Open Shortest Path First IGP
Internet-Draft
Intended status: Standards Track
Expires: July 5, 2018

S. Hegde
Juniper Networks, Inc.
P. Sarkar
H. Gredler
Individual
M. Nanduri
ebay Corporation
L. Jalil
Verizon
January 1, 2018

OSPF Link Overload
draft-ietf-ospf-link-overload-11

Abstract

When a link is being prepared to be taken out of service, the traffic needs to be diverted from both ends of the link. Increasing the metric to the highest metric on one side of the link is not sufficient to divert the traffic flowing in the other direction.

It is useful for routers in an OSPFv2 or OSPFv3 routing domain to be able to advertise a link as being in an overload state to indicate impending maintenance activity on the link. This information can be used by the network devices to re-route the traffic effectively.

This document describes the protocol extensions to disseminate link-overload information in OSPFv2 and OSPFv3.

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 https://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 July 5, 2018.

Copyright Notice

Table of Contents

1. Introduction

When a node is being prepared for a planned maintenance or upgrade, [RFC6987] provides mechanisms to advertise the node being in an overload state by setting all outgoing link costs to MaxLinkMetric (0xffff). These procedures are specific to the maintenance activity on a node and cannot be used when a single link on the node requires maintenance.

In traffic-engineering deployments, LSPs need to be diverted from the link without disrupting the services. [RFC5817] describes requirements and procedures for graceful shutdown of MPLS links. It is useful to be able to advertise the impending maintenance activity on the link and to have LSP re-routing policies at the ingress to route the LSPs away from the link.

Many OSPFv2 or OSPFv3 deployments run on overlay networks provisioned by means of pseudo-wires or L2-circuits. Prior to devices in the underlying network going offline for maintenance, it is useful to divert the traffic away from the node before the maintenance is actually performed. Since the nodes in the underlying network are not visible to OSPF, the existing stub router mechanism described in [RFC6987] cannot be used. An application specific to this use case is described in Section 7.1.

This document provides mechanisms to advertise link-overload state in the flexible encodings provided by OSPFv2 Prefix/Link Attribute Advertisement [RFC7684]. Throughout this document, OSPF is used when the text applies to both OSPFv2 and OSPFv3. OSPFv2 or OSPFv3 is used when the text is specific to one version of the OSPF protocol.

2. Motivation

The motivation of this document is to reduce manual intervention during maintenance activities. The following objectives help to accomplish this in a range of deployment scenarios.

1. Advertise impending maintenance activity so that traffic from both directions can be diverted away from the link.

2. Allow the solution to be backward compatible so that nodes that do not understand the new advertisement do not cause routing loops.

3. Advertise the maintenance activity to other nodes in the network so that LSP ingress routers/controllers can learn of the impending maintenance activity and apply specific policies to re-route the LSPs for traffic-engineering based deployments.

4. Allow the link to be used as last resort link to prevent traffic disruption when alternate paths are not available.

3. Flooding Scope

The link-overload information is flooded in area-scoped Extended Link Opaque LSA [RFC7684]. The Link-Overload sub-TLV MAY be processed by the head-end nodes or the controller as described in the Section 7. The procedures for processing the Link-Overload sub-TLV are described in Section 5.

4. Link-Overload sub-TLV

4.1. OSPFv2 Link-overload sub-TLV

The Link-Overload sub-TLV identifies the link as being in overload state.It is advertised in extended Link TLV of the Extended Link Opaque LSA as defined in [RFC7684].

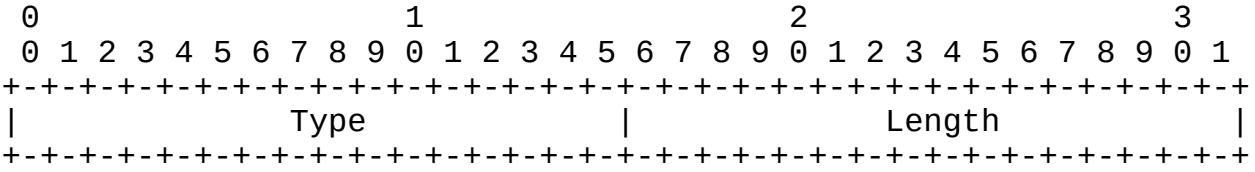

Figure 1: Link-Overload sub-TLV for OSPFv2

Type : TBA (suggested value 7)

Length: 0

4.2. Remote IPv4 Address Sub-TLV

This sub-TLV specifies the IPv4 address of remote endpoint on the link. It is advertised in the Extended Link TLV as defined in [RFC7684]. This sub-TLV is optional and MAY be advertised in area-

scoped Extended Link Opaque LSA to identify the link when there are multiple parallel links between two nodes.

