At first, we do not add any new connection to the network, or remove any existing connection from the network.

Secondly, every router outside of the TTZ is not aware of the TZZ. Even the router directly connecting to the TTZ is not aware of the TTZ.

Furthermore, every router in the network still has a topology view of the network. Except for those internal TTZ routers and links, which are hidden, every router outside of the TTZ has the link state information about all the routers and links in the network.

2. No Service Interruption

For a group of routers and a number of links connecting the routers in an area, they can transfer to work as a TTZ without any service interruptions.

There is not any route change while these routers are migrating to work as a TTZ. Every router in the TTZ "sees" the same network topology (the TTZ topology and the topology outside of the TTZ) and uses it to compute the routes. Thus the routing table on the router will not change.

For every router outside of the TTZ, its routing table will not change either while those routers are migrating to work as a TTZ. Even though there are some new router LSAs for virtualizing TTZ, these LSAs will not change any routes. Each link in any of these LSAs represents a shortest path between two TTZ edge routers within the TTZ.

3. Easy for MPLS TE Tunnel Setup

After a group of routers and links in the network is configured as an OSPF TTZ, a MPLS TE LSP tunnel with an ingress router and an egress router, which are anywhere in the network, can be configured in a way, which is the same as or similar to the way in which a MPLS TE LSP tunnel in one area network is configured.

For example, in the network in Figure 1 above, a MPLS TE LSP tunnel from ingress router R15 to egress router R29 can be configured in the same way as a MPLS TE LSP tunnel in one area network through provisioning the ingress router R15's IP address, the egress router R29's IP address and some constraints for the tunnel on the ingress router R15.

We do not need any PCEs for computing the constrained path for a MPLS TE LSP tunnel in the network with OSPF Topology Transparent Zones (TTZs). After a MPLS TE LSP tunnel with an ingress and egress anywhere in the network with OSPF TTZs is configured, the ingress computes the constrained path for the tunnel from the ingress to the egress in the same way as it computes the constrained path for the tunnel in one area network. The constrained path computed may go through some OSPF TTZs.

For example, in the network in Figure 1 above, the constrained path computed for the tunnel from ingress router R15 to egress router R29 may be from ingress router R15, to edge router R61 of TTZ 600, to edge router R63 of TTZ 600 and then to egress router R29.

As soon as the constrained path for a MPLS TE LSP tunnel is computed or given through configuration, the LSP can be established along the path by the signaling protocol RSVP-TE.

3.2. Multi-Area Network

For a network with multiple areas, it typically needs to be split into more areas when the size of the network becomes larger and larger as the traffic in the network keeps growing.

3.2.1. Issues on Splitting Network into More Areas

What would happen when we split a network with multiple areas into even more areas?

1. Significant Changes on Network Architecture

The changes on network architecture are significant when a network with multiple areas is split into even more areas. In the network before splitting, there is one backbone area, which is surrounded by a number of non backbone areas. Each of the non backbone areas is connected to the backbone area. There is not any direct connection between any two non backbone areas in general.

Splitting the network into more areas is involved in re-arranging a number of routers and links to create new areas and make some of the existing areas to become smaller. Some of the routers and links in a new area may be from the backbone area, and the other from some of the non backbone areas.

In the network after splitting, there is still one backbone area, which must be changed and be surrounded by the new non backbone areas and the existing non backbone areas some of which have been changed. Each of the non backbone areas is connected to the backbone area.

2. Service Interruptions

The services carried by the network may be interrupted while the network is being split from a number of existing areas into even more areas.

3. More Configurations for Tunnel Setup

For reducing the manual configurations for a MPLS TE LSP tunnel crossing multiple areas, we use PCEs to compute the path for the tunnel. Thus more configurations for tunnel setup is needed. We must configure PCEs for each of the new areas and peer relations among the PCEs for the new areas and the PCEs for the existing areas.

3.2.2. Use of TTZ in Multi-Area Network

The issues described above on splitting network into even more areas disappear if we do not split network into more areas and use OSPF TTZ instead.

A TTZ may be applied to a group of routers and links in any area in the network directly. For a group of routers and links connecting the routers in the group in an area, no matter where it resides in the area, we may configure it as an OSPF TTZ as long as each router in the group can reach the other routers in the group through those links.

1. No Significant Changes on Network Architecture

We can see that there is not any significant change on network architecture when an OSPF TTZ is applied to a group of routers and links in an area in the network directly.

At first, we do not add any new connection to the network, or remove any existing connection from the network.

Secondly, every router outside of the TTZ is not aware of the TTZ. Even the router directly connecting to the TTZ is not aware of the TTZ.

Furthermore, every router in the area still has a topology view of the area. Except for those internal TTZ routers and links, which are hidden, every router outside of the TTZ has the link state information about all the routers and links in the area.

2. No Service Interruption

For a group of routers and a number of links connecting the routers in an area, they can transfer to work as a TTZ without any service interruptions.

There is not any route change in the network while those routers are transferring to work as a TTZ.

3. No Extra Configurations for Tunnel Setup

After a group of routers and links in an area in the network is configured as an OSPF TTZ, there is not any extra configuration for supporting setup of a MPLS TE tunnel. We do not need to configure any new PCE since there is not any new area generated after applying a TTZ to a group of routers and links in an area.

3.3. Use of TTZ on Routers in POP

A Point of Presence (POP) comprises the routers in a room, a floor, a building or a group of buildings. These routers are normally in an AS or OSPF area.

We may increase the network scalability significantly through configuring a POP as an OSPF TTZ. When a POP becomes a TTZ, the link state information about every link and every router inside the POP is hidden from a router outside of the POP. Only very small amount of link state information virtualizing the TTZ for the POP is distributed to the router outside of the POP. Thus, the size of the LSDB on every router in the network is reduced significantly, and the speed for the router to compute the shortest path to every destination is increased dramatically.

We may also improve the network availability when we use a TTZ for a POP. In the case that a link or a router in the POP is down, the traffic may be interrupted without using any TTZ for the POP. The link state information about the link or router down needs to be distributed to every router in the network, and every router needs to compute the shortest path to every destination and download the path to its FIB. The traffic is forwarded according to the latest FIB on every router. During this process, there may be inconsistent views on the network topology on different routers.

The traffic interruption is minimized if we use a TTZ for the POP. The link state information about the link or router down is hidden from every router outside of the POP, which is not aware of the link or router down in the POP. Thus every router outside of the POP has the same topology view when the link or router is down. It does not compute the shortest path or download the path to its FIB.

3.4. Use of TTZ on Routers in IPRAN

The IP RAN provides connectivity for IP-based mobile broadband (MBB) from LTE and 4G base stations. The ratio of MBB subscribers to total mobile subscribers keeps growing fast. It is expected to grow to nearly 40% in 2016 from 15% in 2011.

At the end of 2012, China Mobile had deployed more than 500,000 nodes to support MBB services according to PTN Market Research 2013 Frost &

Sullivan. The size of the IP RAN network must seamlessly scale from tens of thousands to hundreds of thousands of nodes.

OSPF TTZ may be used in a big IP RAN network for reducing the partition of the network into more OSPF areas. Thus, the network can smoothly scale to hundreds of thousands of nodes. In addition, OSPF TTZ can make the operation and maintenance of the network easier.

3.5. Use of TTZ on Routers from Same Vendor

In a network, we may separate the routers from different vendors through using TTZ in order to alleviate the possible multi-vendor inter-operability issue. For example, the routers from a same vendor can be configured as a TTZ, and the routers outside of this TTZ are developed by different vendors.

3.6. Use of TTZ on Routers in a Power Saving Group

A power saving group is a set of routers and links, wherein the routers are connected through the links and there is a redundant route or path from a router in the group to another router in the group. The redundant path is within the group. That is that every hop in the redundant path is within the group.

In a power saving group, when the usage of a link within the group between two routers crosses a given threshold value for shutting down the link to save energy, the link will be shut down. This link down in the power saving group will not be distributed to any router outside of the group. The traffic outside of the group will not be affected by the link down inside the group.

From the characteristics of a power saving group, we can see that a power saving group is very suitable to be configured as a TTZ.

4. Security Considerations

This document does not introduce any new security issues.

5. Contributors

Yuanbin Yin
Huawei Technologies
Beijing,
China
Email: yinyuanbin@huawei.com

6. Acknowledgement

The authors would like to thank Alvaro Retana, Acee Lindem, and Dean Cheng for their valuable comments on this draft.

7. References

7.1. 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.
- [RFC2740] Coltun, R., Ferguson, D., and J. Moy, "OSPF for IPv6", RFC 2740, December 1999.

7.2. Informative References

- [RFC2370] Coltun, R., "The OSPF Opaque LSA Option", RFC 2370, July 1998.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC5786] Aggarwal, R. and K. Kompella, "Advertising a Router's Local Addresses in OSPF Traffic Engineering (TE) Extensions", RFC 5786, March 2010.
- [OSPF-TTZ] Chen, H., Li, R., Cauchie, G., So, N., and L. Liu, "OSPF Topology-Transparent Zone", draft-chen-ospf-ttz, Work in Progress, July 2012.

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: huaimo.chen@huawei.com

Renwei Li
Huawei Technologies
2330 Central expressway
Santa Clara, CA
USA

Email: renwei.li@huawei.com

Gregory Cauchie
FRANCE

Email: greg.cauchie@gmail.com

Ning So
Tata Communications
2613 Fairbourne Cir.
Plano, TX 75082
USA

Email: ningso01@gmail.com

Vic Liu
China Mobile
No.32 Xuanwumen West Street, Xicheng District
Beijing, 100053
China

Email: liuzhiheng@chinamobile.com

Lei Liu
UC Davis
USA

Email: liulei.kddi@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 22, 2015

M. Dubrovsky
R. Shrivastava
Cisco Systems
D. Cheng
Huawei Technologies
October 19, 2014

Extensions to OSPF facilitating the deployment of non-backward-
compatible changes.
draft-dubrovsky-ospf-non-compatible-02

Abstract

This document specifies a generic mechanism that facilitates the deployment of non-backward-compatible changes in OSPF protocol. This mechanism allows the OSPF routers to advertise the capability of non-backward-compatible functionality and to make the functionality operational only when supported by all participating routers. Depending on the functionality scope, capability advertisements must be propagated across a link, area or autonomous system (AS). For link and area scope functionality, Router Information Link State Advertisement (LSA) is utilized to propagate the capability information. For the cases when compatibility must be maintained across the whole OSPF autonomous system, new Area Information (AI) LSA is introduced. The AI LSA is a TLV-based analog of Indication-LSA that is used for demand circuit functionality and described in RFC1793.

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 April 22, 2015.

Copyright Notice

Copyright (c) 2014 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	4
2. Method to deploy non-backward-compatible changes	4
3. Area Information LSA	4
3.1. OSPFv2 Area Information (AI) Opaque LSA	5
3.2. OSPFv3 Area Information (AI) LSA	5
3.3. Area Information LSA TLV format	6
3.4. Area Information LSA origination	7
3.4.1. Limiting Area Information LSA origination	7
4. Capability negotiation before adjacency is fully formed	8
5. Backward Compatibility	8
6. IANA Considerations	8
7. Security Considerations	8
8. Acknowledgements	8
9. References	9
9.1. Normative References	9
9.2. Informative References	9
Authors' Addresses	9

1. Introduction

The evolution of OSPF protocol brought up changes that are not backward-compatible. Some of those changes (for example RFC1583Compatibility flag) can cause a routing loop in mixed environments. It therefore requires careful deployment planning, which is difficult to achieve in complex multivendor topologies. Most importantly, the lack of standard extendable mechanism that facilitates the deployment of non-backward-compatible changes obstructs the development of new protocol extensions.

