Routing Area Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 18, 2014

S. Litkowski
B. Decraene
Orange
C. Filsfils
Cisco Systems
P. Francois
IMDEA Networks
February 14, 2014

Microloop prevention by introducing a local convergence delay
draft-litkowski-rtgwg-uloop-delay-03

Abstract

This document describes a mechanism for link-state routing protocols to prevent local transient forwarding loops in case of link failure. This mechanism Proposes a two-steps convergence by introducing a delay between the convergence of the node adjacent to the topology change and the network wide convergence.

As this mechanism delays the IGP convergence it may only be used for planned maintenance or when fast reroute protects the traffic between the link failure and the IGP convergence.

Simulations using real network topologies have been performed and show that local loops are a significant portion (>50%) of the total forwarding loops.

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].

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 18, 2014.

Table of Contents

1. Introduction

Micro-forwarding loops and some potential solutions are well described in [RFC5715]. This document describes a simple targeted mechanism that solves micro-loops local to the failure; based on network analysis, these are a significant portion of the micro-forwarding loops. A simple and easily deployable solution to these local micro-loops is critical because these local loops cause traffic loss after an advanced fast-reroute alternate has been used (see Section 2.1).

Consider the case in Figure 1 where S does not have an LFA to protect its traffic to D. That means that all non-D neighbors of S on the topology will send to S any traffic destined to D if a neighbor did not, then that neighbor would be loop-free. Regardless of the advanced fast-reroute technique used, when S converges to the new topology, it will send its traffic to a neighbor that was not loop-free and thus cause a local micro-loop. The deployment of advanced fast-reroute techniques motivates this simple router-local mechanism to solve this targeted problem. This solution can be work with the various techniques described in [RFC5715].

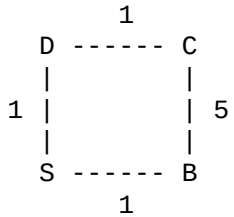


Figure 1

When S-D fails, a transient forwarding loop may appear between S and B if S updates its forwarding entry to D before B.

2. Transient forwarding loops side effects

Even if they are very limited in duration, transient forwarding loops may cause high damage for the network.

2.1. Fast reroute unefficiency

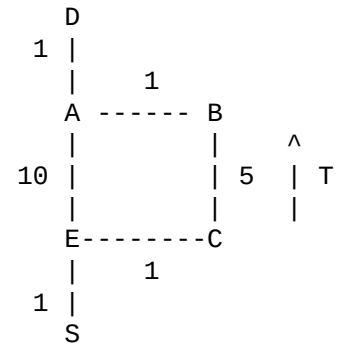


Figure 2 - RSVPTE FRR case

In figure 2, a RSVP-TE tunnel T, provisionned on C and terminating on B, is used to protect against C-B link failure (IGP shortcut activated on C). Primary path of T is C->B and FRR is activated on T providing a FRR bypass or detour using path C->E->A->B. On C, nexthop to D is tunnel T thanks to IGP shortcut. When C-B link fails :

1. C detects the failure, and updates the tunnel path using preprogrammed FRR path, traffic path from S to D is : S->E->C->E->A->B->A->D .