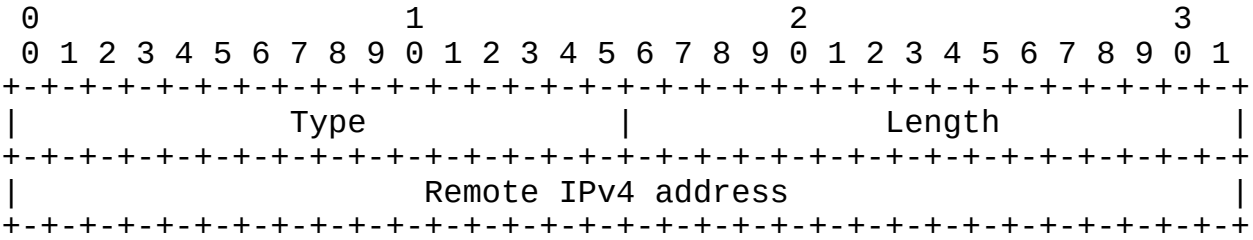

Figure 2: Remote IPv4 Address Sub-TLV

Type : TBA (suggested value 8)

Length: 4

Value: Remote IPv4 address. The remote IP4 address is used to identify the particular link when there are multiple parallel links between two nodes.

4.3. Local/Remote Interface ID Sub-TLV

This sub-TLV specifies local and remote interface identifiers. It is advertised in the Extended Link TLV as defined in [RFC7684]. This sub-TLV is optional and MAY be advertised in area-scoped Extended Link Opaque LSA to identify the link when there are multiple parallel unnumbered links between two nodes. The local interface-id is generally readily available. One of the mechanisms to obtain remote interface-id is described in [RFC4203].

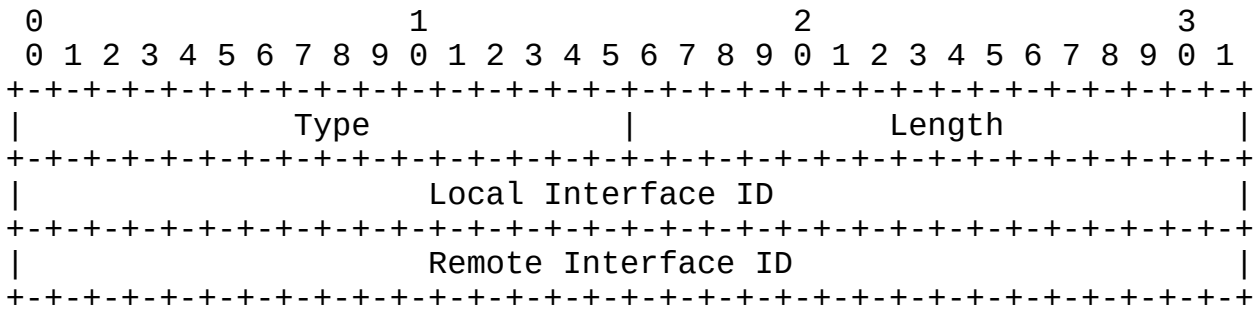

Figure 3: Local/Remote Interface ID Sub-TLV

Type : TBA (suggested value 9)

Length: 8

Value: 4 octets of Local Interface ID followed by 4 octets of Remote interface ID.

4.4. OSPFv3 Link-Overload sub-TLV

The Link Overload sub-TLV is carried in the Router-Link TLV as defined in the [I-D.ietf-ospf-ospfv3-lsa-extend] for OSPFv3. The Router-Link TLV contains the neighbour interface-id and can uniquely identify the link on the remote node.

