OSPF Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 1, 2018

A. Atlas
S. Hegde
C. Bowers
Juniper Networks
J. Tantsura
Individual
Z. Li
Huawei Technologies

June 30, 2017

OSPF Extensions to Support Maximally Redundant Trees
draft-ietf-ospf-mrt-03

Abstract

This document specifies extensions to OSPF to support the distributed computation of Maximally Redundant Trees (MRT). Some example uses of the MRTs include IP/LDP Fast-Reroute and global protection or live-live for multicast traffic. The extensions indicate what MRT profile(s) each router supports. Different MRT profiles can be defined to support different uses and to allow transitioning of capabilities. An extension is introduced to flood MRT-Ineligible links, due to administrative policy.

The need for a mechanism to allow routers to advertise a worst-case FIB compute/install time is well understood for controlling convergence. This specification introduces the Controlled Convergence TLV to be carried in the Router Information LSA.

Status of This Memo

Table of Contents

1. Introduction

This document describes the OSPF extensions necessary to support the architecture that defines how IP/LDP Fast-Reroute can use MRTs [RFC7812]. At least one common standardized algorithm (such as the lowpoint algorithm explained and fully documented in [RFC7811]) is required so that the routers supporting MRT computation consistently

compute the same MRTs. MRT can also be used to protect multicast traffic via either global protection or local protection.[I-D.atlas-rtgwg-mrt-mc-arch]

IP/LDP Fast-Reroute using MRTs can provide 100% coverage for link and node failures in an arbitrary network topology where the failure doesn't split the network. It can also be deployed incrementally inside an OSPF area; an MRT Island is formed of connected supporting routers and the MRTs are computed inside that island.

In the default MRT profile, a supporting router both computes the MRTs and creates the necessary transit forwarding state necessary to provide the two additional forwarding topologies, known as MRT-Blue and MRT-Red.

2. 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]

3. Terminology

For ease of reading, some of the terminology defined in [RFC7812] is repeated here.

network graph: A graph that reflects the network topology where all links connect exactly two nodes and broadcast links have been transformed into the standard pseudo-node representation.

Redundant Trees (RT):   A pair of trees where the path from any node X to the root R along the first tree is node-disjoint with the path from the same node X to the root along the second tree. These can be computed in 2-connected graphs.

Maximally Redundant Trees (MRT):   A pair of trees where the path from any node X to the root R along the first tree and the path from the same node X to the root along the second tree share the minimum number of nodes and the minimum number of links. Each such shared node is a cut-vertex. Any shared links are cut-links. Any RT is an MRT but many MRTs are not RTs.

MRT Island: From the computing router, the set of routers that support a particular MRT profile and are connected via MRT-eligible links.

GADAG: Generalized Almost Directed Acyclic Graph - a graph that is the combination of the ADAGs of all blocks. Transforming a network graph into a GADAG is part of the MRT algorithm.

MRT-Red: MRT-Red is used to describe one of the two MRTs; it is used to described the associated forwarding topology and MT-ID. Specifically, MRT-Red is the decreasing MRT where links in the GADAG are taken in the direction from a higher topologically ordered node to a lower one.

MRT-Blue: MRT-Blue is used to describe one of the two MRTs; it is used to described the associated forwarding topology and MT-ID. Specifically, MRT-Blue is the increasing MRT where links in the GADAG are taken in the direction from a lower topologically ordered node to a higher one.

4. Overview of OSPF Extensions for MRT

There are two separate aspects that need to be advertised in OSPF. Both derive from the need for all routers supporting an MRT profile to compute the same pair of MRTs to each destination. By executing the same algorithm on the same network graph, distributed routers will compute the same MRTs. Convergence considerations are discussed in [RFC7812].

The first aspect that must be advertised is which MRT profile(s) are supported and the associated GADAG Root Selection Priority. The second aspect that must be advertised is any links that are not eligible, due to administrative policy, to be part of the MRTs. This must be advertised consistently across the area so that all routers in the MRT Island use the same network graph.

4.1. Supporting MRT Profiles

An MRT Profile defines the exact MRT Algorithm, the MRT-Red LDP MT-ID, the MRT-Blue LDP MT-ID, and the forwarding mechanisms supported for the transit MRT-Red and MRT-Blue forwarding topologies. Finally, the MRT Profile defines exact behavioral rules such as:

o how reconvergence is handled,

o inter-area forwarding behavior,

A router that advertises support for an MRT Profile MUST provide the specified forwarding mechanism for its MRT-Red and MRT-Blue forwarding topologies. A router that advertises support for an MRT Profile MUST implement an algorithm that produces the same set of MRT-Red and MRT-Blue next-hops for its MRT-Red and MRT-Blue

topologies as is provided by the algorithm specified in the MRT Profile.

