Routing area
Internet-Draft
Intended status: Standards Track
Expires: April 20, 2016

S. Hegde
P. Sarkar
Juniper Networks, Inc.
October 18, 2015

Micro-loop avoidance using SPRING
draft-hegde-rtgwg-microloop-avoidance-using-spring-00

Abstract

When there is a change in network topology either due to a link going down or due to a new link addition, all the nodes in the network need to get the complete view of the network and re-compute the routes. There will generally be a small time window when the forwarding state of each of the nodes is not synchronized. This can result in transient loops in the network, leading to dropped traffic due to over-subscription of links. Micro-looping is generally more harmful than simply dropping traffic on failed links, because it can cause control traffic to be dropped on an otherwise healthy link involved in micro-loop. This can lead to cascading adjacency failures or network meltdown.

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

Table of Contents

1. Introduction

Micro-loops are transient loops that occur during the period of time when some nodes have become aware of a topology change and have changed their forwarding tables in response, but slow routers have not yet modified their forwarding tables. This document provides

mechanisms to prevent micro-loops in the network in the event of link up/down or metric change.The micro-loop prevention mechanism uses the basic principles of near-side tunnelling as described in [RFC5715] sec 6.2.

Micro-loops can be formed involving the PLRs or nodes which are not directly connected to the link/node going down. The nodes which are not directly connected to the node/link going down/up are referred to as remote nodes. The micro-loop prevention mechanism described in this document prevents possible micro-loops involving the remote nodes. A new sub-tlv is defined in ISIS router capability TLV [RFC4971] and OSPF router capability TLV [RFC4970] for discovering support of this feature. The details are described in Section 4. The operational procedures for micro-loop prevention are described in Section 3.

2. Procedures for Micro-loop prevention

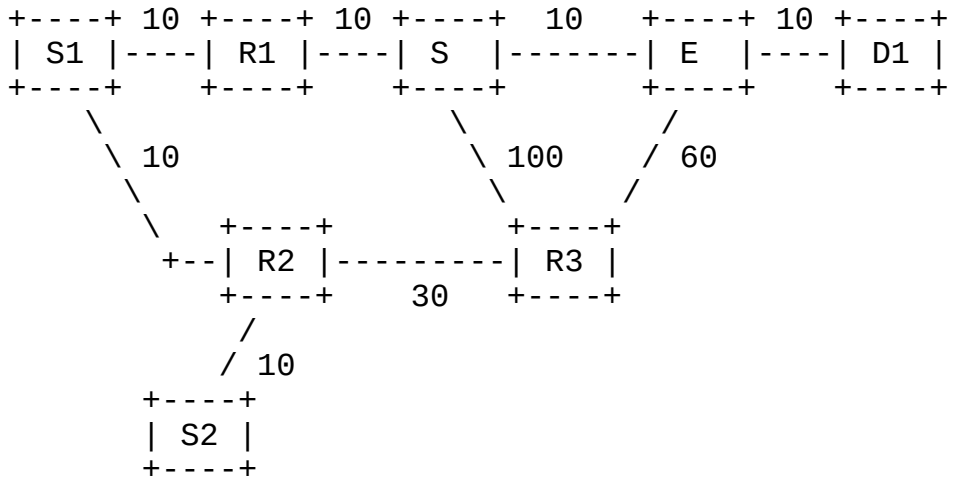

Figure 1: Sample Network

The topology shown in figure 1 illustrates a sample network topology where micro-loops can occur. The symmetric link metrics are shown in the diagram above. The traffic from S1 to D1 takes the path S1->R1->S->E->D1 and traffic from S2 takes the path S2->R2->S1->R1->S->E->D1 in normal operation. When the S->E link goes down, traffic can loop between S1->R2 when the FIB on S1 reflects the shortest path to D1 after the failure and the FIB on R2 reflects the shortest path to D1 before the failure. The mechanisms described in [I-D.litkowski-rtgwg-uloop-delay] do not address micro-loops involving nodes that are not directly attached to the link that has just gone down or come up. For example when S->E link goes down,

S and E are the Point of Local Repair (PLR) and micro-loops formed between S1 and R2 are not handled.