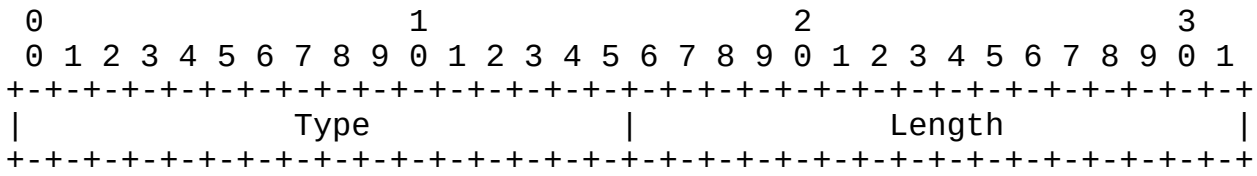

Figure 4: Link-Overload sub-TLV for OSPFv3

Type : TBA (Suggested value 7)

Length: 0

4.5. BGP-LS Link-overload TLV

BGP-LS as defined in [RFC7752] is a mechanism to distribute network information to external entities using BGP routing protocol. link-overload is an imporatant link information that the external entities can use for various usecases as defined in Section 7. BGP Link NLRI

is used to carry the link information. a new TLV called Link-Overload is defined to describe the link attribute corresponding to link-overload state.

4.6. Distinguishing parallel links

Figure 5: Parallel Linkls

Consider two routers A and B connected with two parallel point-to-point interfaces. I.w and I.x represent the Interface address on Router A's side and I.y and I.z represent Interface addresses on Router B's side. The extended link opaque LSA as described in [RFC7684] describes links using link-type, Link-ID and Link-data. For ex. Link with address I.w is described as below on Router A.

Link-type = Point-to-point

Link-ID: Router-ID B

Link-Data = I.w

A third node (controller or head-end) in the network cannot distinguish the Interface on router B which is connected to this particular Interface with the above information. Interface with address I.y or I.z could be chosen due to this ambiguity. In such cases Remote-IPv4 Address sub-TLV should be originated and added to the extended link-TLV. The usecases as described in Section 7 require controller or head-end nodes to interpret the link-overload information and hence the need for the RemoteIPv4 address sub-TLV. I.y is carried in the extended-link-TLV which unambiguously identifies the interface on the remote side. OSPFv3 Router-link-TLV as described in [I-D.ietf-ospf-ospfv3-lsa-extend] contains Interface ID and neighbor's Interface-ID which can uniquely identify connecting interface on the remote side and hence OSPFv3 does not require seperate Remote-IPv6 address to be advertised along with OSPFv2-link-overload-sub-TLV.

5. Elements of procedure

As defined in [RFC7684] every link on the node will have a separate Extended Link Opaque LSA. The node that has the link to be taken out of service SHOULD advertise the Link-Overload sub-TLV in the Extended Link TLV of the Extended Link Opaque LSA as defined in [RFC7684] for OSPFv2. The Link-Overload sub-TLV indicates that the link identified by the sub-TLV is overloaded. The Link-Overload information is advertised as a property of the link and is flooded across the area. This information can be used by ingress routers or controllers to take special actions. An application specific to this use case is described in Section 7.2.

The precise action taken by the remote node at the other end of the link identified as overloaded depends on the link type.

5.1. Point-to-point links

The node that has the link to be taken out of service MUST set metric of the link to MaxLinkMetric (0xffff) and re-originate its router-LSA. The TE metric SHOULD be set to MAX-TE-METRIC (0xfffffffe) and the node SHOULD re-originate the corresponding TE Link Opaque LSAs. When a Link-Overload sub-TLV is received for a point-to-point link, the remote node MUST identify the local link which corresponds to the overloaded link and set the metric to MaxLinkMetric (0xffff)and the remote node MUST re-originate its router-LSA with the changed metric. The TE metric SHOULD be set to MAX-TE-METRIC (0xfffffffe) and the TE opaque LSA for the link SHOULD be re-originated with new value.

The Extended link opaque LSAs and the Extended link TLV are not scoped for multi-topology [RFC4915]. In multi-topology deployments [RFC4915], the Link-Overload sub-TLV advertised in an Extended Link opaque LSA corresponds to all the topologies which include the link. The receiver node SHOULD change the metric in the reverse direction for all the topologies which include the remote link and re-originate the router-LSA as defined in [RFC4915].

When the originator of the Link-Overload sub-TLV purges the Extended Link Opaque LSA or re-originates it without the Link-Overload sub-TLV, the remote node must re-originate the appropriate LSAs with the metric and TE metric values set to their original values.

5.2. Broadcast/NBMA links

Broadcast or NBMA networks in OSPF are represented by a star topology where the Designated Router (DR) is the central point to which all other routers on the broadcast or NBMA network logically connect. As a result, routers on the broadcast or NBMA network advertise only

their adjacency to the DR. Routers that do not act as DR do not form or advertise adjacencies with each other. For the Broadcast links, the MaxLinkMetric on the remote link cannot be changed since all the neighbors are on same link. Setting the link cost to MaxLinkMetric would impact paths going via all neighbors.