A router MAY indicate support for multiple MRT Profiles. A router computes its local MRT Island for each separate MRT Profile that the router supports. Supporting multiple MRT Profiles also provides a mechanism for transitioning from one profile to another. Different uses of MRT forwarding topologies may behave better on different MRT profiles.

The default MRT Profile is defined in [RFC7812]. Its behavior is intended to support IP/LDP unicast and multicast fast-reroute.

4.2. GADAG Root Selection

One aspect of the MRT algorithms is that the selection of the GADAG root can affect the alternates and the traffic through that GADAG root. Therefore, it is important to provide an operator with control over which router will play the role of GADAG root. A measure of the centrality of a node may help determine how good a choice a particular node is.

The GADAG Root Selection Policy (defined as part of an MRT profile) may make use of the GADAG Root Selection Priority value advertised in the MRT Profile TLV of the Router Information LSA. For example, the GADAG Root Selection Policy for the default MRT profile is the following: Among the routers in the MRT Island and with the highest priority advertised, an implementation MUST pick the router with the highest Router ID to be the GADAG root.

4.3. Triggering an MRT Computation

When an MRT Computation is triggered, it is triggered for a given MRT Profile in a given area. First, the associated MRT Island is determined. Then, the GADAG Root is selected. Finally, the actual MRT algorithm is run to compute the transit MRT-Red and MRT-Blue topologies. Additionally, the router MAY choose to compute MRT-FRR alternates or make other use of the MRT computation results.

Prefixes can be attached and detached and have their associated MRT-Red and MRT-Blue next-hops computed without requiring a new MRT computation.

5. MRT Profile TLV in Router Information LSA

A router may advertise an MRT Profile TLV to indicate support for one or more MRT Profiles. The MRT Profile TLV is advertised within the

OSPF router information LSA which is defined for both OSPFv2 and OSPFv3 in [RFC7770]. The RI LSA MUST have area scope.

Note that the presence of the MRT Profile TLV indicates support for a given MRT profile in the default topology (MT-ID = 0). The extensions in this document do not define a method to advertise support for MRT profiles in topologies with non-zero MT-ID.

```
 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              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Profile ID(1) |GADAG Priority |         Reserved              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         ..............                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Profile ID(n) |GADAG Priority |         Reserved              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

TYPE:  TBA-MRT-OSPF-1 (To Be Allocated by IANA)
LENGTH: 4 * (number of Profiles)
Profile ID :  1 byte
GADAG Priority: 1 byte
```

MRT Profile TLV in Router Information LSA

Each Profile ID listed indicates support for a given MRT Profile, as defined in [RFC7812]. A Profile ID value of 0 corresponds to the Default MRT profile.

The GADAG Priority is the GADAG Root Selection Priority associated with the advertising router in the MRT Island for the associated MRT Profile, as indicated by the Profile ID. An implementation SHOULD send a default value of 128 for the GADAG Root Selection Priority if another value is not explicitly configured.

The length of this TLV depends on the number of MRT profiles supported. The ordering of the profiles inside the TLV is not significant. Multiple appearances of this TLV is not an error.

An advertising router MUST NOT advertise the same Profile ID multiple times in one or more TLVs. If a receiving router receives multiple appearances of the same Profile ID for the same router, it MUST treat the advertising router as NOT supporting the MRT Profile associated

with that Profile ID. This is the case even if the multiple appearances of the same Profile ID have the same GADAG Priority values. However, other Profile IDs advertised by the same advertising router that are not repeated should continue to be honored by the receiving router. The receiving router SHOULD also log an error message regarding the multiple appearances of the same Profile ID for the same router.

6. Advertising MRT-ineligible links for MRT

Due to administrative policy, some otherwise eligible links in the network topology may need to be excluded from the network graph upon which the MRT algorithm is run. Since the same network graph must be used across the area, it is critical for OSPF to flood which links to exclude from the MRT calculation. This is done by introducing a new MRT-Ineligible Link sub-TLV. For OSPFv2, this sub-TLV is carried in the Extended Link TLV defined in [I-D.ietf-ospf-prefix-link-attr]. For OSPFv3, this sub-TLV is carried in the Router-Link TLV defined in [I-D.ietf-ospf-ospfv3-lsa-extend].

If a link is marked by administrative policy as MRT-Ineligible, then a router MUST flood the OSPFv2 Extended Link TLV (or OSPFv3 Router-Link TLV) for that link, including the MRT-Ineligible Link sub-TLV. The OSPVv2 Extended Link TLV and OSPFv3 Router-Link TLV have area wide scope.