The basic principle of the solution is to send the traffic on tunnelled paths for a certain time period until all the nodes in the network process the event and update their forwarding plane. When the link S->E goes down, all the nodes in the network tunnel the traffic to the nearest PLR. The PLR S maintains the FRR ([RFC5286]) backup path until all other nodes in the network converge and forwards the traffic to the affected destinations via the back-up path. This document assumes 100% backup coverage for the destinations via various FRR mechanisms. This document describes the procedures corresponding to the traffic flow from sources (S nodes) to the destination nodes (D nodes). The procedures equally apply to the D nodes being source and S nodes being destination.

As soon as a node learns of the topology change, it modifies its FIB to use loop-free tunnelled paths for the affected traffic, and it starts a "convergence delay timer". When the "convergence delay timer" expires, the node modifies its FIB to use the SPF path based on the changed topology. The use of tunnelled paths during the convergence period ensures that (barring other topology changes) all traffic affected by the topology change travels on a loop-free path.

After all the nodes in the network converge to actual SPF path,PLR converges to SPF path and updates the FIB. This micro-loop prevention mechanism delays the time it takes for routing to converge to the optimal paths in the new topology by a factor of 3 but the convergence time is deterministic and completely avoids micro-loops.

In principle, near-side tunnelling could be accomplished using labels distributed via LDP. However, since the application requires that any given router have the potential to create a tunnel to nearly every other router in the IGP domain, a large number of targeted LDP sessions would be needed to learn the FEC-label bindings distributed by the PLRs. SPRING provides a more efficient method for distributing shortest path labels for this application, since any router can compute the locally significant FEC-label bindings for any other router without the need for targeted LDP sessions.

3. Detailed Solution based on SPRING

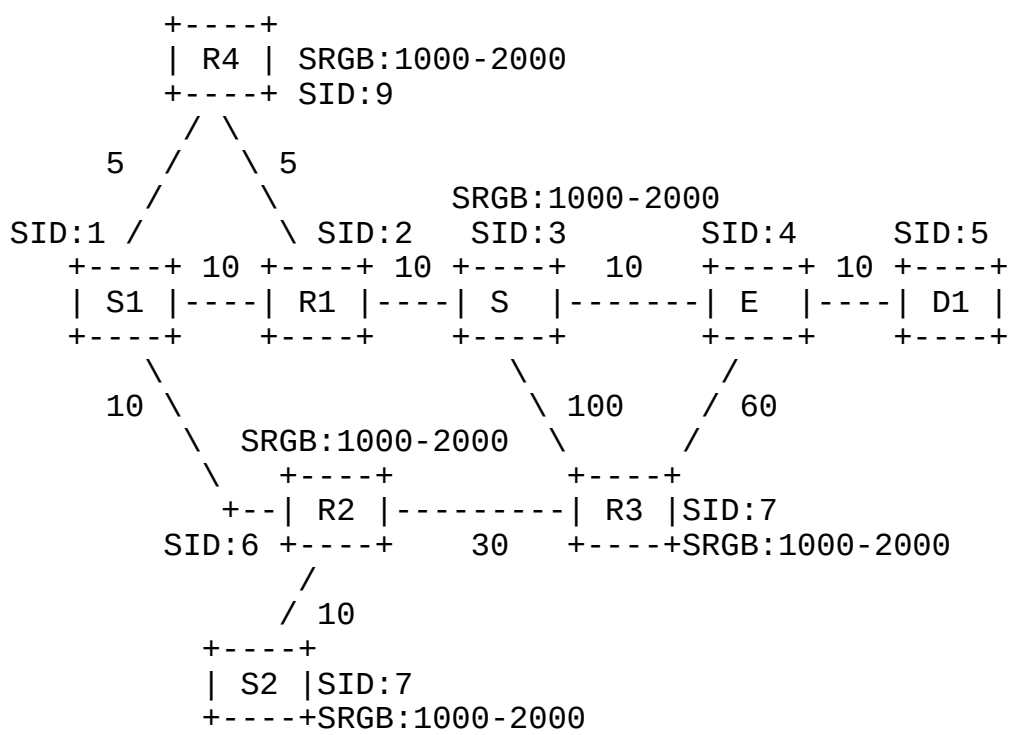

Figure 2: Sample SR Network

The above sample topology is provided with basic SPRING configurations of SRGB and the indices corresponding to each node. Each node has an SRGB 1000-2000 configured on the node. Same SRGB on all nodes is used for simplifying the example and the procedures are equally applicable when there is different SRGB configured on multiple nodes. Each node is provisioned with a MAX_CONVERGENCE_DELAY value that corresponds to its RIB to FIB convergence time. The information for support of the micro-loop prevention feature and the MAX_CONVERGENCE_DELAY value are flooded across the IGP domain (ISIS level/OSPF area). Each node in the IGP domain sets the MAX_CONVERGENCE_DELAY to the maximum of the values received in the domain.