As a solution for the above described problems, this document proposes an extendable mechanism, which guarantees that the non-backward-compatible functionality is turned on only when supported by all participating routers.

The proposed mechanism is not new; the existing demand circuit functionality [DEMAND] uses the same approach. This document simply makes the solution generic.

2. Method to deploy non-backward-compatible changes

Each participating router advertises the capability of functionality that it supports in the Router Information LSA as described in RFC 4970 [OSPF-CAP]. Routers only turn on a new functionality when it is supported by every router within the functionality scope. The routers revert back to their original behavior as soon as one incompatible device is detected.

The scope of functionality could be link, area or AS wide. For link and area wide, the router accordingly originates a link or area scope RI LSA. For AS functionality, an area scope RI LSA is used. To propagate compatibility information across area borders, a new LSA type Area Information is introduced.

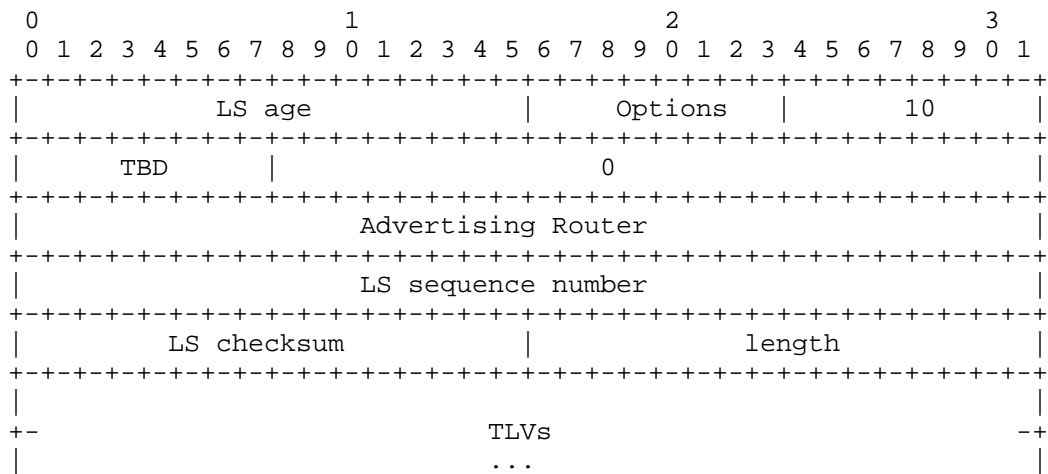
3. Area Information LSA

The Area Border Router inserts a particular capability TLV into an Area Information (AI) LSA to signal that at least one router in the attached areas does not support the functionality. Therefore, the presence of a particular TLV in AI LSA signals the opposite case to the presence of the same TLV in RI LSA. The AI LSA origination algorithm is very similar to the algorithm of Indication-LSA origination [DEMAND] and outlined below in Section 3.4. The AI LSA format is very similar to RI LSA [OSPF-CAP]. In OSPFv2, the AI LSA will be implemented with a new opaque LSA type ID. In OSPFv3, the AI

LSA will be implemented with a new LSA type function code. In both protocols, the AI LSA will have an area flooding scope. The exact format of AI LSA is outlined in the sections 3.1 and 3.2.

3.1. OSPFv2 Area Information (AI) Opaque LSA

OSPFv2 routers will advertise an area-scoped Opaque-LSA [OPAQUE]. The OSPFv2 Area Information LSA has a Link-State type of 10 indicating that the flooding scope is area-local, an Opaque type of TBD and Opaque ID of 0.



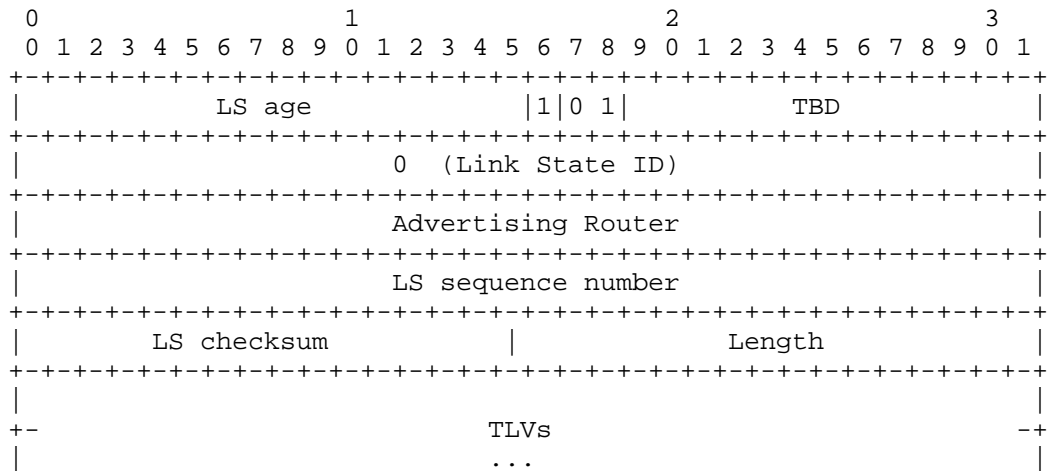
OSPFv2 Area Information Opaque LSA

The format of the TLVs within the body of an AI LSA is defined in Section 3.3.

3.2. OSPFv3 Area Information (AI) LSA

The OSPFv3 Area Information LSA has a function code of TBD while the S1/S2 bits are set to 1/0, indicating the area flooding scope for the LSA.

The U bit is set indicating that the OSPFv3 AI LSA should be flooded even if it is not understood. The Link State ID (LSID) value for this LSA is 0. This is unambiguous since an OSPFv3 router will only advertise a single AI LSA per flooding scope.

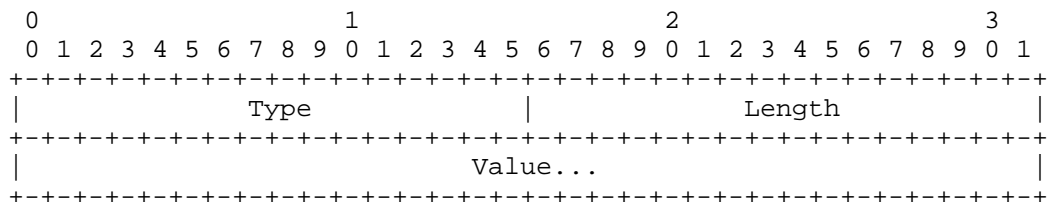


OSPFv3 Area Information LSA

The format of the TLVs within the body of an AI LSA is defined in Section 3.3.

3.3. Area Information LSA TLV format

The format of the TLVs within the body of an AI LSA is exactly the same as the corresponding RI LSA TLV format, which in turn is the same as the format used by the Traffic Engineering Extensions to OSPF [TE]. The LSA payload consists of one or more nested Type/Length/Value (TLV) triplets. The format of each TLV is:



TLV Format

The Length field defines the length of the value portion in octets (thus a TLV with no value portion would have a length of 0). The TLV is padded to a 4-octet alignment; padding is not included in the length field (so a 3-octet value would have a length of 3, but the total size of the TLV would be 8 octets). Nested TLVs are also 32-

bit aligned. For example, a 1-byte value would have the length field set to 1, and 3 octets of padding would be added to the end of the value portion of the TLV. Unrecognized types are ignored.

When new Area Information LSA TLV is defined, the specification **MUST** explicitly state whether the TLV is applicable to OSPFv2 only, OSPFv3 only, or both OSPFv2 and OSPFv3.

3.4. Area Information LSA origination

Through the origination of AI LSAs, information about the existence of incapable routers propagates from non-backbone areas, to the backbone area and from there to all other areas. The following two cases summarize the requirements for an area border router to originate AI LSAs:

1. Suppose an area border router (Router X) is connected to a non-backbone OSPF area (Area A). Furthermore, assume that Area A has an incapable router i.e. a router LSA without corresponding RI LSA TLV. Then Router X should originate the AI LSAs into all other directly connected areas, including the backbone area, in accordance with the guidelines of Section 3.4.1.

2. Suppose an area border router (Router X) is connected to the backbone OSPF area (Area 0.0.0.0). Furthermore, assume that the backbone has an indication of an existing incapable device via either

- a) the existence of a router LSA without corresponding RI LSA TLV

or

- b) AI LSAs that have been originated by routers other than Router X. Then Router X should originate AI LSAs into all other directly connected non-backbone areas, keeping the guidelines of Section 3.4.1 in mind.

3.4.1. Limiting Area Information LSA origination

The following guidelines should be observed by an area border router (Router X) when originating AI LSAs in order to limit their number. First, AI LSAs with corresponding TLV are not originated into an Area A when A has incapable routers; i.e. router LSAs without corresponding RI LSA TLV. Secondly, if another area border router has originated an AI LSA with corresponding TLV into Area A, and that area border router has a higher OSPF Router ID than Router X (same tie-breaker as for forwarding the address origination; see Section 12.4.4.1 of [OSPF]), then Router X should not originate an AI LSA with corresponding TLV into Area A.

4. Capability negotiation before adjacency is fully formed

For negotiating link scope capability before adjacency is fully formed, link local signaling [LLS] should be used instead of RI LSA. An example of such a functionality would be a modification to OSPF adjacency formation FSM.

5. Backward Compatibility

The mechanism is backward compatible with the existing OSPF specification. Setting the U bit in OSPFv3 AI LSA allows LSA propagation even if some routers in the area can not decode the LSA content. The Opaque LSA specification [OPAQUE] also guarantees the propagation of OSPFv2 AI LSA, even if the content is not understood by some of the transit routers.

6. IANA Considerations

The following IANA assignments are to be made from existing registries:

The OSPFv2 opaque LSA option type TBD will need to be reserved for the OSPFv2 AI opaque LSA via IETF Consensus.

OSPFv3 LSA Function Codes TBD will need to be reserved for the OSPFv3 Area Information (AI) LSA via Standards Action.

Both Standards Action and IETF Consensus registration procedures are described in the update of RFC 2434 [I-D.narten-iana-considerations-rfc2434bis].

7. Security Considerations

This document describes a generic mechanism for deployment of non-backward-compatible changes and it introduces Area-Information LSA for AS scope compatibility. The security considerations for those entities are as critical as the topology information currently advertised by the base OSPF protocol. Security considerations for the base OSPF protocol are covered in [OSPF] and [OSPFV3].

8. Acknowledgements

The author would like to acknowledge the helpful comments of Cisco OSPF Development team.

This memo is a product of the OSPF Working Group.

9. References

9.1. Normative References

- [LLS] Zinin, A., Roy, A., Nguyen, L., Friedman, B., and D. Yeung, "OSPF Link-Local Signaling", RFC 5613, August 2009.
- [OPAQUE] Berger, L., Bryskin, I., Zinin, A., and R. Coltun, "The OSPF Opaque LSA Option", RFC 5250, July 2008.
- [OSPF] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [OSPF-CAP] Lindem, A., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.
- [OSPFV3] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [TE] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.