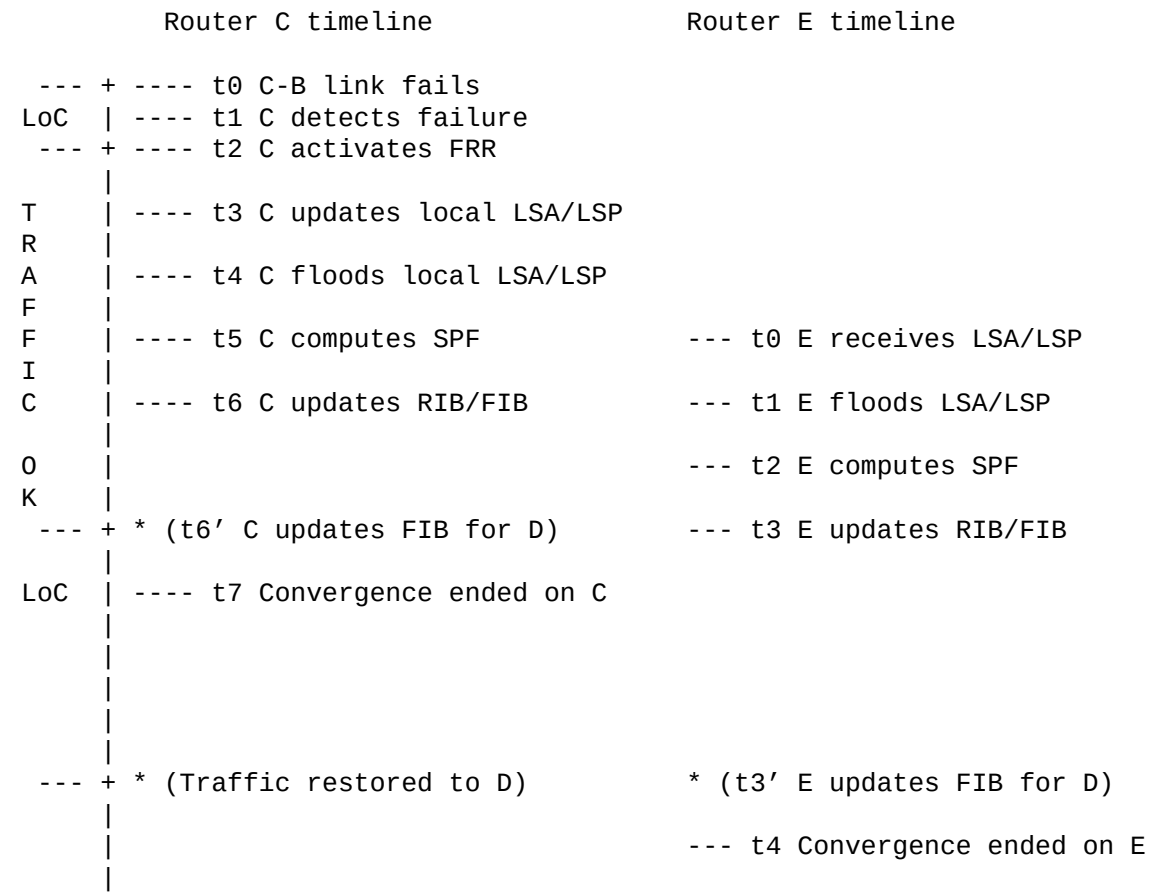
2. In parallel, on router C, both IGP convergence and TE tunnel convergence (tunnel path recomputation) are occuring :

   * T path is recomputed : C->E->A->B

   * IGP path to D is recomputed : C->E->A->D

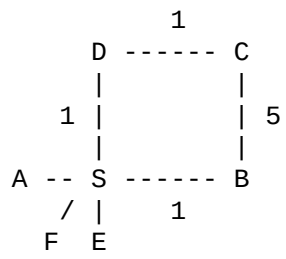
3. On C, tail-end of the TE tunnel (router B) is no more on SPT to D, so C does not encapsulate anymore the traffic to D using the tunnel T and update forwarding entry to D using nexthop E.

If C updates its forwarding entry to D before router E, there would be a transient forwarding loop between C and E until E has converged.



The issue described here is completely independent of the fast-reroute mechanism involved (TE FRR, LFA/rLFA, MRT ...). Fast-reroute is working perfectly but ensures protection, by definition, only until the PLR has converged. When implementing FRR, a service provider wants to guarantee a very limited loss of connectivity time. The previous example shows that the benefit of FRR may be completely lost due to a transient forwarding loop appearing when PLR has converged. Delaying FIB updates after IGP convergence may permit to keep fast-reroute path until neighbor has converged and preserve customer traffic.

2.2. Network congestion



In the figure above, as presented in Section 1, when link S-D fails, a transient forwarding loop may appear between S and B for destination D. The traffic on S-B link will constantly increase due to the looping traffic to D. Depending on TTL of packets, traffic rate destinated to D and bandwidth of link, the S-B link may be congestioned in few hundreds of milliseconds and will stay overloaded until the loop is solved.

Congestion introduced by transient forwarding loops are problematic as they are impacting traffic that is not directly concerned by the failing network component. In our example, the congestion of S-B link will impact customer traffic that is not directly concerned by the failure : e.g. A to B, F to B, E to B. Class of services may be implemented to mitigate the congestion but some traffic not directly concerned by the failure would still be dropped as a router is not able to identify looped traffic from normal traffic.

3. Overview of the solution

This document defines a two-step convergence initiated by the router detecting the failure and advertising the topological changes in the IGP. This introduces a delay between the convergence of the local router and the network wide convergence. This delay is positive in case of "down" events and negative in case of "up" events.

This ordered convergence, is similar to the ordered FIB proposed defined in [RFC6976], but limited to only one hop distance. As a consequence, it is simpler and becomes a local only feature not requiring interoperability; at the cost of only covering the transient forwarding loops involving this local router. The proposed mechanism also reuses some concept described in [I-D.ietf-rtgwg-microloop-analysis] with some limitation.