3.1. Link-down event

When the S->E link goes down, all the nodes in the network receive the event via IGP database flooding. Each node supporting the micro-loop prevention mechanism specified in this document SHOULD perform the steps below.

1. The PLRs (S and E) perform FRR local repair for destinations affected by the failure of the link. Each computing node identifies the destinations affected by the topology change.In

the example above, the destination D1 is affected by S->E link down for nodes S1,R1,R2, and R4. For S2, although the path to D1 changes there is no change in the immediate next-hop and hence its not necessary for S2 to perform any specific actions to prevent micro-loops.

2. For each affected destination, identify the nearest PLR advertising the change. The link-down event is advertised by both S and E. S is the nearest PLR for the nodes S1,R1,R2, and R4.

3. Let the S->E link down event occurs at time T0.

4. Start a timer T1 = max (all MAXIMUM_CONVERGENCE_DELAY) at all non-PLR nodes with affected destinations.

5. Start a timer T2 = 2 * T1 at the PLR.

6. For IP routes, modify the FIB for the affected destinations so that the nearest PLR's node-sid is pushed on the packet's label stack. For MPLS ingress and transit routes, modify the FIB for the affected destinations with a two label stack, the inner label corresponding to the destination and the outer label corresponding to the nearest PLR.

7. In the case of ECMP paths to the nearest PLR, both tunnelled paths are used. S1 has ECMP paths to the destination D1 and both the paths are impacted. Both the paths are modified to carry two label stacks containing the nearest PLR on top and the destination label at the bottom.

8. After the expiry of timer T1 all the non-PLR nodes modify their FIBs to use the shortest path as computed by the IGP, and they no longer push the node-SID of the nearest PLR on the packets.

9. After the expiry of T2, the PLR converges and updates the FIB to represent shortest path.

The ingress MPLS routes at various nodes for destination D1 at specified time intervals is mentioned below.

Node	Before T0	T0-T1	T1-T2	After T2
S1	Push 1005, Fwd to R1	Push 1005, 1003(top), Fwd to R1	Push 1005, Fwd to R2	Push 1005, Fwd to R2
	Push 1005, Fwd to R4	Push 1005, 1003(top), Fwd to R4		
S2	Push 1005, Fwd to R2	Push 1005, Fwd to R2	Push 1005, Fwd to R2	Push 1005, Fwd to R2
R1	Push 1005, Fwd to S	Push 1005, Fwd to S	Push 1005, Fwd to R4	Push 1005, Fwd to R4
			Push 1005, Fwd to S1	Push 1005, Fwd to S1
R2	Push 1005, Fwd to S1	Push 1005, 1003(top), Fwd to S1	Push 1005, Fwd to R3	Push 1005, Fwd to R3
R3	Push 1005, Fwd to E	Push 1005, 1003(top), Fwd to E	Push 1005, Fwd to E	Push 1005, Fwd to E
R4	Push 1005, Fwd to R1	Push 1005, 1003(top), Fwd to R1	Push 1005, Fwd to S1	Push 1005, Fwd to S1
S	Push 1005, Fwd to E	Push 1005, Fwd to R3 *	Push 1005, Fwd to R3 *	Push 1005, Fwd to R1
	Push 1005, Fwd to R3 *			Push 1005, Fwd to R3 *
E	Pop, Fwd to D1	Pop, Fwd to D1	Pop, Fwd to D1	Pop, Fwd to D1

* - Indicates backup path.

Figure 3: Sample MPLS ingress RIB

The corresponding MPLS transit routes at various nodes at specified time interval is shown below.