9.2. Informative References

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

Authors' Addresses

Mike Dubrovsky
Cisco Systems
510 McCarthy Blvd.
Milpitas, CA 95035
USA

Email: mdubrovs@cisco.com

Rashmi Shrivastava
Cisco Systems
10 West Tasman Drive
San Jose, CA 95134
USA

Email: rashmi@cisco.com

Dean Cheng
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

Email: dean.cheng@huawei.com

Network Working Group
Internet Draft
Intended status: Proposed Standard
Expires: July 2015

S. Giacalone
Unaffiliated

D. Ward
Cisco Systems

J. Drake
Juniper Networks

A. Atlas
Juniper Networks

S. Previdi
Cisco Systems

January 09, 2015

OSPF Traffic Engineering (TE) Metric Extensions
draft-ietf-ospf-te-metric-extensions-11.txt

Abstract

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

This document describes common extensions to RFC 3630 "Traffic Engineering (TE) Extensions to OSPF Version 2" and RFC 5329 "Traffic Engineering Extensions to OSPF Version 3" to enable network performance information to be distributed in a scalable fashion. The information distributed using OSPF TE Metric Extensions can then be used to make path selection decisions based on network performance.

Note that this document only covers the mechanisms by which network performance information is distributed. The mechanisms for measuring network performance information or using 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 July 9, 2015.

Copyright Notice

Copyright (c) 2015 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.....	4
2. Conventions used in this document.....	5
3. TE Metric Extensions to OSPF TE.....	5
4. Sub-TLV Details.....	7
4.1. Unidirectional Link Delay Sub-TLV.....	7
4.1.1. Type.....	7
4.1.2. Length.....	7

4.1.3. A bit.....	7
4.1.4. Reserved.....	7
4.1.5. Delay Value.....	7
4.2. Min/Max Unidirectional Link Delay Sub-TLV.....	8
4.2.1. Type.....	8
4.2.2. Length.....	8
4.2.3. A bit.....	8
4.2.4. Reserved.....	8
4.2.5. Min Delay.....	9
4.2.6. Reserved.....	9
4.2.7 Max Delay.....	9
4.3. Unidirectional Delay Variation Sub-TLV.....	9
4.3.1. Type.....	10
4.3.2. Length.....	10
4.3.3. Reserved.....	10
4.3.4. Delay Variation.....	10
4.4. Unidirectional Link Loss Sub-TLV.....	10
4.4.1. Type.....	11
4.4.2. Length.....	11
4.4.3. A bit.....	11
4.4.4. Reserved.....	11
4.4.5. Link Loss.....	11
4.5. Unidirectional Residual Bandwidth Sub-TLV.....	11
4.5.1. Type.....	12
4.5.2. Length.....	12
4.5.3. Residual Bandwidth.....	12
4.6. Unidirectional Available Bandwidth Sub-TLV.....	12
4.6.1. Type.....	13
4.6.2. Length.....	13
4.6.3. Available Bandwidth.....	13
4.7. Unidirectional Utilized Bandwidth Sub-TLV.....	13
4.7.1. Type.....	14
4.7.2. Length.....	14
4.7.3. Utilized Bandwidth.....	14
5. Announcement Thresholds and Filters.....	14
6. Announcement Suppression.....	15
7. Network Stability and Announcement Periodicity.....	15
8. Enabling and Disabling Sub-TLVs.....	16
9. Static Metric Override.....	16
10. Compatibility.....	16
11. Security Considerations.....	16
12. IANA Considerations.....	17
13. References.....	17
13.1. Normative References.....	17
13.2. Informative References.....	18
14. Acknowledgments.....	19
15. Author's Addresses.....	19

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., link propagation delay) is becoming as critical to data path selection as other metrics.

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 network performance information (such as link propagation delay) in a cost-effective and scalable way.

This document describes extensions to OSPFv2 and OSPFv3 TE (hereafter called "OSPF TE Metric Extensions"), that can be used to distribute network performance information (viz link propagation delay, delay variation, link loss, residual bandwidth, available bandwidth, and utilized bandwidth).

The data distributed by OSPF TE Metric Extensions is meant to be used as part of the operation of the routing protocol (e.g., by replacing cost with link propagation delay or considering bandwidth as well as cost), by enhancing CSPF, or for use by a PCE [RFC4655] or an Alto server [RFC7285]. With respect to CSPF, the data distributed by OSPF TE Metric Extensions 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 distribute network performance information. The methods for measuring that information or acting on it once it is distributed are outside the scope of this document. A method for measuring loss and delay in an MPLS network is described in [RFC6374].

While this document does not specify the method for measuring network performance information, any measurement of link propagation delay SHOULD NOT vary significantly based upon the offered traffic load and hence SHOULD NOT include queuing delays. For a forwarding adjacency (FA) [RFC4206], care must be taken that measurement of the link propagation delay avoids significant queuing delay; this can be accomplished in a variety of ways, e.g., measuring with a traffic class that experiences minimal queuing or summing the measured link propagation delay of the links on the FA's path.

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. TE Metric Extensions to OSPF TE

This document defines new OSPF TE sub-TLVs that are used to distribute network performance information. The extensions in this document build on the ones provided in OSPFv2 TE [RFC3630] and OSPFv3 TE [RFC5329].

OSPFv2 TE LSAs [RFC3630] are opaque LSAs [RFC5250] with area flooding scope while OSPFv3 Intra-Area-TE-LSAs have their own LSA type, also with area flooding scope; both consist of a single TLV with one or more nested sub-TLVs. The Link TLV is common to both and describes the characteristics of a link between OSPF neighbors.

This document defines several additional sub-TLVs for the Link TLV:

Type	Length	Value
TBD1	4	Unidirectional Link Delay
TBD2	8	Min/Max Unidirectional Link Delay
TBD3	4	Unidirectional Delay Variation
TBD4	4	Unidirectional Link Loss
TBD5	4	Unidirectional Residual Bandwidth
TBD6	4	Unidirectional Available Bandwidth
TBD7	4	Unidirectional Utilized Bandwidth

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 move 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) The node determines 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 failback) it wishes (including nothing).

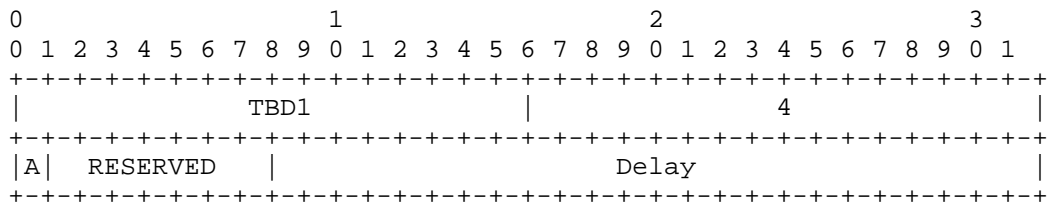
The A bit was intentionally omitted from some sub-TLVs to help mitigate oscillations. See section 7. 1. for more information.

Link delay, delay variation, and link loss MUST be encoded as integers. Consistent with existing OSPF TE specifications [RFC3630], residual, available, and utilized bandwidth MUST be encoded in IEEE single precision floating point [IEEE754]. Link delay and delay variation MUST be in units of microseconds, link loss MUST be a percentage, and bandwidth MUST be in units of 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 advertising node to its neighbor (i.e., the forward path delay). 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.

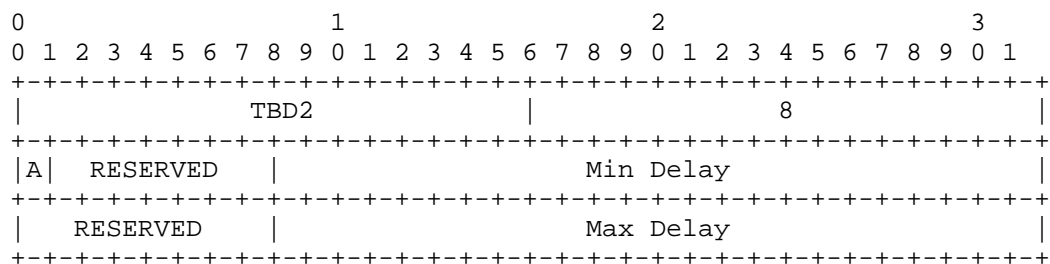
4.1.5. Delay Value

This 24-bit field carries the average link delay over a configurable interval in micro-seconds, encoded as an integer value. When set to the maximum value 16,777,215 (16.777215 sec), then the delay is at least that value and may be larger. If there is no value to send

(unmeasured and not statically specified), then the sub-TLV should not be sent or be withdrawn.

4.2. Min/Max Unidirectional Link Delay Sub-TLV

This sub-TLV advertises the minimum and maximum delay values between two directly connected OSPF neighbors. The delay advertised by this sub-TLV MUST be the delay from the advertising node to its neighbor (i.e., the forward path delay). 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 8.

4.2.3. A bit

This field represents the Anomalous (A) bit. The A bit is set when one or more measured values exceed a 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.

4.2.4. Reserved

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


```

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

4.3.1. Type

This sub-TLV has a type of TBD3.

4.3.2. Length

The length is 4.

4.3.3. Reserved

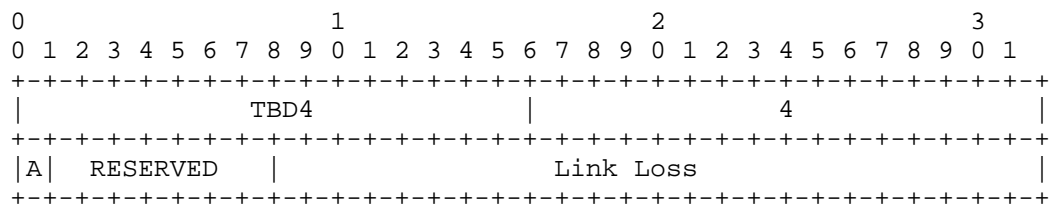
This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

4.3.4. 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.

4.4. 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 advertising node to its neighbor (i.e., the forward path loss). The format of this sub-TLV is shown in the following diagram:




```

|-----Residual Bandwidth-----|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

4.5.1. Type

This sub-TLV has a type of TBD5.

4.5.2. Length

The length is 4.

4.5.3. 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.

The calculation of Residual Bandwidth is different than that of Unreserved Bandwidth [RFC3630]. Residual Bandwidth subtracts tunnel reservations from Maximum Bandwidth (i.e., the link capacity) [RFC3630] and provides an aggregated remainder across QoS classes. Unreserved Bandwidth [RFC3630], on the other hand, is subtracted from the Maximum Reservable Bandwidth (the bandwidth that can theoretically be reserved) [RFC3630] and provides per-QoS-class remainders. Residual Bandwidth and Unreserved Bandwidth [RFC3630] can be used concurrently, and each has a separate use case (e.g., the former can be used for applications like Weighted ECMP while the latter can be used for call admission control).

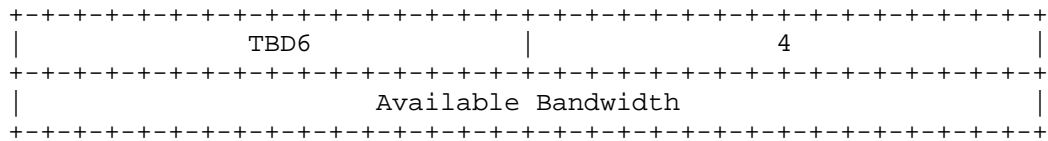
4.6. Unidirectional Available Bandwidth Sub-TLV

This TLV advertises the available bandwidth between two directly connected OSPF neighbors. The available bandwidth advertised by this sub-TLV MUST be the available bandwidth from the advertising node to its neighbor. The format of this sub-TLV is shown in the following diagram:

```

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

```



4.6.1. Type

This sub-TLV has a type of TBD6.