Note that a router that advertises support for MRT with the MRT Profile TLV MUST also support receipt of the MRT-Ineligible Link sub-TLVs. This ensures that all routers participating in a given MRT Island have the same view of the links included in the MRT Island.

6.1. MRT-Ineligible Link sub-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              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

TYPE:  TBA-MRT-OSPF-2 in OSPFv2 Extended Link TLV
       TBA-MRT-OSPF-3 in OSPFv3 Router-Link TLV
       (To Be Allocated by IANA)
LENGTH: 0
```

MRT-Ineligible Link sub-TLV

This zero-length sub-TLV can appear in the OSPFv2 Extended Link TLV or the OSPFv3 Router-Link TLV. Its presence indicates that the associated link MUST NOT be used in the MRT calculation for all profiles.

7. Worst-Case Network Convergence Time

As part of converging the network after a single failure, Section 12.2 of [RFC7812] describes the need to wait for a configured or advertised period for all routers to be using their new SPTs. Similarly, some proposals to avoid micro-forwarding loops during convergence[RFC5715] require determining the maximum among all routers in the area of the worst-case route computation and FIB installation time. More details on the specific reasoning and need for flooding it are given in [I-D.atlas-bryant-shand-lf-timers].

```
    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            |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |         Reserved              |    FIB compute/install time   |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

   TYPE:  TBA-MRT-OSPF-4 (To Be Allocated by IANA)
   LENGTH: 4
   FIB compute/install time:  This is the worst-case time the router
       may take to compute and install all OSPF routes in the area
       after a change to a stable network.  The value is
       in milliseconds.