Node	Incoming Label	Before T0	T0-T1	T1-T2	After T2
S1	1005	Push 1005, Fwd to R1	Push 1005, 1003(top), Fwd to R1	Push 1005, Fwd to R2	Push 1005, Fwd to R2
		Push 1005, Fwd to R4	Push 1005, 1003(top), Fwd to R4		
	1003	Push 1003, Fwd to R1	Push 1003, Fwd to R1	Push 1003, Fwd to R2	Push 1003, Fwd to R2
S2	1005	Push 1005, Fwd to R2	Push 1005, Fwd to R2	Push 1005, Fwd to R2	Push 1005, Fwd to R2
	1003	Push 1003, Fwd to R1	Push 1003, Fwd to R1	Push 1003, Fwd to R2	Push 1003, Fwd to R2
R1	1005	Push 1005, Fwd to S	Push 1005, Fwd to S	Push 1005, Fwd to R4	Push 1005, Fwd to R4
				Push 1005, Fwd to S1	Push 1005, Fwd to S1
	1003	Push 1003, Fwd to S	Push 1003, Fwd to S	Push 1003, Fwd to S	Push 1003, Fwd to S
R2	1005	Push 1005, Fwd to	Push 1005, 1003(top), Fwd to S1	Push 1005, Fwd to R3	Push 1005, Fwd to R3

		S1			
	1003	Push 1003, Fwd to S1	Push 1003, Fwd to S1	Push 1003, Fwd to S1	Push 1003, Fwd to S1
R3	1005	Push 1005, Fwd to E	Push 1005, 1003(top), Fwd to E	Push 1005, Fwd to E	Push 1005, Fwd to E
	1003	Push 1003, Fwd to R2	Push 1003, Fwd to R2	Push 1003, Fwd to R2	Push 1003, Fwd to R2
R4	1005	Push 1005, Fwd to R1	Push 1005, 1003(top), Fwd to R1	Push 1005, Fwd to S1	Push 1005, Fwd to S1
	1003	Push 1003, Fwd to R1	Push 1003, Fwd to R1	Push 1003, Fwd to R1	Push 1003, Fwd to R1
S	1005	Push 1005, Fwd to E	Push 1005, Fwd to R3 *	Push 1005, Fwd to R3 *	Push 1005, Fwd to R1
		Push 1005, Fwd to R3 *			Push 1005, Fwd to R3 *
	1003	--	--	--	--
E	1005	Pop, Fwd to D1	Pop, Fwd to D1	Pop, Fwd to D1	Pop, Fwd to D1

* - Indicates backup path.

Figure 4: Sample MPLS transit RIB

3.2. Link-up event

When a new-link is added to the network, the PLR needs to update the FIB before it announces the change. First the PLR converges, updates the FIB as per the new-link based topology and then announces the new-link addition to the rest of the network. The other network nodes SHOULD follow the procedure exactly same as described in sec 3.1. They SHOULD update their FIB to tunnel the traffic to the closest node corresponding to the change.After MAX_CONVERGENCE_DELAY the nodes SHOULD update the FIB with the shortest path next-hops.

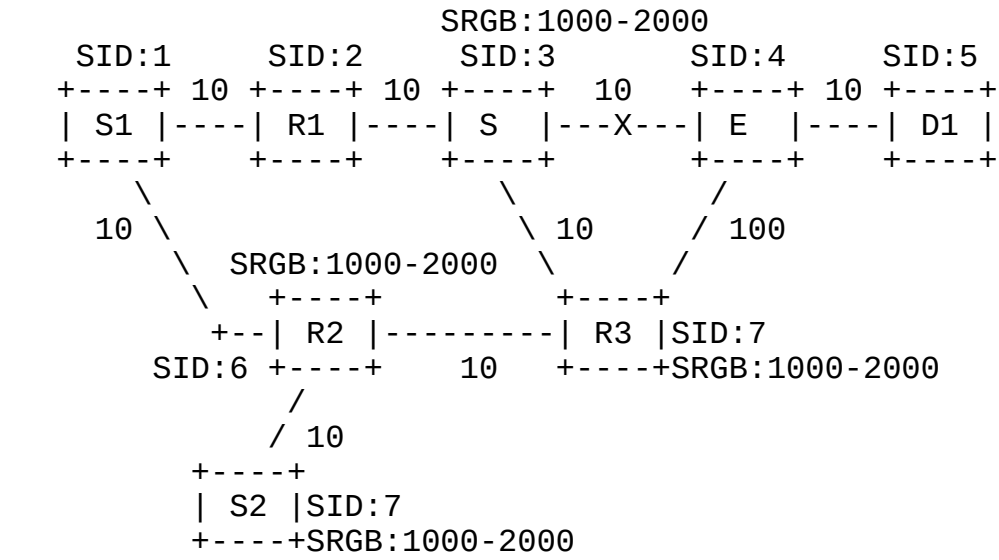

Figure 5: Sample SR Network

In the figure above, when the S->E link is added (or restored back),

1. PLR S processes the event and programs the FIB with new path for the affected destinations.

2. PLR delays flooding the event for MAX_CONVERGENCE_DELAY interval. This step prevents possible local micro-loop between S and R3.