4.6.2. Length

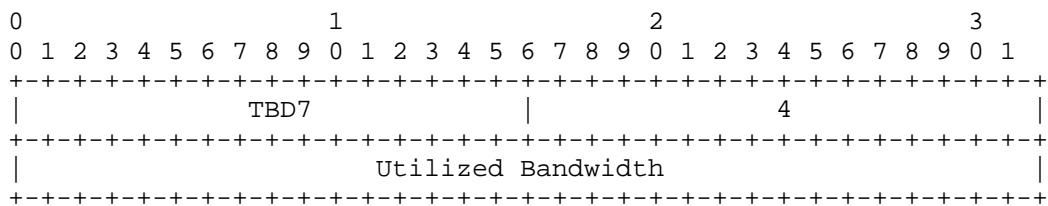
The length is 4.

4.6.3. 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 4.5.) 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.

4.7. Unidirectional Utilized Bandwidth Sub-TLV

This Sub-TLV advertises the bandwidth utilization between two directly connected OSPF neighbors. The bandwidth utilization advertised by this sub-TLV MUST be the bandwidth from the advertising node to its neighbor. The format of this Sub-TLV is shown in the following diagram:



4.7.1. Type

This sub-TLV has a type of TBD7.

4.7.2. Length

The length is 4.

4.7.3. Utilized Bandwidth

This field carries the bandwidth utilization 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, bandwidth utilization represents the actual utilization of the link (i.e., as measured by the advertising node). For a bundled link, bandwidth utilization is defined to be the sum of the component link bandwidth utilizations.

5. Announcement Thresholds and Filters

The values advertised in all sub-TLVs (except min/max delay and residual bandwidth) MUST represent an average over a period or be obtained by a filter that is reasonably representative of an average. For example, a rolling average is one such filter.

Min and max delay MAY be the lowest and/or highest measured value over a measurement interval or MAY make use of a filter, or other technique to obtain a reasonable representation of a min and max value representative of the interval with compensation for outliers.

The measurement interval, any filter coefficients, and any advertisement intervals MUST be configurable for each sub-TLV.

In addition to the measurement intervals governing re-advertisement, implementations SHOULD provide for each sub-TLV configurable accelerated advertisement thresholds, such that:

1. If the measured parameter falls outside a configured upper bound for all but the min delay metric (or lower bound for min delay metric only) and the advertised sub-TLV is not already outside that bound or,
2. If the difference between the last advertised value and current measured value exceed a configured threshold then,

3. The advertisement is made immediately.
4. For sub-TLVs which include an A-bit (except min/max delay), an additional threshold SHOULD be included corresponding to the threshold for which the performance is considered anomalous (and sub-TLVs with the A bit are sent). The A-bit is cleared when the sub-TLV's performance has been below (or re-crosses) this threshold for an advertisement interval(s) to permit fail back.

To prevent oscillations, only the high threshold or the low threshold (but not both) may be used to trigger any given sub-TLV that supports both.

Additionally, once outside of the bounds of the threshold, any re-advertisement of a measurement within the bounds would remain governed solely by the measurement interval for that sub-TLV.

6. Announcement Suppression

When link performance values change by small amounts that fall under thresholds that would cause the announcement of a sub-TLV, implementations SHOULD suppress sub-TLV re-advertisement and/or lengthen the period within which they are refreshed.

Only the accelerated advertisement threshold mechanism described in section 5 may shorten the re-advertisement interval.

All suppression and re-advertisement interval back-off timer features SHOULD be configurable.

7. Network Stability and Announcement Periodicity

Sections 5 and 6 provide configurable mechanisms to bound the number of re-advertisements. Instability might occur in very large networks if measurement intervals are set low enough to overwhelm the processing of flooded information at some of the routers in the topology. Therefore care should be taken in setting these values.

Additionally, the default measurement interval for all sub-TLVs should be 30 seconds.

Announcements must also be able to be throttled using configurable inter-update throttle timers. The minimum announcement periodicity is

1 announcement per second. The default value should be set to 120 seconds.

Implementations should not permit the inter-update timer to be lower than the measurement interval.

Furthermore, it is recommended that any underlying performance measurement mechanisms not include any significant buffer delay, any significant buffer induced delay variation, or any significant loss due to buffer overflow or due to active queue management.

8. Enabling and Disabling Sub-TLVs

Implementations MUST make it possible to individually enable or disable the advertisement of each sub-TLV.

9. Static Metric Override

Implementations SHOULD permit the static configuration and/or manual override of dynamic measurements for each sub-TLV in order to simplify migration and to mitigate scenarios where dynamic measurements are not possible.

10. Compatibility

As per [RFC3630], an unrecognized TLV should be silently ignored. I.e., it should not be processed but it should be included in LSAs sent to OSPF neighbors.

11. Security Considerations

This document does not introduce security issues beyond those discussed in [RFC3630]. OSPFv2 HMAC-SHA [RFC5709] provides additional protection for OSPFv2. OSPFv3 IPsec [RFC4552] and OSPFv3 Authentication Trailer [RFC7166] provide additional protection for OSPFv3.

OSPF KARP [RFC6863] provides an analysis of OSPFv2 and OSPFv3 routing security and OSPFv2 Security Extensions [OSPFSEC] provides extensions designed to address the identified gaps in OSPFv2.

12. IANA Considerations

IANA maintains the registry for the Link TLV sub-TLVs. OSPF TE Metric Extensions will require one new type code for each sub-TLV defined in this document, as follows:

Type	Description
------	-------------

TBD1	Unidirectional Link Delay
------	---------------------------

TBD2	Min/Max Unidirectional Link Delay
------	-----------------------------------

TBD3	Unidirectional Delay Variation
------	--------------------------------

TBD4	Unidirectional Link Loss
------	--------------------------

TBD5	Unidirectional Residual Bandwidth
------	-----------------------------------

TBD6	Unidirectional Available Bandwidth
------	------------------------------------

TBD7	Unidirectional Utilized Bandwidth
------	-----------------------------------

13. References

13.1. 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.
- [RFC5329] Ishiguro, K., Manral, V., Davey, A., Lindem, A., "Traffic Engineering Extensions to OSPF Version 3", RFC 5329, September 2009.

- [IEEE754] Institute of Electrical and Electronics Engineers,
"Standard for Floating-Point Arithmetic", IEEE Standard
754, August 2008.

13.2. Informative References

- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., Swallow, G., "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4206] Kompella, K., Rekhter, Y., "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC4552] Gupta, M., Melam, M., "Authentication/Confidentiality for OSPFv3", RFC 4552, June 2006.
- [RFC4655] Farrel, A., Vasseur, J.-P., Ash, J., "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC5250] Berger, L., Bryskin I., Zinin, A., Coltun, R., "The OSPF Opaque LSA Option", RFC 5250, July 2008.
- [RFC5709] Bhatia, M., Manral, V., Fanto, M., White, R., Barnes, M., Li, T., Atkinson, R., "OSPFv2 HMAC-SHA Cryptographic Authentication", RFC 5709, October 2009.
- [RFC6374] Frost, D., Bryant, S., "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.
- [RFC6863] Hartman, S., Zhang, D., "Analysis of OSPF Security According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6863, March 2013.
- [RFC7166] Bhatia, M., Manral, V., Lindem, A., "Supporting Authentication Trailer for OSPFv3", RFC 7166, March 2014.
- [RFC7285] Almi, R., Penno, R., Yang, Y., Kiesel, S., Previdi, S., Roome, W., Shalunov, S., Woundy, R., "Application-Layer Traffic Optimization (ALTO) Protocol", RFC 7285, September 2014.
- [OSPFSEC] Bhatia, M., Hartman, S., Zhang, D., Lindem, A., "Security Extensions for OSPFv2 when using Manual Key Management",

draft-ietf-ospf-security-extension-manual-keying, Work in Progress.

14. Acknowledgments

The authors would like to recognize Nabil Bitar, Edward Crabbe, Don Fedyk, Acee Lindem, David McDysan, and Ayman Soliman for their contributions to this document.

The authors would also like to acknowledge Curtis Villamizar for his significant comments and direct content collaboration.

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

15. Author's Addresses

Spencer Giacalone
Unaffiliated

Email: spencer.giacalone@gmail.com

Dave Ward
Cisco Systems
170 West Tasman Dr.
San Jose, CA 95134, USA

Email: dward@cisco.com

John Drake
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089, USA

Email: jdrake@juniper.net

Alia Atlas
Juniper Networks

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

Email: akatlas@juniper.net

Stefano Previdi
Cisco Systems
Via Del Serafico 200
00142 Rome
Italy

Email: sprevidi@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 21, 2013

Z. Li
M. Chen
Huawei
February 17, 2013

Routing Extensions for Discovery of Role-based MPLS Label Switching
Router (MPLS LSR) Traffic Engineering (TE) Mesh Membership
draft-li-ccamp-role-based-automesh-00

Abstract

A Traffic Engineering (TE) mesh-group is defined as a group of Label Switch Routers (LSRs) that are connected by a full mesh of TE LSPs. Routing (OSPF and IS-IS) extensions for discovery Multiprotocol Label Switching (MPLS) LSR TE mesh membership has been defined to automate the creation of mesh of TE LSPs.

This document introduces a role-based TE mesh-group that applies to the scenarios where full mesh TE LSPs are not necessary and TE LSPs setup depends on the roles of LSRs in a TE mesh-group. Interior Gateway Protocol (IGP) routing extensions for automatic discovery of role-based TE mesh membership are defined accordingly.

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 August 21, 2013.

Copyright Notice

Copyright (c) 2013 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. Role-based TE Mesh Group	3
3. Role-based TE-MESH-GROUP TLV Formats	4
3.1. OSPF Role-based TE-MESH-GROUP TLV Format	4
3.2. IS-IS Role-based TE-MESH-GROUP Sub-TLV Format	7
4. Elements of Procedure	9
4.1. OSPF	9
4.2. IS-IS	10
5. Backward Compatibility	11
6. IANA Considerations	11
6.1. OSPF	11
6.2. IS-IS	12
7. Security Considerations	12
8. Acknowledgements	12
9. References	12
9.1. Normative References	12
9.2. Informative References	13
Authors' Addresses	13

1. Introduction

A TE mesh-group [RFC4972] is defined as a group of LSRs that are connected by a full mesh of TE LSPs. [RFC4972] specifies Intermediate System-to-Intermediate System (IS-IS) and Open Shortest Path First (OSPF) extensions to provide an automatic discovery of the set of LSR members of a TE mesh-group in order to automate the creation of such mesh of TE LSPs. This is called "auto-mesh TE" or "auto-mesh". The auto-mesh TE largely simplifies the configurations and deployments of TE LSPs.

In some scenarios, it may not necessary to establish full mesh TE LSPs among all the LSRs of a TE mesh-group. An example of the scenarios is the mobile backhaul networks, TE LSPs are normally setup between the Cell Site Gateways(CSGs) and the Radio Network Controller (RNC) Site Gateways(RSGs). TE LSPs between/among CSGs and TE LSPs between RSGs are not necessary. And normally, the amount of CSGs is very large in the real deployments, with the auto-mesh mechanism defined[RFC4972], full mesh TE LSPs will be established among the CSGs and RSGs, this will result in large amount of unnecessary TE LSPs established among CSGs and between RSGs. This may not be scale for the CSG devices and is waste of network resources.