4. Specification

4.1. Definitions

This document will refer to the following existing IGP timers:

- LSP_GEN_TIMER: to batch multiple local events in one single local LSP update. It is often associated with damping mechanism to slowdown reactions by incrementing the timer when multiple consecutive events are detected.

- SPF_TIMER: to batch multiple events in one single computation. It is often associated with damping mechanism to slowdown reactions by incrementing the timer when the IGP is instable.

- IGP_LDP_SYNC_TIMER: defined in [RFC5443] to give LDP some time to establish the session and learn the MPLS labels before the link is used.

This document introduces the following two new timers :

- ULOOP_DELAY_DOWN_TIMER: slowdown the local node convergence in case of link down events.

- ULOOP_DELAY_UP_TIMER: slowdown the network wide IGP convergence in case of link up events.

4.2. Current IGP reactions

Upon a change of status on an adjacency/link, the existing behavior of the router advertising the event is the following:

1. UP/Down event is notified to IGP.

2. IGP processes the notification and postpones the reaction in LSP_GEN_TIMER msec.

3. Upon LSP_GEN_TIMER expiration, IGP updates its LSP/LSA and floods it.

4. SPF is scheduled in SPF_TIMER msec.

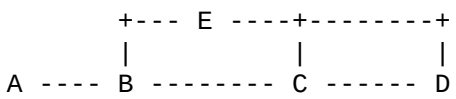
5. Upon SPF_TIMER expiration, SPF is computed and RIB/FIB are updated.

4.3. Local events

The mechanisms described in this document assume that there has been a single failure as seen by the IGP area/level. If this assumption is violated (e.g. multiple links or nodes failed), then standard IP

convergence MUST be applied. There are three types of single failures: local link, local node, and remote failure.

Example :



Let B be the computing router when the link B-C fails. B updates its local LSP/LSA describing the link B->C as down, C does the same, and both start flooding their updated LSP/LSAs. During the SPF_TIMER period, B and C learn all the LSPs/LSAs to consider. B sees that C is flooding as down a link where B is the other end and that B and C are describing the same single event. Since B receives no other changes, B can determine that this is a local link failure.

[Editor s Note: Detection of a failed broadcast link involves additional complexity and will be described in a future version.]

If a router determines that the event is local link failure, then the router may use the mechanism described in this document.

Distinguishing local node failure from remote or multiple link failure requires additional logic which is future work to fully describe. To give a sense of the work necessary, if node C is failing, routers B,E and D are updating and flooding updated LSPs/ LSAs. B would need to determine the changes in the LSPs/LSAs from E and D and see that they all relate to node C which is also the far-end of the locally failed link. Once this detection is accurately done, the same mechanism of delaying local convergence can be applied.

4.4. Local delay

4.4.1. Link down event

Upon an adjacency/link down event, this document introduces a change in step 5 in order to delay the local convergence compared to the network wide convergence: the node SHOULD delay the forwarding entry updates by ULOOP_DELAY_DOWN_TIMER. Such delay SHOULD only be introduced if all the LSDB modifications processed are only reporting down local events . Note that determining that all topological change are only local down events requires analyzing all modified LSP/LSA as a local link or node failure will typically be notified by multiple nodes. If a subsequent LSP/LSA is received/updated and a new SPF computation is triggered before the expiration of ULOOP_DELAY_DOWN_TIMER, then the same evaluation SHOULD be performed.

As a result of this addition, routers local to the failure will converge slower than remote routers. Hence it SHOULD only be done for non urgent convergence, such as for administrative de-activation (maintenance) or when the traffic is Fast ReRouted.

4.4.2. Link up event

Upon an adjacency/link up event, this document introduces the following change in step 3 where the node SHOULD:

- Firstly build a LSP/LSA with the new adjacency but setting the metric to MAX_METRIC . It SHOULD flood it but not compute the SPF at this time. This step is required to ensure the two way connectivity check on all nodes when computing SPF.

- Then build the LSP/LSA with the target metric but SHOULD delay the flooding of this LSP/LSA by SPF_TIMER + ULOOP_DELAY_UP_TIMER. MAX_METRIC is equal to MaxLinkMetric (0xFFFF) for OSPF and 2^24-2 (0xFFFFFE) for IS-IS.

- Then continue with next steps (SPF computation) without waiting for the expiration of the above timer. In other word, only the flooding of the LSA/LSP is delayed, not the local SPF computation.

As as result of this addition, routers local to the failure will converge faster than remote routers.