3. Once PLR floods the event, non PLR nodes in the network identify the destinations affected by the database change. This is done by SPF computation and examining the next-hop change. The destination D1 is affected by S->E link up for nodes S1, R1, R2 and R3.

4. For each affected destination, identify the nearest PLR advertising the change. The link-up event is advertised by both

S and E. S is the nearest PLR for the nodes S1,R1,R2 and R3. When there are ECMP paths to the destination and a new ECMP path is added, the new ECMP path follows the micro-loop prevention mechanisms and tunnels the traffic towards nearest PLR.

5. Start a timer T3 = max (all MAXIMUM_CONVERGENCE_DELAY) at all non-PLR nodes.

6. For IP routes, update the FIB for the affected destinations so that the nearest PLR's node-sid is pushed on the packet's label stack. For MPLS ingress and transit router update the path with two label stack, the inner label corresponding to the destination and the outer label corresponding to the nearest PLR. This step prevents the possible remote micro-loop between S1 and R2.

7. After the expiry of timer T3 all the non-PLR nodes perform global convergence and update the FIB to represent the shortest path.

Other management events like metric change are handled similar to the link-down/link-up cases for metric increase/metric decrease cases respectively.

3.3. Computation of nearest PLR

When a network event is received by a node via the IGP database change notification, a node has to compute the nearest PLR corresponding to that advertisement. The first database change advertisement may be received from any of the PLRs, nearest or farthest.

3.3.1. Link down event

When a link goes down, IGPs generate a fresh LSP/Router LSA with the affected link removed. The computing node has to identify the missing link by walking over the LSP/LSA and compare the contents with an older version. Once the affected link is identified, the cost to reach both ends of the link should be examined. The nearest PLR is chosen based on the cost to reach the ends.

3.3.2. Node down event

When a node goes down, it is identified by the neighbouring nodes via link-down event. the neighbouring routers generate a fresh LSP/ Router LSA with the affected link removed. The computing node has to identify the missing link by walking over the LSP/LSA and compare the contents with an older version. Once the affected link is identified, the cost to reach both ends of the link should be

examined. The nearest PLR is chosen based on the cost to reach the ends.

When an advertisement from the farthest node is received before the nearest node, it is possible that the node that went down is chosen as the nearest PLR, as the node that went down might be still lingering in the database. In such cases node protection mechanisms for the deceased node at the previous-hop should prevent traffic loss. The details of such a mechanism is outside the scope of this document.

3.4. Handling multiple network events

[RFC5715] sec 6.2 describes mechanisms to handle the SRLG failures. If the received failure advertisement is part of an SRLG advertised in the IGP TE advertisement, the links on the path sharing same SRLG are identified and the tunnel is built with multiple label stack corresponding to nearest PLR of each SRLG member.

When a failure is received, and the failure does not belong to the same SRLG as the already on-going micro-loop prevention, the micro-loop prevention procedures MUST be aborted and the normal convergence procedures SHOULD be followed.

3.4.1. Handling SRLG failures

Consider a sample network as shown above with S->E and S1->R1 belonging to same SRLG group. The symmetric link metrics are shown in the figure and the SRGB is 1000-2000 on all nodes. When the S->E link goes down, all the links belonging to the same SRLG are considered to be down and the route is modified to carry multiple node-sids along the path.