The node that has the link to be taken out of service MUST set metric of the link to MaxLinkMetric (0xffff) and re-originate the Router-LSA. The TE metric SHOULD be set to MAX-TE-METRIC(0xfffffffe) and the node SHOULD re-originate the corresponding TE Link Opaque LSAs. For a broadcast link, the two part metric as described in [RFC8042] is used. The node originating the Link-Overload sub-TLV MUST set the metric in the Network-to-Router Metric sub-TLV to MaxLinkMetric (0xffff) for OSPFv2 and OSPFv3 and re-originate the corresponding LSAs. The nodes that receive the two-part metric should follow the procedures described in [RFC8042]. The backward compatibility procedures described in [RFC8042] should be followed to ensure loop free routing.

5.3. Point-to-multipoint links

Operation for the point-to-multipoint links is similar to the point-to-point links. When a Link-Overload sub-TLV is received for a point-to-multipoint link the remote node MUST identify the neighbour which corresponds to the overloaded link and set the metric to MaxLinkMetric (0xffff). The remote node MUST re-originate the router-LSA with the changed metric for the correponding neighbor.

5.4. Unnumbered interfaces

Unnumbered interface do not have a unique IP address and borrow their address from other interfaces. [RFC2328] describes procedures to handle unnumbered interfaces in the context of the router-LSA. We apply a similar procedure to the Extended Link TLV advertising the Link-Overload sub-TLV in order to handle unnumbered interfaces. The link-data field in the Extended Link TLV includes the Local interface-id instead of the IP address. The Local/Remote Interface ID sub-TLV MUST be advertised when there are multiple parallel unnumbered interfaces between two nodes. One of the mechanisms to obtain the interface-id of the remote side are defined in [RFC4203].

5.5. Hybrid Broadcast and P2MP interfaces

Hybrid Broadcast and P2MP interfaces represent a broadcast network modeled as P2MP interfaces. [RFC6845] describes procedures to handle these interfaces. Operation for the Hybrid interfaces is similar to the P2MP interfaces. When a Link-Overload sub-TLV is received for a hybrid link, the remote node MUST identify the neighbor which

corresponds to the overloaded link and set the metric to MaxLinkMetric (0xffff). All the remote nodes connected to originator MUST re-originate the router-LSA with the changed metric for the neighbor.

6. Backward compatibility

The mechanisms described in the document are fully backward compatible. It is required that the node adverting the Link-Overload sub-TLV as well as the node at the remote end of the overloaded link support the extensions described herein for the traffic to diverted from the overloaded link. If the remote node doesn't support the capability, it will still use the overloaded link but there are no other adverse effects. In the case of broadcast links using two-part metrics, the backward compatibility procedures as described in [RFC8042] are applicable.

7. Applications

7.1. Pseudowire Services

Many service providers offer pseudo-wire services to customers using L2 circuits. The IGP protocol that runs in the customer network would also run over the pseudo-wire to create a seamless private network for the customer. Service providers want to offer overload functionality when the PE device is taken-out for maintenance. The provider should guarantee that the PE is taken out for maintenance only after the service is successfully diverted on an alternate path. There can be large number of customers attached to a PE node and the remote end-points for these pseudo-wires are spread across the service provider's network. It is a tedious and error-prone process to change the metric for all pseudo-wires in both directions. The link-overload feature simplifies the process by increasing the metric on the link in the reverse direction as well so that traffic in both directions is diverted away from the PE undergoing maintenance. The Link-Overload feature allows the link to be used as a last resort link so that traffic is not disrupted when alternative paths are not available.