If this mechanism is used in cooperation with "LDP IGP Synchronization" as defined in [RFC5443] then the mechanism defined in RFC 5443 is applied first, followed by the mechanism defined in this document. More precisely, the procedure defined in this document is applied once the LDP session is considered "fully operational" as per [RFC5443].

5. Applicability

As previously stated, the mechanism only avoids the forwarding loops on the links between the node local to the failure and its neighbor. Forwarding loops may still occur on other links.

5.1. Applicable case : local loops

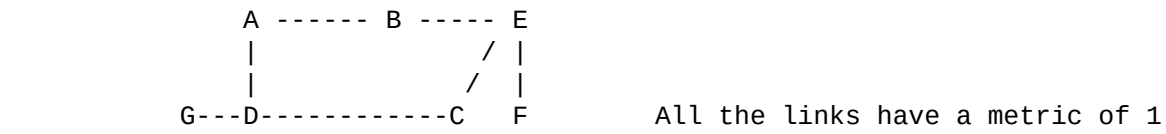


Figure 2

Let us consider the traffic from G to F. The primary path is G->D->C->E->F. When link CE fails, if C updates its forwarding entry for F before D, a transient loop occurs. This is sub-optimal as C has FRR enabled and it breaks the FRR forwarding while all upstream routers are still forwarding the traffic to itself.

By implementing the mechanism defined in this document on C, when the CE link fails, C delays the update of his forwarding entry to F, in order to let some time for D to converge. FRR keeps protecting the traffic during this period. When the timer expires on C, forwarding entry to F is updated. There is no transient forwarding loop on the link CD.

5.2. Non applicable case : remote loops

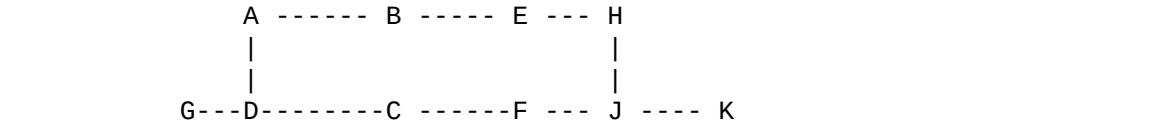


Figure 3

Let us consider the traffic from G to K. The primary path is G->D->C->F->J->K. When the CF link fails, if C updates its forwarding entry to K before D, a transient loop occurs between C and D.

By implementing the mechanism defined in this document on C, when the link CF fails, C delays the update of his forwarding entry to K, letting time for D to converge. When the timer expires on C, forwarding entry to F is updated. There is no transient forwarding loop between C and D. However, a transient forwarding loop may still occur between D and A. In this scenario, this mechanism is not enough to address all the possible forwarding loops. However, it does not create additional traffic loss. Besides, in some cases -such as when the nodes update their FIB in the following order C, A, D, for example because the router A is quicker than D to converge- the mechanism may still avoid the forwarding loop that was occuring.

6. Simulations

Simulations have been run on multiple service provider topologies. So far, only link down event have been tested.

Topology	Gain
T1	71%
T2	81%
T3	62%
T4	50%
T5	70%
T6	70%
T7	59%
T8	77%

Table 1: Number of Repair/Dst that may loop

We evaluated the efficiency of the mechanism on eight different service provider topologies (different network size, design). The benefit is displayed in the table above. The benefit is evaluated as follows:

- We consider a tuple (link A-B, destination D, PLR S, backup nexthop N) as a loop if upon link A-B failure, the flow from a router S upstream from A (A could be considered as PLR also) to D may loop due to convergence time difference between S and one of his neighbor N.

- We evaluate the number of potential loop tuples in normal conditions.

- We evaluate the number of potential loop tuples using the same topological input but taking into account that S converges after N.

- Gain is how much loops (remote and local) we succeed to suppress.

On topology 1, 71% of the transient forwarding loops created by the failure of any link are prevented by implementing the local delay. The analysis shows that all local loops are obviously solved and only remote loops are remaining.

7. Deployment considerations

Transient forwarding loops have the following drawbacks :

- Limit FRR efficiency : even if FRR is activated in 50msec, as soon as PLR has converged, traffic may be affected by a transient loop.

- It may impact traffic not directly concerned by the failure (due to link congestion).

This local delay proposal is a transient forwarding loop avoidance mechanism (like OFIB). Even if it only address local transient loops, , the efficiency versus complexity comparison of the mechanism makes it a good solution. It is also incrementally deployable with incremental benefits, which makes it an attractive option for both vendors to implement and Service Providers to deploy. Delaying convergence time is not an issue if we consider that the traffic is protected during the convergence.