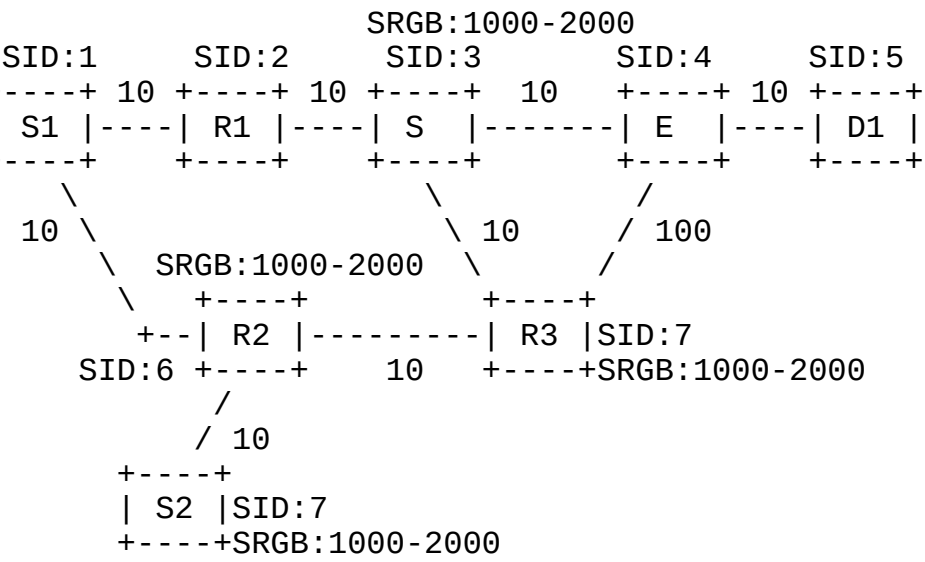

Figure 6: Sample Network with SRLG links

1. when the S->E link goes down, S and E generate the link down event, update their Router-LSA/ LSP and flood the updated information across the IGP domain.

2. The nodes in the IGP domain process the link-down event for affected destinations.If there are any other links with same SRLG on the path to destination, the nearest PLRs for those links are identified. For destination D1, R2 identifies two PLRs S1 and S for the S->E link down event.

3. The nodes build the tunnelled path having multiple labels for each of the identified links. for ex, R2 builds a stack containing node-sid of S1 and S. The tunnelled path at R2 looks as shown in Figure 7 below.

Node	Destination Prefix	Label Operation
R2	D1	Push 1005, 1003(top), Fwd to S1

Figure 7: Sample ingress RIB for SRLG failure handling

4. The procedures as described in sec 3.1 for the link-down event is followed to achieve micro-loop free convergence.

3.5. Handling ECMP

When a network event is received, if the the change causes only one of the ECMP paths to change, then the micro-loop prevention mechanisms described in sec 3.1 and 3.2 are applied to the changed path only. As described in section 3.1 and 3.2 , if there is an ECMP path to the nearest PLR, then all ECMP paths are used to tunnel the traffic during convergence.

3.6. Recognizing same network event

When a link goes down, both the ends of the link report the event by updating their LSP/LSA and flood it across the IGP domain. It is possible that the same network event being reported by two nodes is perceived as two different network events by the nodes in the IGP domain. The nodes processing the network events SHOULD evaluate if the received multiple events correspond to a single event by comparing the both ends of the reported link and also by looking at the previous event for which micro-loop prevention is being performed. If the event is same then micro-loop prevention procedures MUST be allowed to continue and MUST NOT be aborted.

Node down or new node addition events are reported by removing a link or adding a new link by all the adjacent nodes. In addition Node up event also comprises of a new LSA advertisement. The criteria to recognize if the event is same is to look at both ends of the changed link. If one end of the changed link maps to previously reported events and the other end of the link (advertising router) changes for each successive event, then the event is SHOULD be recognized as a new node addition or a node deletion. Micro-loop procedures MUST be allowed to continue and MUST NOT be aborted.

3.7. Partial deployment Considerations

The micro-loop mechanisms described in this document, are very effective and safe when all the nodes in the network support this feature and apply it when a network event happens. However, in some topologies, when all the nodes do not support the micro-loop prevention mechanism, the time duration of the loop can increase when only some nodes apply the procedures described in this document and some nodes do not.

For example, consider the sample topology described in the figure below.