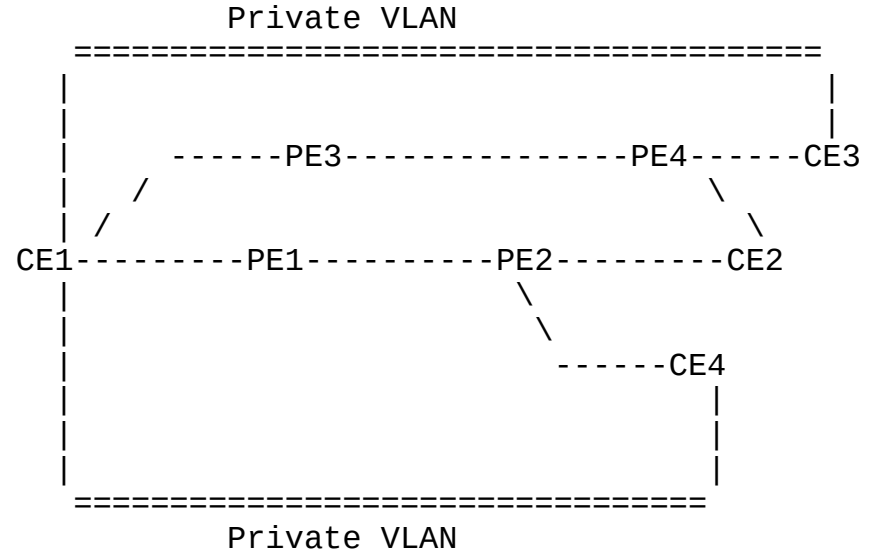

Figure 6: Pseudowire Services

In the example shown in Figure 6, when the PE1 node is going out of service for maintenance, service providers set the PE1 to overload state. The PE1 going in to overload state triggers all the CEs connected to the PE (CE1 in this case) to set their pseudowire links passing via PE1 to link-overload state. The mechanisms used to communicate between PE1 and CE1 is outside the scope of this document. CE1 sets the link-overload state on its private VLAN connecting CE3, CE2 and CE4 and changes the metric to MAX_METRIC and re-originates the corresponding LSA. The remote end of the link at CE3, CE2, and CE4 also set the metric on the link to MaxLinkMetric and the traffic from both directions gets diverted away from the pseudowires.

7.2. Controller based Traffic Engineering Deployments

In controller-based deployments where the controller participates in the IGP protocol, the controller can also receive the link-overload information as a warning that link maintenance is imminent. Using this information, the controller can find alternate paths for traffic which uses the affected link. The controller can apply various policies and re-route the LSPs away from the link undergoing maintenance. If there are no alternate paths satisfying the traffic engineering constraints, the controller might temporarily relax those constraints and put the service on a different path. Increasing the link metric alone does not specify the maintenance activity as the metric could increase in events such as LDP-IGP synchronisation. An explicit indication from the router using the link-overload sub-TLV is needed to inform the Controller or head-end routers.

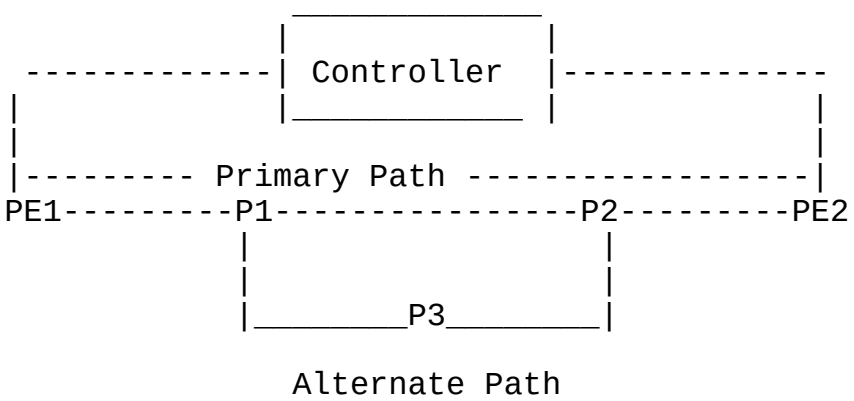

Figure 7: Controller based Traffic Engineering

In the above example, PE1->PE2 LSP is set-up to satisfy a constraint of 10 Gbps bandwidth on each link. The links P1->P3 and P3->P2 have only 1 Gbps capacity and there is no alternate path satisfying the bandwidth constraint of 10Gbps. When P1->P2 link is being prepared for maintenance, the controller receives the link-overload information, as there is no alternate path available which satisfies the constraints, the controller chooses a path that is less optimal and temporarily sets up an alternate path via P1->P3->P2. Once the traffic is diverted, the P1->P2 link can be taken out of service for maintenance/upgrade.

7.3. L3VPN Services and sham-links

Many service providers offer L3VPN services to customers and CE-PE links run OSPF [RFC4577]. When PE is taken out of service for maintenance, all the links on the PE can be set to link-overload state which will gurantee that the traffic to/from dual-homed CEs gets diverted. The interaction between OSPF and BGP is outside the scope of this document. [RFC6987] based mechanism with summaries and externals advertised with high metrics could also be used to achieve the same functionality when implementations support high metrics advertisement for summaries and externals.