So, there are requirements to optimize the auto-mesh TE hence to reduce the unnecessary TE LSPs. This document introduces a "role-based auto-mesh TE" or "role-based auto-mesh" where the setup of the TE LSPs are dependent on the roles of the LSRs within a TE mesh-group. Therefore, besides the discovery of the membership of a TE mesh-group, it needs to discover the role of each node in the TE mesh-group.

Another scenario to which the role-based auto-mesh TE can apply is the Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Point-to-Multipoint (P2MP) TE LSP[RFC4875] scenario. For a RSVP-TE P2MP TE LSP, the root LSR has to know all the leaf LSRs before signaling the P2MP TE LSP. The automatic discovery mechanisms defined in this document can be used to discover the leaf LSRs for P2MP TE LSPs.

This document defines IGP routing extensions to automatically discover of the members and their roles of a TE mesh-group.

2. Role-based TE Mesh Group

A role-based TE mesh-group is a special TE mesh-group where TE LSPs will not be established among all member LSRs. LSRs in a role-based TE mesh-group will have different roles. The TE LSPs setup depends

on the roles of the LSRs in a TE mesh-group. This document introduces two types of roles: Hub-Spoke LSRs, and Root-Leaf LSRs. So there would be Hub-Spoke TE mesh-group and Root-Leaf TE mesh-group.

For a Hub-Spoke TE mesh-group, an LSR can be either a Hub or Spoke LSR in a group, but cannot be both. The rules for Hub-Spoke TE mesh-group are as follows:

TE LSPs SHOULD only be setup between Spoke and Hub LSR.

TE LSPs MUST NOT be setup between/among Spoke LSRs.

TE LSPs MUST NOT be setup between/among Hub LSRs.

For a Root-Leaf TE mesh-group, an LSR can be a Root, a Leaf or both a Root and Leaf LSR. Once the membership and roles are determined, the root LSRs can signal the P2MP TE LSPs toward all the Leaf LSRs. There may be multiple P2MP TE LSPs within a TE mesh-group.

How to signal the TE LSPs is out the scope of this document.

3. Role-based TE-MESH-GROUP TLV Formats

3.1. OSPF Role-based TE-MESH-GROUP TLV Format

The OSPF Role-based TE-MESH-GROUP TLV is used to advertise that an LSR joins/leaves a TE mesh-group and the role of the LSR in the TE mesh-group. The OSPF Role-based TE-MESH-GROUP TLV format for IPv4 (Figure 2) and IPv6 (Figure 3) is as follows:

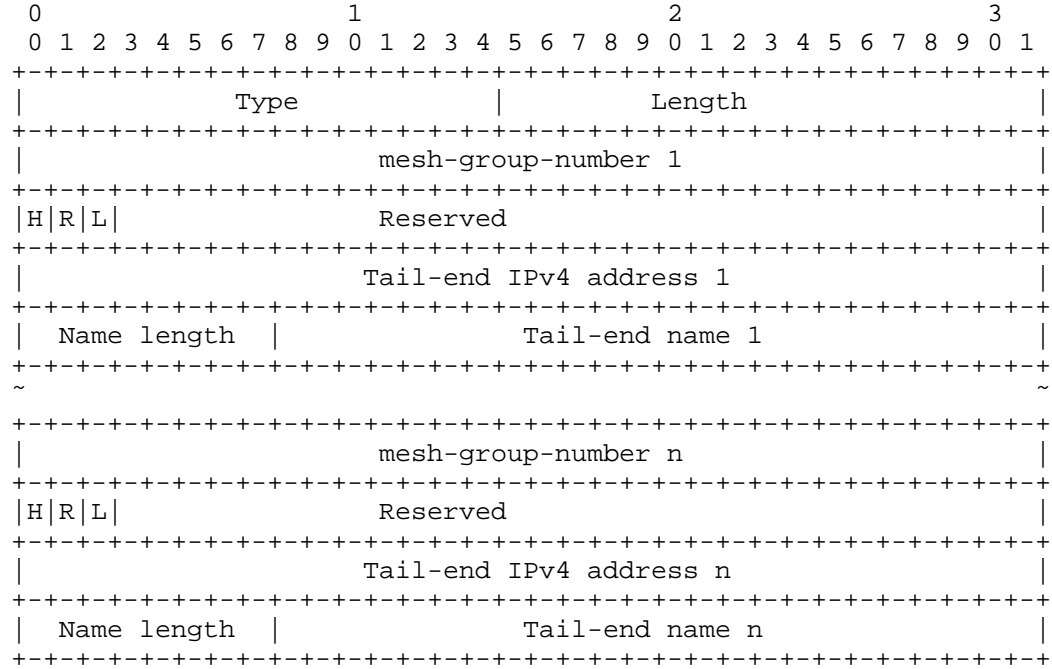


Figure 2 - OSPF TE-MESH-GROUP TLV format (IPv4 Address)

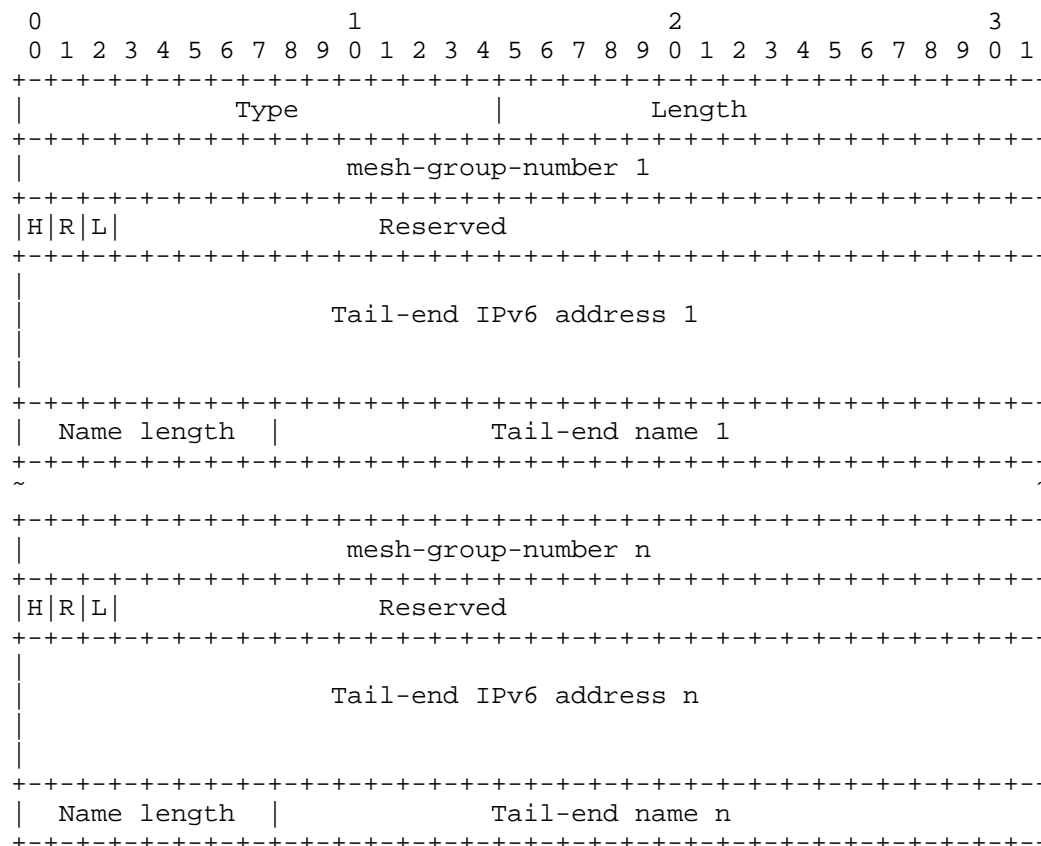


Figure 3 - OSPF TE-MESH-GROUP TLV format (IPv6 Address)

The Type of OSPF TE-MESH-GROUP TLV for IPv4 is TBD1, the value of the Length is variable.

The Type of OSPF TE-MESH-GROUP TLV for IPv6 is TBD2, the value of the Length is variable.

The OSPF Role-based TE-MESH-GROUP TLV may contain one or more role-based mesh-group entries. And each entry corresponds to a TE mesh-group. The definition of the mesh-group-number, the Tail-end address, the Name length and the Tail-end name in each role-based mesh group entry is the same as that of OSPF TE-MESH-GROUP TLV defined in [RFC4972].

In addition, for each mesh group entry, an four-octet flag field is introduced and three flags are defined in this document. Other bits are reserved for future use and MUST be set to zero when sent, and

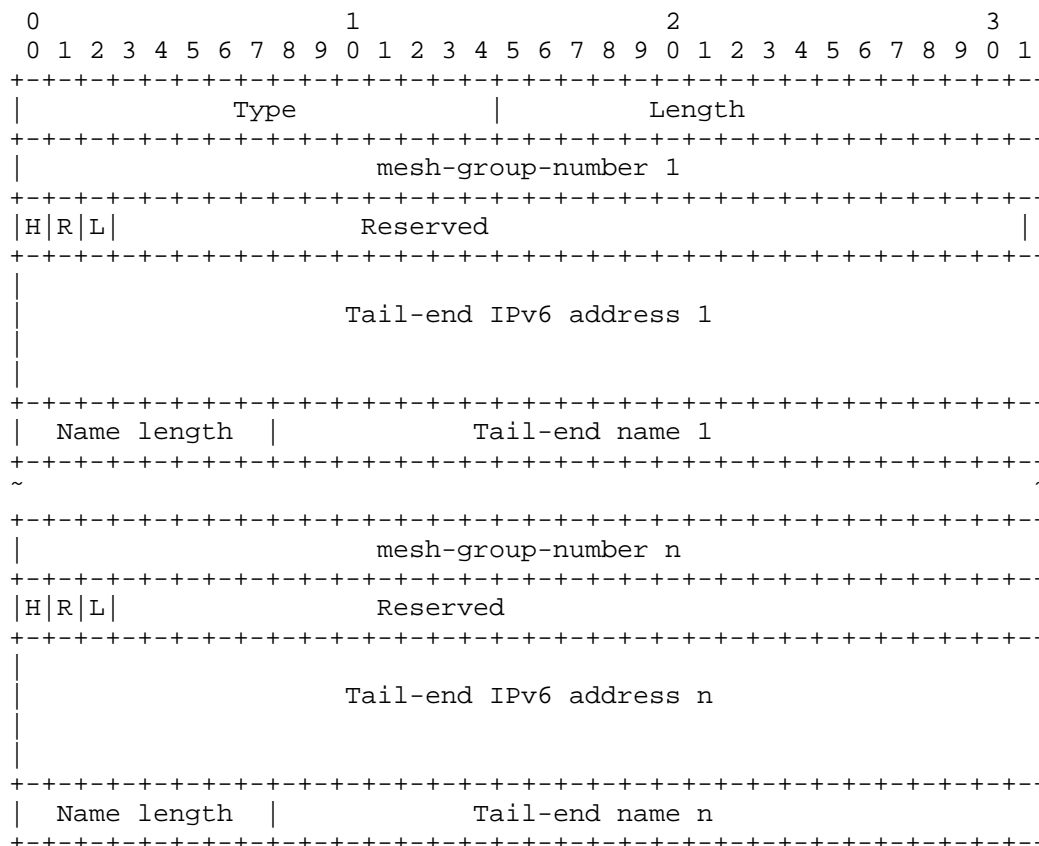


Figure 5 - IS-IS TE-MESH-GROUP sub-TLV format (IPv6 Address)

The Type of IS-IS TE-MESH-GROUP sub-TLV for IPv4 is TBD3, the value of the Length is variable.