        Controlled Convergence TLV in Router Information LSA
```

The Controlled Convergence TLV is carried in the Router Information LSA and flooded with area-wide scope. The FIB compute/install time value sent by a router SHOULD be an estimate taking into account network scale or real-time measurements, or both. Advertisements SHOULD be dampened to avoid frequent communication of small changes in the FIB compute/install time.

A router receiving the Controlled Convergence TLV SHOULD estimate the network convergence time as the maximum of the FIB compute/install times advertised by the routers in an area, including itself. In order to account for routers that do not advertise the Controlled Convergence TLV, a router MAY use a locally configured minimum network convergence time as a lower bound on the computed network

convergence time. A router MAY use a locally configured maximum network convergence time as an upper bound on the computed network convergence time.

8. Backwards Compatibility

The MRT Profile TLV, the MRT-Ineligible Link sub-TLV, the OSPFv3 MRT-Ineligible Link sub-TLV, and the Controlled Convergence TLV are defined in this document. A router that does not understand the MRT Profile TLV will ignore it. A router that does not advertise an MRT Profile TLV with a Profile ID may do so either because it doesn't understand the MRT Profile TLV, or because it understands these extensions, but chooses not to advertise support for any MRT profile. Routers that support the MRT Profile TLV will treat either case in the same manner, by excluding the router not advertising the MRT Profile from the particular MRT Island.

The MRT-Ineligible Link sub-TLVs will be ignored by a router that doesn't understand MRT, and a router supporting MRT must support receipt of the MRT-Ineligible Link sub-TLVs.

Finally, applications that utilize the Controlled Convergence TLV can use local configuration to account for routers that do not understand the Controlled Convergence TLV.

8.1. Handling MRT Capability Changes

When a router that is running a version of software supporting MRT is downgraded to software that does not support MRT, it is important that the routers still running MRT do not continue to use the Router Information LSA (RI LSA) containing the MRT Profile TLV advertised by the downgraded router before the downgrade. As long as the downgraded router supports Opaque LSAs, the downgraded router will purge the old RI LSA containing the MRT Profile TLV that it originated before the downgrade. This will occur when the downgraded router receives the self-originated RI LSA after restarting, as described in Section 13.4 and 14.1 of [RFC2328]. This behavior is clearly required when the downgraded router supports the RI LSA.

It is also reasonable to expect this behavior even when the software on the downgraded router does not understand the RI LSA. Although this precise behavior is not explicitly described in [RFC2328] , it is reasonable to infer from the documents. As long as the downgraded router supports Opaque LSAs, it is required to flood link-state type 10 (area-local scope) Opaque LSAs, even those that it does not understand [RFC5250]. So, when a restarting router receives a self-originated link-state type 10 Opaque LSA whose Option Type it does not recognize, it can (in principle) flood it or purge it. Purging

an unknown self-originated Opaque LSA is the most reasonable thing to do.

9. Implementation Status

[RFC Editor: please remove this section prior to publication.]

Please see [RFC7812] for details on implementation status.

10. Security Considerations

This OSPF extension is not believed to introduce new security concerns. It relies upon the security architecture already provided for Router LSAs and Router Information LSAs.

11. Acknowledgements

The authors would like to thank Anil Kumar SN for his suggestions and review.

12. IANA Considerations

12.1. MRT Profile and Controlled Convergence TLVs

IANA is requested to allocate values for the following OSPF Router Information TLV Types [RFC7770]: MRT Profile TLV (TBA-MRT-OSPF-1), and Controlled Convergence TLV (TBA-MRT-OSPF-4). The requested entries in the OSPF Router Information (RI) TLVs registry are shown below.

Type Value	Capabilities	Reference
TBA-MRT-OSPF-1	MRT Profile TLV	[This draft]
TBA-MRT-OSPF-4	Controlled Convergence TLV	[This draft]

12.2. MRT-Ineligible Link sub-TLV

IANA is requested to allocate a value from the OSPF Extended Link TLV sub-TLV registry defined in [I-D.ietf-ospf-prefix-link-attr] for the MRT-Ineligible Link sub-TLV (TBA-MRT-OSPF-2). The OSPF Extended Link TLV sub-TLV registry after implementing the above request is shown below.

Value	Description	Reference
0	Reserved	[prefix-link-attr-draft]
TBA-MRT-OSPF-2	MRT-Ineligible Link sub-TLV	[This draft]
2-32767	Unassigned	[prefix-link-attr-draft]
32768-33023	Reserved for Experimental Use	[prefix-link-attr-draft]
33024-65535	Reserved	[prefix-link-attr-draft]

IANA is requested to allocate a value from the OSPFv3 Extended-LSA sub-TLV registry [I-D.ietf-ospf-ospfv3-lsa-extend] for the MRT-Ineligible Link sub-TLV (TBA-MRT-OSPF-3). The OSPFv3 Extended-LSA sub-TLV registry has not yet been created by IANA.

13. References

13.1. Normative References

[I-D.ietf-ospf-ospfv3-lsa-extend]
Lindem, A., Roy, A., Goethals, D., Vallem, V., and F. Baker, "OSPFv3 LSA Extendibility", draft-ietf-ospf-ospfv3-lsa-extend-14 (work in progress), April 2017.

[I-D.ietf-ospf-prefix-link-attr]
Psenak, P., Gredler, H., rjs@rob.sh, r., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", draft-ietf-ospf-prefix-link-attr-13 (work in progress), August 2015.

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

[RFC5250] Berger, L., Bryskin, I., Zinin, A., and R. Coltun, "The OSPF Opaque LSA Option", RFC 5250, DOI 10.17487/RFC5250, July 2008, <http://www.rfc-editor.org/info/rfc5250>.

[RFC7770] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 7770, DOI 10.17487/RFC7770, February 2016, <http://www.rfc-editor.org/info/rfc7770>.

[RFC7811] Enyedi, G., Csaszar, A., Atlas, A., Bowers, C., and A. Gopalan, "An Algorithm for Computing IP/LDP Fast Reroute Using Maximally Redundant Trees (MRT-FRR)", RFC 7811, DOI 10.17487/RFC7811, June 2016, <http://www.rfc-editor.org/info/rfc7811>.

[RFC7812] Atlas, A., Bowers, C., and G. Enyedi, "An Architecture for IP/LDP Fast Reroute Using Maximally Redundant Trees (MRT-FRR)", RFC 7812, DOI 10.17487/RFC7812, June 2016, <http://www.rfc-editor.org/info/rfc7812>.

13.2. Informative References

[I-D.atlas-bryant-shand-lf-timers]
K, A. and S. Bryant, "Synchronisation of Loop Free Timer Values", draft-atlas-bryant-shand-lf-timers-04 (work in progress), February 2008.

[I-D.atlas-rtgwg-mrt-mc-arch]
Atlas, A., Kebler, R., Wijnands, I., Csaszar, A., and G. Envedi, "An Architecture for Multicast Protection Using Maximally Redundant Trees", draft-atlas-rtgwg-mrt-mc-arch-02 (work in progress), July 2013.

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

[RFC5715] Shand, M. and S. Bryant, "A Framework for Loop-Free Convergence", RFC 5715, DOI 10.17487/RFC5715, January 2010, <http://www.rfc-editor.org/info/rfc5715>.

Authors' Addresses

Alia Atlas
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: akatlas@juniper.net

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

Email: shraddha@juniper.net

Chris Bowers
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
USA

Email: cbowers@juniper.net

Jeff Tantsura
Individual
USA

Email: jefftant.ietf@gmail.com

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

Email: lizhenbin@huawei.com