Another useful usecase is when ISPs provide sham-link services to customers [RFC4577]. When PE goes out of service for maintenance, all sham-links on the PE can be set to link-overload state and traffic can be divered from both ends without having to touch the configurations on the remote end of the sham-links.

7.4. Hub and spoke deployment

OSPF is largely deployed in Hub and Spoke deployments with a large number of spokes connecting to the Hub. It is a general practice to deploy multiple Hubs with all spokes connecting to these Hubs to achieve redundancy. The [RFC6987] mechanism can be used to divert the spoke-to-spoke traffic from the overloaded hub router. The traffic that flows from spokes via the hub into an external network may not be diverted in certain scenarios.When a Hub node goes down for maintenance, all links on the Hub can be set to link-overload state and traffic gets divered from the spoke sites as well without having to make configuration changes on the spokes.

8. Security Considerations

This document does not introduce any further security issues other than those discussed in [RFC2328] and [RFC5340].

9. IANA Considerations

This specification updates one OSPF registry:

OSPFv2 Extended Link TLV Sub-TLVs

i) Link-Overload Sub-TLV - Suggested value 7

ii) Remote IPv4 Address Sub-TLV - Suggested value 8

iii) Local/Remote Interface ID Sub-TLV - Suggested Value 9

OSPFv3 Extended-LSA sub-TLV Registry

i) Link-Overload sub-TLV - suggested value 7

BGP-LS Link NLRI Registry [RFC7752]

i)Link-Overload TLV - Suggested 1101

10. Acknowledgements

Thanks to Chris Bowers for valuable inputs and edits to the document. Thanks to Jeffrey Zhang, Acee Lindem and Ketan Talaulikar for inputs. Thanks to Karsten Thomann for careful review and inputs on the applications where link-overload is useful.

11. References

11.1. Normative References

[I-D.ietf-ospf-ospfv3-lsa-extend]
Lindem, A., Mirtorabi, S., Roy, A., and F. Baker, "OSPFv3 LSA Extendibility", draft-ietf-ospf-ospfv3-lsa-extend-10 (work in progress), May 2016.

[RFC6845] Sheth, N., Wang, L., and J. Zhang, "OSPF Hybrid Broadcast and Point-to-Multipoint Interface Type", RFC 6845, DOI 10.17487/RFC6845, January 2013, <https://www.rfc-editor.org/info/rfc6845>.

[RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <https://www.rfc-editor.org/info/rfc7684>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

[RFC8042] Zhang, Z., Wang, L., and A. Lindem, "OSPF Two-Part Metric", RFC 8042, DOI 10.17487/RFC8042, December 2016, <https://www.rfc-editor.org/info/rfc8042>.

11.2. Informative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC4203] Kompella, K., Ed. and Y. Rekhter, Ed., "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, DOI 10.17487/RFC4203, October 2005, <https://www.rfc-editor.org/info/rfc4203>.

[RFC4577] Rosen, E., Psenak, P., and P. Pillay-Esnault, "OSPF as the Provider/Customer Edge Protocol for BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4577, DOI 10.17487/RFC4577, June 2006, <https://www.rfc-editor.org/info/rfc4577>.

[RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, DOI 10.17487/RFC4915, June 2007, <https://www.rfc-editor.org/info/rfc4915>.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[RFC5817] Ali, Z., Vasseur, JP., Zamfir, A., and J. Newton, "Graceful Shutdown in MPLS and Generalized MPLS Traffic Engineering Networks", RFC 5817, DOI 10.17487/RFC5817, April 2010, <https://www.rfc-editor.org/info/rfc5817>.

[RFC6987] Retana, A., Nguyen, L., Zinin, A., White, R., and D. McPherson, "OSPF Stub Router Advertisement", RFC 6987, DOI 10.17487/RFC6987, September 2013, <https://www.rfc-editor.org/info/rfc6987>.

Authors' Addresses

Shraddha Hegde
Juniper Networks, Inc.
Embassy Business Park
Bangalore, KA 560093
India

Email: shraddha@juniper.net

Pushpasis Sarkar
Individual

Email: pushpasis.ietf@gmail.com

Hannes Gredler
Individual

Email: hannes@gredler.at

Mohan Nanduri
ebay Corporation
2025 Hamilton Avenue
San Jose, CA 98052
US

Email: mnanduri@ebay.com

Luay Jalil
Verizon

Email: luay.jalil@verizon.com