The Type of IS-IS TE-MESH-GROUP sub-TLV for IPv6 is TBD4, the value of the Length is variable.

The IS-IS Role-based TE-MESH-GROUP sub-TLV may contain one or more role-based mesh-group entries. And each entry corresponds to a TE mesh-group. The definition of the fields, mesh-group-number, Tail-end address, Name length and Tail-end name in each role-based mesh group entry is the same as that of IS-IS TE-MESH-GROUP sub-TLV defined in [RFC4972].

The H, R and L bits are defined as in Section 3.1 of this document.

4. Elements of Procedure

The OSPF Role-based TE-MESH-GROUP TLV is carried within the OSPF Routing Information LSA and the IS-IS Role-based TE-MESH-GROUP sub-TLV is carried within the IS-IS Router capability TLV. As such, elements of procedure are inherited from those defined in [RFC4970] and [RFC4971] for OSPF and IS-IS respectively. Specifically, a router MUST originate a new LSA/LSP whenever the content of this information changes, or whenever required by regular routing procedure (e.g., updates).

The Role-based TE-MESH-GROUP TLV is OPTIONAL and MUST NOT include more than one of each of the IPv4 instances or the IPv6 instance. If either the IPv4 or the IPv6 OSPF Role-based TE-MESH-GROUP TLV occurs more than once within the OSPF Router Information LSA, only the first instance is processed, subsequent TLV(s) SHOULD be silently ignored. Similarly, if either the IPv4 or the IPv6 IS-IS Role-based TE-MESH-GROUP sub-TLV occurs more than once within the IS-IS Router capability TLV, only the first instance is processed, subsequent TLV(s) SHOULD be silently ignored.

4.1. OSPF

The Role-based TE-MESH-GROUP TLV is advertised within an OSPF Router Information opaque LSA (opaque type of 4, opaque ID of 0) for OSPFv2 [RFC2328] and within a new LSA (Router Information LSA) for OSPFv3 [RFC5340]. The Router Information LSAs for OSPFv2 and OSPFv3 are defined in [RFC4970].

A router MUST originate a new OSPF router information LSA whenever the content of any of the advertised TLV changes or whenever required by the regular OSPF procedure (LSA update (every LSRefreshTime)). If an LSR desires to join or leave a particular TE mesh group or an LSR desires to change its role in a mesh group, it MUST originate a new OSPF Router Information LSA comprising the updated Role-based TE-MESH-GROUP TLV. In the case of a join, a new entry will be added to the role-based TE-MESH-GROUP TLV; if the LSR leaves a mesh-group, the corresponding entry will be removed from the role-based TE-MESH-GROUP TLV; if the LSR changes its role in the mesh group, the corresponding entry will be updated in the role-based TE-MESH-GROUP TLV. Note that these operations can be performed in the context of a single LSA update. An implementation SHOULD be able to detect any change to a previously received role-based TE-MESH-GROUP TLV from a specific LSR.

As defined in [RFC5250] for OSPFv2 and in [RFC5340] for OSPFv3, the flooding scope of the Router Information LSA is determined by the LSA Opaque type for OSPFv2 and the values of the S1/S2 bits for OSPFv3.

For OSPFv2 Router Information opaque LSA:

- Link-local scope: type 9;
- Area-local scope: type 10;
- Routing-domain scope: type 11. In this case, the flooding scope is equivalent to the Type 5 LSA flooding scope.

For OSPFv3 Router Information LSA:

- Link-local scope: OSPFv3 Router Information LSA with the S1 and S2 bits cleared;
- Area-local scope: OSPFv3 Router Information LSA with the S1 bit set and the S2 bit cleared;
- Routing-domain scope: OSPFv3 Router Information LSA with S1 bit cleared and the S2 bit set.

A router may generate multiple OSPF Router Information LSAs with different flooding scopes.

The Role-based TE-MESH-GROUP TLV may be advertised within an Area-local or Routing-domain scope Router Information LSA, depending on the MPLS TE mesh group profile:

- If the MPLS TE mesh-group is contained within a single area (all the LSRs of the mesh-group are contained within a single area), the Role-based TE-MESH-GROUP TLV MUST be generated within an Area-local Router Information LSA.
- If the MPLS TE mesh-group spans multiple OSPF areas, the TE Role-based mesh-group TLV MUST be generated within a Routing-domain scope router information LSA.

When the router receives Role-based TE-MESH-GROUP TLV, it SHOULD setup MPLS TE LSPs according rules which defined in the Section 3.

4.2. IS-IS

The Role-based TE-MESH-GROUP sub-TLV is advertised within the IS-IS Router CAPABILITY TLV defined in [RFC4971].

An IS-IS router MUST originate a new IS-IS LSP whenever the content of any of the advertised sub-TLV changes or whenever required by regular IS-IS procedure (LSP updates). If an LSR desires to join or leave a particular TE mesh group or an LSR desires to change its role

in a mesh group, it MUST originate a new LSP comprising the refreshed IS-IS Router capability TLV comprising the updated TE-MESH-GROUP sub-TLV. In the case of a join, a new entry will be added to the TE-MESH-GROUP sub-TLV; if the LSR leaves a mesh-group, the corresponding entry will be deleted from the TE-MESH-GROUP sub-TLV; if the LSR changes its role in the mesh group, the corresponding entry will be updated in the role-based TE-MESH-GROUP sub-TLV. Note that these operations can be performed in the context of a single update. An implementation SHOULD be able to detect any change to a previously received TE-MESH-GROUP sub-TLV from a specific LSR.

If the flooding scope of a Role-based TE-MESH-GROUP sub-TLV is limited to an IS-IS level/area, the sub-TLV MUST NOT be leaked across level/area and the S flag of the Router CAPABILITY TLV MUST be cleared. Conversely, if the flooding scope of a Role-based TE-MESH-GROUP sub-TLV is the entire routing domain, the TLV MUST be leaked across IS-IS levels/areas, and the S flag of the Router CAPABILITY TLV MUST be set. In both cases, the flooding rules specified in [RFC4971] apply.

As specified in [RFC4971], a router may generate multiple IS-IS Router CAPABILITY TLVs within an IS-IS LSP with different flooding scopes.

When the router receives Role-based TE-MESH-GROUP sub-TLV, it SHOULD setup MPLS TE LSPs according rules which defined in the Section 3.

5. Backward Compatibility

For a role-based TE mesh-group, if there are some LSRs only supporting mechanisms defined [RFC4972], all the LSRs of the mesh-group MUST process as defined in [RFC4972]. The operators should avoid to add an LSR that does not support role-based auto-mesh TE to a role-based TE mesh-group.

6. IANA Considerations

6.1. OSPF

The registry for the Router Information LSA is defined in [RFC4970]. IANA assigned a new OSPF TLV code-point for the Role-based TE-MESH-GROUP TLVs carried within the Router Information LSA.

Value	TLV	References
-----	-----	-----
TBD1	Role-based TE-MESH-GROUP TLV (IPv4)	this document
TBD2	Role-based TE-MESH-GROUP TLV (IPv6)	this document

6.2. IS-IS

The registry for the Router Capability TLV is defined in [RFC4971]. IANA assigned a new IS-IS sub-TLV code-point for the TE-MESH-GROUP sub-TLVs carried within the IS-IS Router Capability TLV.

Value	Sub-TLV	References
-----	-----	-----
TBD3	Role-based TE-MESH-GROUP sub-TLV (IPv4)	this document
TBD4	Role-based TE-MESH-GROUP sub-TLV (IPv6)	this document

7. Security Considerations

The function described in this document does not create any new security issues for the OSPF and IS-IS protocols, the security considerations described in [RFC4972] apply here.

8. Acknowledgements

The authors would like to thank Loa Andersson for his valuable comments.

The authors would also like to thank the authors of [RFC4972] from where we have taken most of the elements procedures.

9. References

9.1. 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.
- [RFC4970] Lindem, A., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.
- [RFC4971] Vasseur, JP., Shen, N., and R. Aggarwal, "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", RFC 4971, July 2007.
- [RFC4972] Vasseur, JP., Leroux, JL., Yasukawa, S., Previdi, S., Psenak, P., and P. Mabbey, "Routing Extensions for Discovery of Multiprotocol (MPLS) Label Switch Router (LSR) Traffic Engineering (TE) Mesh Membership", RFC 4972,

July 2007.

[RFC5250] Berger, L., Bryskin, I., Zinin, A., and R. Coltun, "The OSPF Opaque LSA Option", RFC 5250, July 2008.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.

9.2. Informative References

[RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

Authors' Addresses

Zhenbin Li
Huawei

Email: lizhenbin@huawei.com

Mach(Guoyi) Chen
Huawei

Email: mach.chen@huawei.com

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: August 22, 2013

Z. Li
L. Zhang
Y. Liu
Huawei Technologies
February 18, 2013

OSPF Extensions for Automatic Computation of MPLS Traffic Engineering
Path Using Traffic Engineering Layers and Areas
draft-li-ospf-auto-mbb-te-path-00

Abstract

As the network scale expands, especially in the mobile backhaul network, automatic computation of MPLS Traffic Engineering (TE) path becomes very important. But owing to requirements on the MPLS TE path, explicit path or affinity property has to be introduced for the path computation. This causes the complexity of MPLS TE path design. The document proposes an architecture and corresponding OSPF extensions to improve automation on computation of MPLS TE path. MPLS TE networks are divided into different traffic engineering layers and areas according to the characteristics of the network topology. MPLS TE path can compute automatically based on traffic engineering layers and areas to satisfy major requirements to bear mobile network services.

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 August 22, 2013.

Copyright Notice

Copyright (c) 2013 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. Problem Statement	3
2.1. Mobile Backhaul Network and Service Deployment	3
2.2. Weakness of Existing MPLS TE Path Computation	4
3. Architecture of MPLS TE Auto Path Computation	6
3.1. Concept of TL and TA	6
3.2. TL and TA Information Flooding	8
3.3. Enhanced CSPF Algorithm Based on TL and TA	8
3.3.1. An Example of Enhanced CSPF Algorithm Based on TL and TA	9
4. OSPF Extensions	10
4.1. OSPF TA TLV and TL TLV Format	10
4.2. Elements of Procedure	11
4.3. Backward Compatibility	12
5. IANA Considerations	12
6. Security Considerations	12
7. Normative References	12
Authors' Addresses	13

1. Introduction

As the network scale expands, especially in the mobile backhaul network, automatic computation of MPLS TE path becomes very important. Since the mobile traffic has high SLA (Service Level Agreement) requirement, MPLS TE is introduced to provide bandwidth guarantee and traffic protection. On the other hand, in order to provide traffic engineering properties, constraints such as explicit path or affinity property has to be specified for a MPLS TE tunnel. This causes that the path design is very complex. For example, when explicit path is specified for a MPLS TE tunnel in a large scale network, many hops along the MPLS TE path have to be specified. This operation is cumbersome and error-prone. In addition, if new nodes are introduced in the network, a lot of configuration of existing explicit paths has to be changed.

This document proposes an architecture and corresponding OSPF extensions to improve automation on computation of MPLS TE path. MPLS TE layers and areas are introduced according to the characteristics of the network topology. MPLS TE path can compute more automatically based on MPLS TE layers and areas to reduce the operation expense greatly.