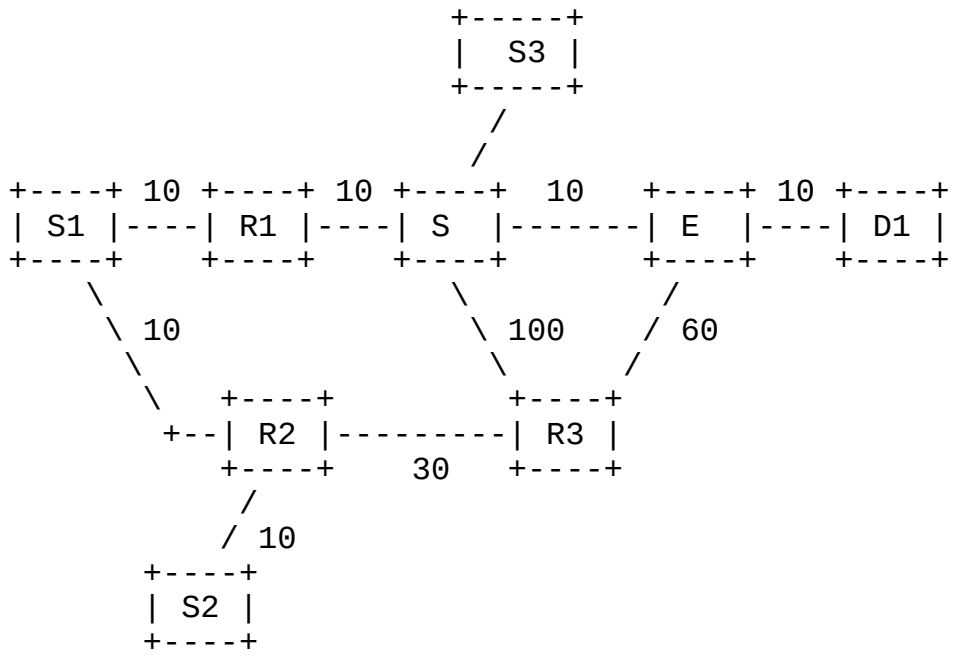

Figure 8: Sample Network with partial deployment

In this topology, S1, S2, and S3 are traffic sources and D1 is the destination. For each of the sources, Figure 9 shows the path before the failure (the before path) and the path after the failure (the post convergence path)..

Src	Dest	Original Path	Post-Convergence Path
S1	D1	S1->R1->S->E->D1	S1->R2->R3->E->D1
S2	D1	S2->R2->S1->R1->S->E->D1	S2->R2->R3->E->D1
S3	D1	S3->S->E->D1	S3->S->R1->S1->R2->R3->E->D1

Figure 9: Traffic flow in normal operation and post convergence path with S->E link down

In the above topology, if the PLR S does not support the micro-loop prevention mechanism but all other nodes support and apply this mechanism, then there is a possibility that the duration of traffic looping is higher than when the micro-loop prevention mechanisms are not applied at all. To mitigate this issue, protocol extensions to negotiate the support of this feature in the IGP domain is needed.

Section 4 describes the protocol mechanisms to advertise the support of this feature in OSPF and ISIS.

However, in certain deployments and topologies, it MAY be safe to apply the micro-loop prevention procedures even when all the nodes in the network do not support this feature, especially in topologies where the post convergence path from PLR does not traverse the nodes in P space of the PLR with respect to the the node or link being protected.

4. Protocol Procedures

4.1. OSPF

[RFC4970], defines Router Information (RI) LSA which may be used to advertise properties of the originating router. Payload of the RI LSA consists of one or more nested Type/Length/Value (TLV) triplets. This document defines a new TLV Micro-loop prevention support TLV which has following format:

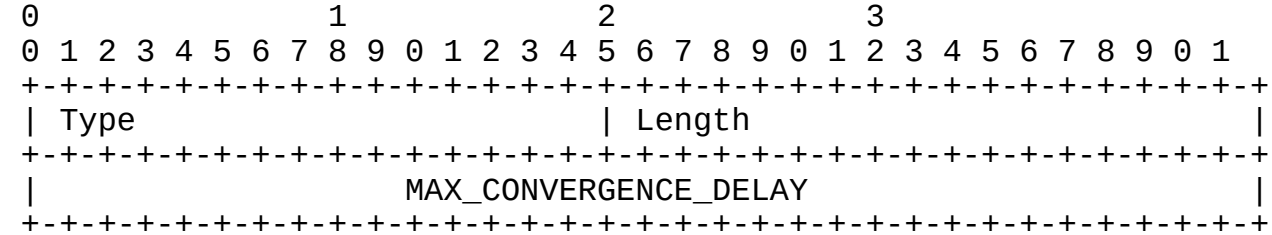

Figure 10: OSPF micro-loop prevention support TLV

Type : TBA, Suggested value 15

Length: 4

Value: Integer number indicating the maximum convergence delay in milliseconds. The delay SHOULD include convergence time for the IGP prefixes on the node.