8. Comparison with other solutions

As stated in Section 3, our solution reuses some concepts already introduced by other IETF proposals but tries to find a tradeoff between efficiency and simplicity. This section tries to compare behaviors of the solutions.

8.1. PLSN

PLSN ([I-D.ietf-rtgwg-microloop-analysis]) describes a mechanism where each node in the network tries a avoid transient forwarding loops upon a topology change by always keeping traffic on a loop-free path for a defined duration (locked path to a safe neighbor). The locked path may be the new primary nexthop, another neighbor, or the old primary nexthop depending how the safety condition is satisified.

PLSN does not solve all transient forwarding loops (see [I-D.ietf-rtgwg-microloop-analysis] Section 4 for more details).

Our solution reuse some concept of PLSN but in a more simple fashion :

- PLSN has 3 different behavior : keep using old nexthop, use new primary nexthop if safe, or use another safe nexthop, while our solution only have one : keep using the current nexthop (old primary, or already activated FRR path).

- PLSN may cause some damage while using a safe nexthop which is not the new primary nexthop in case the new safe nexthop does not enough provide enough bandwidth (see [I-D.ietf-rtgwg-lfa-manageability]). Our solution may not experience this issue as the service provider may have control on the FRR path being used preventing network congestion.

- PLSN applies to all nodes in a network (remote or local changes), while our mechanism applies only on the nodes connected to the topology change.

8.2. OFIB

OFIB ([RFC6976]) describes a mechanism where convergence of the network upon a topology change is made ordered to prevent transient forwarding loops. Each router in the network must deduce the failure type from the LSA/LSP received and compute/apply a specific FIB update timer based on the failure type and its rank in the network considering the failure point as root.

This mechanism permit to solve all the transient forwarding loop in a network at the price of introducing complexity in the convergence process that may require strong monitoring by the service provider.

Our solution reuses the OFIB concept but limits it to the first hop that experience the topology change. As demonstrated, our proposal permits to solve all the local transient forwarding loops that represents a high percentage of all the loops. Moreover limiting the mechanism to one hop permit to keep the network-wide convergence behavior.

9. Security Considerations

This document does not introduce change in term of IGP security. The operation is internal to the router. The local delay does not increase the attack vector as an attacker could only trigger this mechanism if he already has be ability to disable or enable an IGP link. The local delay does not increase the negative consequences as if an attacker has the ability to disable or enable an IGP link, it can already harm the network by creating instability and harm the traffic by creating forwarding packet loss and forwarding loss for the traffic crossing that link.

10. Acknowledgements

We wish to thanks the authors of [RFC6976] for introducing the

concept of ordered convergence: Mike Shand, Stewart Bryant, Stefano Previdi, and Olivier Bonaventure.

11. IANA Considerations

This document has no actions for IANA.

12. References

12.1. Normative References

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

[RFC5443] Jork, M., Atlas, A., and L. Fang, "LDP IGP Synchronization", RFC 5443, March 2009.

[RFC5715] Shand, M. and S. Bryant, "A Framework for Loop-Free Convergence", RFC 5715, January 2010.

12.2. Informative References

[I-D.ietf-rtgwg-lfa-manageability] Litkowski, S., Decraene, B., Filsfils, C., Raza, K., Horneffer, M., and p. psarkar@juniper.net, "Operational management of Loop Free Alternates", draft-ietf-rtgwg-lfa-manageability-03 (work in progress), February 2014.

[I-D.ietf-rtgwg-microloop-analysis] Zinin, A., "Analysis and Minimization of Microloops in Link-state Routing Protocols", draft-ietf-rtgwg-microloop-analysis-01 (work in progress), October 2005.

[I-D.ietf-rtgwg-remote-lfa] Bryant, S., Filsfils, C., Previdi, S., Shand, M., and S. Ning, "Remote LFA FRR", draft-ietf-rtgwg-remote-lfa-04 (work in progress), November 2013.

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

[RFC6571] Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free

Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.

[RFC6976] Shand, M., Bryant, S., Previdi, S., Filsfils, C., Francois, P., and O. Bonaventure, "Framework for Loop-Free Convergence Using the Ordered Forwarding Information Base (oFIB) Approach", RFC 6976, July 2013.

Authors' Addresses

Stephane Litkowski
Orange

Email: stephane.litkowski@orange.com

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

Clarence Filsfils
Cisco Systems

Email: cfilsfil@cisco.com

Pierre Francois
IMDEA Networks

Email: pierre.francois@imdea.org