2. Problem Statement

2.1. Mobile Backhaul Network and Service Deployment

Mobile multimedia devices such as smartphones are ubiquitous now which runs a wide variety of bandwidth-intensive applications and causes unprecedented growth in mobile data traffic. The huge growth is challenging legacy network infrastructure. There are two obvious solutions to cope with the growing bandwidth:

-- Increase the radio wireless interface bandwidth

-- Increase more cell sites: more LTE eNodeBs and associated Cell Site Gateways(CSGs) are added in the networks. This causes the network scale expands fast and has much effect on the service provision.

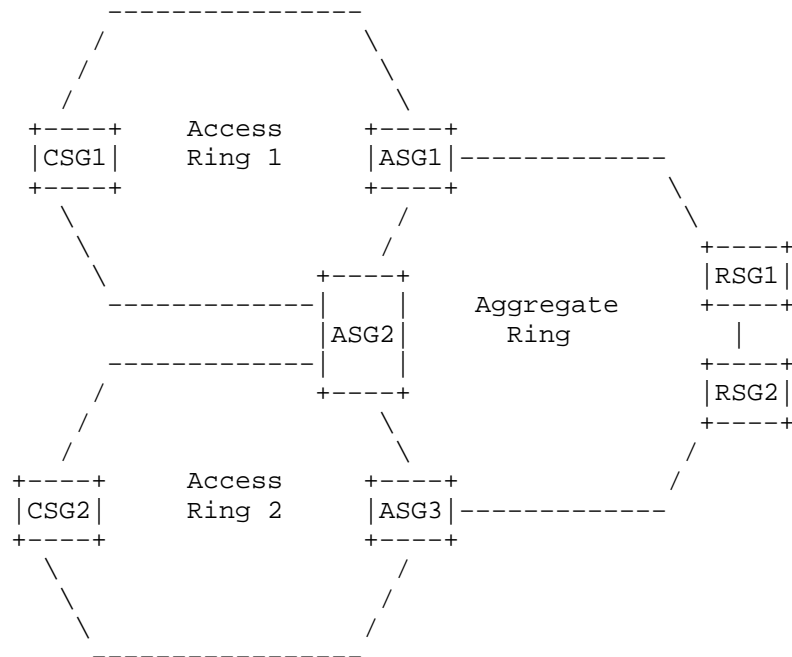


Figure 1 Mobile Backhaul Network

The topology of mobile backhaul network is shown in the Figure 1. It usually adopts ring topology to save fiber resource and it is divided into the aggregate network and the access network. Cell Site Gateways(CSGs) connects the eNodeBs and RNC site gateways(RSGs) connects the RNCs. The mobile traffic is transported from CSGs to RSGs. The network takes a typical aggregate traffic model that more than one access rings will attach to one pair of aggregate site gateways(ASGs) and more than one aggregate rings will attach to one pair of RSGs.

2.2. Weakness of Existing MPLS TE Path Computation

Since the mobile traffic has high SLA (Service Level Agreement) requirement, MPLS TE is introduced to provide bandwidth guarantee and traffic protection. As the network scale expands, automation becomes more and more important to reduce the effort of service provision. But the path design becomes complex inevitably owing to guarantee traffic engineering properties. There are following two primary requirements for MPLS TE path computation:

1. Completely disjointed primary and backup LSP

MPLS TE Hot-standby feature is introduced to implement traffic protection. That is, primary LSP and backup LSP are setup at the same time for one MPLS TE tunnel. In order to achieve higher protection, it is required that the primary and backup LSP should not share any nodes and links. Thus when failure happens in the primary path, the backup LSP can always take over the traffic.

According to current SPF(Shortest Path First) algorithm, if there is no other constraints, it may be difficult to satisfy above path computation requirement. For example, in figure 1 the primary path computed from CSG1 to RSG1 may be CSG1->ASG2->ASG1->RSG1. Since the primary path passes through both ASG2 and ASG1, the backup path cannot be disjointed completely from the primary path. In fact, it is apparent that the two completely disjointed paths exist from CSG1 to RSG1 in the figure 1.

2. Avoid passing through different access rings

When the mobile traffic is transported from the CSG to the RSG, it is expected that the path would not pass through multiple access rings. Since the bandwidth of the access ring is always designed to satisfy requirement of its own, if mobile traffic from other access ring passes through, the access ring is prone to be overloaded which will cause traffic loss owing to traffic congestion.

When automatic path computation is done for MPLS TE tunnels, it may be inevitable that the path will pass through multiple access rings. For example, in figure 1 the primary path computed from CSG1 to RSG2 may be CSG1->ASG2->CSG2->ASG3->RSG2 instead of CSG1->ASG2->ASG3->RSG2.

There are two possible solutions to satisfy requirements described above:

The first one is to set reasonable link cost. For example, the cost of the key link between ASG1 and ASG2 can be set as a large value, then the primary LSP will not be calculated to pass through the key link and the backup LSP can be disjointed from the primary LSP completely. The cost of the access ring can also be larger than the aggregate ring to avoid that the traffic will pass through unexpected rings.

The second one is to use explicit-path or affinity property to achieve better path design. When explicit path is used, it has to designate the exact nodes or links which the primary LSP and the backup LSP go through. When affinity property is used, it can divide different rings with different colors and the primary LSP and backup LSP can be setup with different affinity property.

The two methods can satisfy the two requirements of path computation. But as we know the mobile backhaul network faces more frequent topology change than the fixed network. Adding and deleting of eNodeB will change the access ring topology and which will change the hops and cost for mobile traffic from the source to the destination. It will be very complex and time-consuming to adjust the cost for a large scale network or change explicit path or affinity property for a great deal of MPLS TE tunnels. It is necessary to propose a more automatic way to satisfy the requirements.

3. Architecture of MPLS TE Auto Path Computation

3.1. Concept of TL and TA

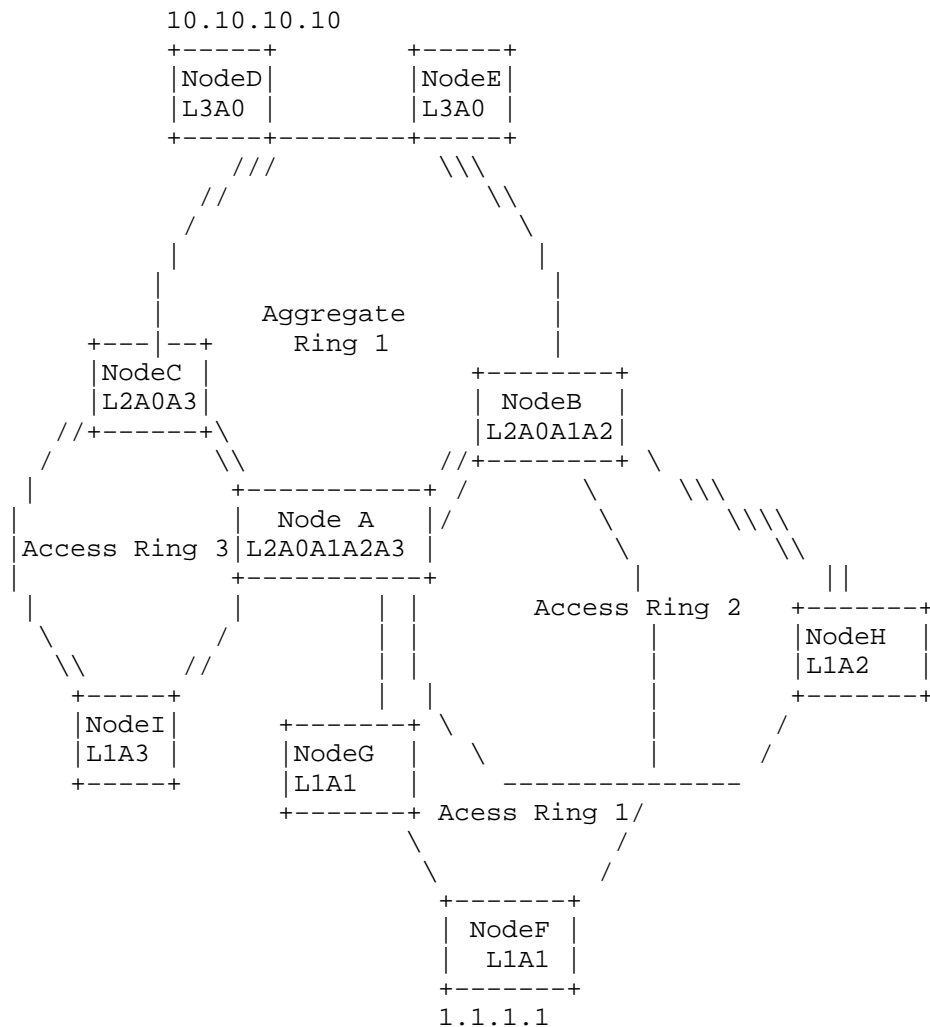


Figure 2 Definition of TAs and TLs

New network constraints are introduced to improve automation of MPLS TE path computation. As the figure above shows, the mobile backhaul network can be divided into multiple layers and multiple areas. The layers and areas can be designated easily according to the natural physical topology. We propose two concepts below:

- o TE Layer (TL): It indicates the physical layer of the node in the network. The TL value should be increased from the access ring to the aggregate ring layer by layer. The TL values from the access ring to the aggregate ring can be not continuous. They just

reflect the relation of the different layers. In order to accommodate future network expansion, it is better that the lowest TL value should not start from the 0 or 1.

- o TE Area(TA): It indicates the physical ring of the node. All nodes of the physical ring forms a natural area. TA value must be unique in the whole network. TA is designed mostly according to the physical topology with the aim to separate the obvious physical areas. One node can have multiple TA values when it belongs to multiple rings.

TL and TA are defined for every node instead of every link to reduce the effort of configuration and operation. TA and TL indicates the network layer and area which one node belongs to. TL and TA value should be set for the node before the path of the TE LSP is calculated just like that the cost of the link should be set before the routes are calculated. TL and TA are only defined for MPLS TE path computation according to the natural topology of the mobile network. They have no relationship with IGP area or level.

3.2. TL and TA Information Flooding

After the TL and TA value are set for the node, the TL and TA information of this node should be flooded through IGP. When all nodes TL and TA information are flooded, every node in this route region will have the whole TL and TA information which will be added to the TEDB for TE LSP calculation. When a TE LSP requires path computation in a source node, a new enhanced CSPF algorithm based on TL and TA will be used to calculate the optimal path automatically.

3.3. Enhanced CSPF Algorithm Based on TL and TA

The enhanced CSPF algorithm based on TL and TA can calculate the TE path more automatically comparing with the existing CSPF algorithm. In order to achieve more automatic path computation, some new rules are introduced for the CSPF algorithm.

We assume that:

- o The high layer is TL high(TLh), the low layer is TL low(TLl);
- o The source node of the LSP has the TA value TAs, the destination node of LSP has the TA value TAd, the passed node has the TA value TAp.

The rules for the enhanced CSPF algorithm are as follows:

- o Rule 1: If the destination node of the LSP is not in the same TA as the source node or the passed node, the node in the different layer will be the potential next-hop for the LSP path calculation.
- o Rule 2: One LSP's TL track can not include TLh->TLl->TLh, this means that the LSP cannot pass through the low layer twice.
- o Rule 3: If the LSP reach a node that in the same TA as the destination node, the LSP must be calculated in this TA only.
- o Rule 4: If the LSP reach a node that among more than one TAs, the node in different TA should be prior to be the next hop. This rule ensures that the primary and backup LSPs would not pass the same links.