4.2. ISIS

[RFC4971], defines Router capability TLV which may be used to advertise properties of the originating router. This document defines a new sub-TLV Micro-loop prevention support sub-TLV which has following format:

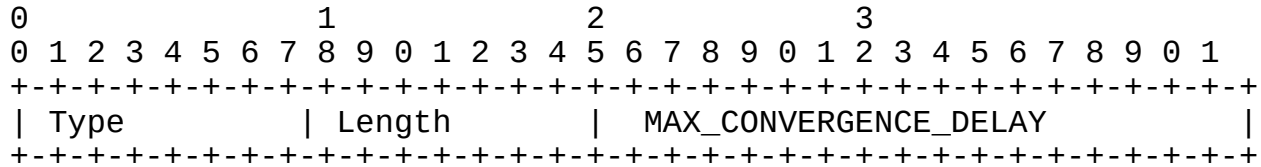

Figure 11: OSPF micro-loop prevention support TLV

The Router Capability TLV specifies flags that control its advertisement. The Micro-loop prevention support sub-TLV MUST be propagated throughout the level and SHOULD NOT be advertised across level boundaries. Therefore Router Capability TLV distribution flags SHOULD be set accordingly, i.e.: the S flag in the Router Capability TLV [RFC4971] MUST be unset.

Type : TBA, Suggested value 5

Length: 2

Value: Integer number indicating the maximum convergence delay in milliseconds. The delay SHOULD include convergence time for the IGP prefixes on edge node.

4.3. Elements of procedure

The micro-loop prevention support sub-TLV MUST be advertised only when the feature is enabled.When all the nodes in the IGP domain advertise this sub-TLV, a node supporting this feature MUST perform the micro-loop prevention procedures as described in this document. The micro-loop prevention mechanisms are applied within the OSPF area or ISIS level.

When there are one or more nodes in the IGP domain which do not support this feature, a node MAY perform micro-loop prevention procedures. Near side tunnelling mechanism ensures that when a group of nodes support this feature, traffic sourced from these set of nodes do not suffer micro-loop. A manageability interface SHOULD be provided to support micro-loop prevention in case of partial feature deployment.

5. Security Considerations

This document does not introduce any further security issues other than those discussed in [RFC2328] ,[RFC5340] , [ISO10589] and [RFC1195]

6. IANA Considerations

This specification updates one OSPF registry: OSPF Router Information (RI) TLVs Registry

i) TBD - Micro-loop prevention support sub-TLV

This specification updates one ISIS registry: ISIS Router capability TLVs (TLV 242) Registry

i) TBD - Micro-loop prevention support sub-TLV

7. Acknowledgments

Thanks to Chris Bowers, Hannes Gredler and Eric Rosen for valuable inputs.

8. References

8.1. Normative References

[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>.

[RFC4970] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, DOI 10.17487/RFC4970, July 2007, <http://www.rfc-editor.org/info/rfc4970>.

[RFC4971] Vasseur, JP., Ed., Shen, N., Ed., and R. Aggarwal, Ed., "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", RFC 4971, DOI 10.17487/RFC4971, July 2007, <http://www.rfc-editor.org/info/rfc4971>.

8.2. Informative References

[I-D.litkowski-rtgwg-uloop-delay] Litkowski, S., Decraene, B., Filsfils, C., and P. Francois, "Microloop prevention by introducing a local convergence delay", draft-litkowski-rtgwg-uloop-delay-04 (work in progress), October 2015.

[ISO10589] "Intermediate system to Intermediate system intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473), ISO/IEC 10589:2002, Second Edition.", Nov 2002.

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, DOI 10.17487/RFC1195, December 1990, <http://www.rfc-editor.org/info/rfc1195>.

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

[RFC5286] Atlas, A., Ed. and A. Zinin, Ed., "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, DOI 10.17487/RFC5286, September 2008, <http://www.rfc-editor.org/info/rfc5286>.

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

[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

Shraddha Hegde
Juniper Networks, Inc.
Exora Business Park
Bangalore, KA 560037
India

Email: shraddha@juniper.net

Pushpasis Sarkar
Juniper Networks, Inc.
Exora Business Park
Bangalore, KA 560037
India

Email: psarkar@juniper.net; pushpasis.ietf@gmail.com