Since these rules are applied to calculate both the primary and secondary path automatically, rules for determining which is the primary or the secondary should also be introduced. The rules are as follows:

- o Rule 5: The LSP which passes fewer TLs will be the primary LSP.
- o Rule 6: If the two LSPs passes the same TLs, the one with shorter metric in every layer from high to low will be the main LSP

3.3.1. An Example of Enhanced CSPF Algorithm Based on TL and TA

As the figure above shows, the TL and TA values are designed for every node and the flooding has completed. Now the primary LSP and the backup LSP should setup from the source node(1.1.1.1) to the destination node(10.10.10.10), the path calculation is as follows:

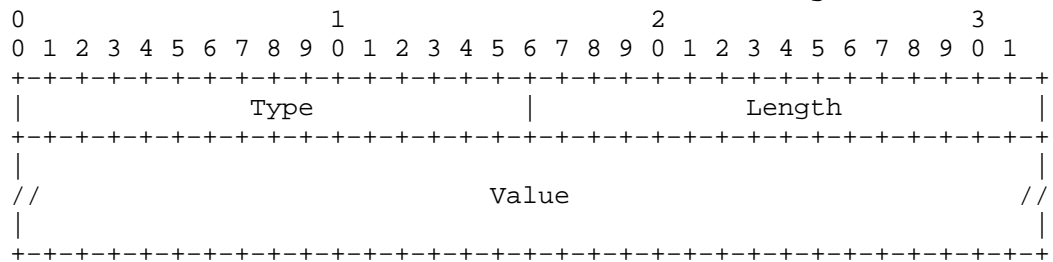
1. The source node(1.1.1.1) is TL1TA1 and the destination node(10.10.10.10) is TL3TA0. The LSP path should be calculated towards the node with higher TL value in TA1, according to Rule 1. The candidate nodes are NodeA and NodeB and we assume that the algorithm will choose NodeB as the next hop according to the cost.
2. After get NodeB, there are three candidate nodes for the next hop which are NodeA and NodeE and NodeH. Node H will be excluded according to Rule 2, because it will cause the LSP to pass through TL2->TL1->TL2, that means the LSP will pass another access ring which is on the same low layer as the source node.
3. NodeB is in TA0, which is the same as the destination node, so we can only choose NodeA or Node E, according to Rule 3

4. TA1 has been passed, so the NodeB in TA1 is excluded according to rule4
5. Node E is the best appropriate choice according to the Rules.As a result ,we can get a path NodeF->NodeB->NodeE->NodeD
6. The other path is calculated according to the rules with the nodes and links passed by the first path excluded. So we can get the other path NodeF->NodeA->NodeC->NodeD.
7. Then we will select the primary path from these two paths. According to the rule5 and rule6, the path NodeF->NodeA->NodeC->NodeD is determined as the primary LSP and the path NodeF->NodeB->NodeE->NodeD is the backup LSP.

4. OSPF Extensions

4.1. OSPF TA TLV and TL TLV Format

The OSPF TA TLV and TL TLV are used to advertise the TA and TL a node belongs to. The OSPF TA TLV and TL TLV (advertised in an OSPF router information LSA defined in [RFC4970]) has the following format:



Where

Type: identifies the TLV type

Length: the length of the value field in octets

The format of the TA TLV and TL TLV are the same as the TLV format used by the Traffic Engineering Extensions to OSPF (see [RFC3630]).

TLV:

OSPFv2 TA TYPE: TBD,

OSPFv2 TL TYPE: TBD,

OSPFv3 TA TYPE: TBD

OSPFv3 TL TYPE: TBD

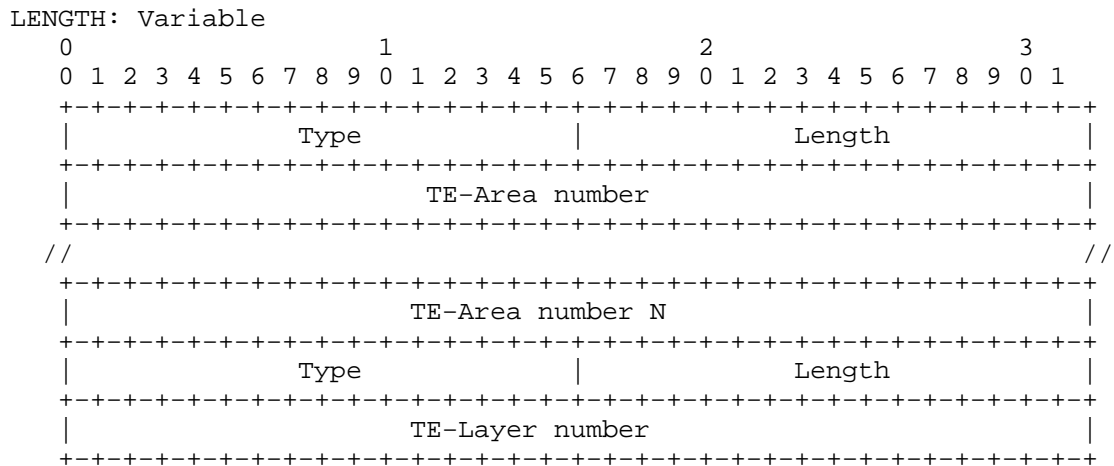


Figure 1 - OSPF TE-Area Group TLV and TE-Layer Group TLV format

4.2. Elements of Procedure

The OSPF TA and TL TLV is carried within the OSPF Routing Information LSA. Specifically, a router MUST originate a new LSA whenever the content of this information changes, or whenever required by regular routing procedure (e.g., updates). The OSPF TLVs are OPTIONAL and MUST NOT included more than one instance. If either of the TLVs occurs more than once within the OSPF Router Information LSA, only the first instance is processed, subsequent TLV(s) SHOULD be silently ignored.

When the TA or TL of a node change, a new router information LSA SHOULD be advertised. The flood scope is OSPF Area using type 10 LSA or Routing-domain scope using type 11 LSA.

As defined in [RFC2370] for OSPFv2 and in [RFC2740]for OSPFv3, the flooding scope of the Router Information LSA is determined by the LSA Opaque type for OSPFv2 and the values of the S1/S2 bits for OSPFv3.

The TA TLV and TL TLV may be advertised within an Area-local or Routing-domain scope Router Information LSA, depending on the MPLS TE profile:

- If the MPLS TE Area and Layer are contained within a single area, the TA TLV and TL TLV MUST be generated within an Area-local Router Information LSA.
- If the MPLS TE Area and Layer spans multiple OSPF areas, the TA TLV and TL TLV MUST be generated within a Routing-domain scope router

information LSA.

4.3. Backward Compatibility

The TLVs defined in this document do not introduce any interoperability issue. For OSPF, a router not supporting the TLV SHOULD just silently ignore the TLV as specified in [RFC2370].

5. IANA Considerations

The registry for the Router Information LSA is defined in [RFC4970]. IANA assigned a new OSPF TLV code-point for the OSPF-TE-Attributes TLVs carried within the Router Information LSA.

Value	Sub-TLV	References
-----	-----	-----
TBD	OSPF-TE-Area TLV (IPv4)	RFC 4970
TBD	OSPF-TE-Layer TLV (IPv4)	RFC 4970
TBD	OSPF-TE-Area TLV (IPv6)	RFC 4970
TBD	OSPF-TE-Layer TLV (IPv6)	RFC 4970

6. Security Considerations

TBD.

7. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2370] Coltun, R., "The OSPF Opaque LSA Option", RFC 2370, July 1998.
- [RFC2740] Coltun, R., Ferguson, D., and J. Moy, "OSPF for IPv6", RFC 2740, December 1999.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC4970] Lindem, A., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.

Authors' Addresses

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Li Zhang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: monica.zhangli@huawei.com

Yuanjiao Liu
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: liuyuanjiao@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 17, 2014

G. Yan
J. Yang
Z. Li
Huawei Technologies
October 14, 2013

OSPF Extensions for MPLS Green Traffic Engineering
draft-li-ospf-ext-green-te-01

Abstract

The energy-saving is one important topic in the world, and now most of technologies for energy-saving focus on the hardware design instead of the energy saving design based on the whole network. This document proposes OSPF extensions to synchronize the energy consumption parameter of each node in the network. These parameters can be used for the energy saving design.

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 RFC2119 [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 April 17, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

Authors' Addresses

Gang Yan
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: yangang@huawei.com

Jianjun Yang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: jack.yangjianjun@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com


```

+-----+
|               OSPF Hold Timer Value               |
+-----+

```

3. Routers must set the HELLO and Dead interval values carried in the OSPF HELLO packet to zero.

4. To stop advertising the asymmetric hold timer value, routers will simply revert back to advertising configured (non-zero) values of HELLO and dead interval in the OSPF Hello packet.

The mechanism of sending HELLOs would remain as specified in Sec 9.5 of RFC 2328.

2.2 Receiving HELLOs with Asymmetric Hold Timer Values.

The processing of an incoming HELLO packet with the L-bit set, and containing the extended options set for alternate HELLOs would follow the specification in Sec 10.5 of RFC 2328 with one modification.

1. Routers that recognize this new extended options will set the value of the neighbor dead interval to the value specified in the LLS block TLV, but only if BOTH the HELLO and dead interval are set to zero in the OSPF HELLO packet.

2. Routers that do not recognize the extended options would drop adjacency as it will not match with the configured (or default) HELLO or dead interval as specified in Sec 10.5 of RFC 2328.

Note that, routers can stop appending the LLS block in their HELLO, and the neighbors will simply (re)start using the value specified in the HELLO packet.

4 Discussion

It is possible to use Bidirectional Forwarding Detection (BFD) [RFC 5880] to alleviate some of the concerns in the use-cases identified above. Relying entirely on BFD without OSPF HELLOs is not a possibility given that OSPF HELLOs are still used for discovery of neighbors. The BFD approach has its own shortcomings such as limited cross-vendor and cross-platform support and also performance implications, especially with increasing scale requirements. In any case, BFD can be made to work in conjunction with the proposal in this document to achieve the best possible network performance. It is intended that the proposal for asymmetric hold timer would work independent of BFD deployment considerations, and could also help in new applications where it may be desirable to support asymmetric and possibly dynamic dead interval values (e.g., OSPFv3 Auto-Configuration, [OSPFV3-AUTOCONFIG]).

5 Acknowledgements

The authors would like to thank Paul Wells for careful review of this document. We would also like to thank Anton Smirnov for reviewing this document and bringing the BFD alternative to our attention.

6 Backward Compatibility

No modifications to OSPF packet formats are proposed here. The new EO-TLV introduced here is standard OSPF because LLS-incapable routers will not consider the extra data after the packet; i.e., the LLS data block will be ignored by routers that do not support the LLS extension.

Email: anandmkr@cisco.com

Hasmit Grover
Cisco Systems
170 W Tasman Drive
San Jose, CA 95138
US

Email: hasmit@cisco.com

Abhay Roy
Cisco Systems
170 W Tasman Drive
San Jose, CA 95138
US

Email: akr@cisco.com

- [RFC4813] Friedman, B., Nguyen, L., Roy, A., Yeung, D., and A. Zinin, "OSPF Link-Local Signaling", RFC 4813, March 2007.
- [RFC4970] Lindem, A., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.

Authors' Addresses

Gang Yan
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: yangang@huawei.com

Yuanjiao Liu
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: liuyuanjiao@huawei.com

Xudong Zhang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhangxudong@huawei.com