Routing Area Working Group
Internet-Draft
Intended status: Informational
Expires: July 20, 2019

S. Litkowski
Orange Business Service
B. Decraene
Orange
M. Horneffer
Deutsche Telekom

January 16, 2019

Link State protocols SPF trigger and delay algorithm impact on IGP micro-loops

draft-ietf-rtgwg-spf-uloop-pb-statement-10

Abstract

A micro-loop is a packet forwarding loop that may occur transiently among two or more routers in a hop-by-hop packet forwarding paradigm.

In this document, we are trying to analyze the impact of using different Link State IGP (Interior Gateway Protocol) implementations in a single network, with respect to micro-loops. The analysis is focused on the SPF (Shortest Path First) delay algorithm.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 20, 2019.

Table of Contents

1. Introduction 2
2. Problem statement 4
3. SPF trigger strategies 5
4. SPF delay strategies 6
4.1. Two steps SPF delay 6
4.2. Exponential backoff 7
5. Mixing strategies 8
6. Benefits of standardized SPF delay behavior 12
7. Security Considerations 13
8. Acknowledgements 13
9. IANA Considerations 13
10. References 14
10.1. Normative References 14
10.2. Informative References 14
Authors' Addresses 15

1. Introduction

Link State IGP protocols are based on a topology database on which the SPF algorithm is run to find a consistent set of non-looping routing paths.

Specifications like IS-IS ([RFC1195]) propose some optimizations of the route computation (See Appendix C.1 of [RFC1195]) but not all the implementations follow those non-mandatory optimizations.

We will call "SPF triggers", the events that would lead to a new SPF computation based on the topology.

Link State IGP protocols, like OSPF ([RFC2328]) and IS-IS ([RFC1195]), are using multiple timers to control the router behavior

in case of churn: SPF delay, PRC (Partial Route Computation) delay, LSP (Link State Packet) generation delay, LSP flooding delay, LSP retransmission interval...

Some of those timers (values and behavior) are standardized in protocol specifications, while some are not. The SPF computation related timers have generally remained unspecified.

For non standardized timers, implementations are free to implement them in any way. For some standardized timers, we can also see that rather than using static configurable values for such timer, implementations may offer dynamically adjusted timers to help control the churn.

We will call "SPF delay", the timer that exists in most implementations that specifies the required delay before running SPF computation after a SPF trigger is received.

A micro-loop is a packet forwarding loop that may occur transiently among two or more routers in a hop-by-hop packet forwarding paradigm. We can observe that these micro-loops are formed when two routers do not update their Forwarding Information Base (FIB) for a certain prefix at the same time. The micro-loop phenomenon is described in [I-D.ietf-rtgwg-microloop-analysis].

Two micro-loop mitigation techniques have been defined by IETF. [RFC6976] has not been widely implemented, presumably due to the complexity of the technique. [RFC8333] has been implemented. However, it does not prevent all micro-loops that can occur for a given topology and failure scenario.

In multi-vendor networks, using different implementations of a link state protocol may favor micro-loops creation during the convergence process due to discrepancies of timers. Service Providers are already aware to use similar timers (values and behavior) for all the network as a best practice, but sometimes it is not possible due to limitations of implementations.

This document will present reasons for service providers to have consistent implementations of Link State protocols across vendors. We are particularly analyzing the impact of using different Link State IGP implementations in a single network in regards of micro-loops. The analysis is focused on the SPF delay algorithm.

[RFC8405] defines a solution that partially addresses this problem statement and this document captures the reasoning of the provided solution.

2. Problem statement

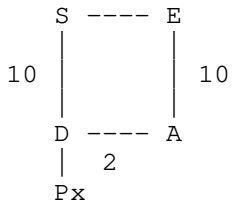


Figure 1 - Network topology suffering from micro-loops

Figure 1 represents a small network composed of four routers (S,D,E and A).Router S uses primarily the SD link to reach the prefixes behind router D (named Px). When the SD link fails, the IGP convergence occurs. If S converges before E, S will forward the traffic to Px through E, but as E has not converged yet, E will loop back traffic to S, leading to a micro-loop.

The micro-loop appears due to the asynchronous convergence of nodes in a network when an event occurs.

Multiple factors (or a combination of these factors) may increase the probability for a micro-loop to appear:

- the delay of failure notification: the greater the time gap between E and S being advised of the failure, the more a micro-loop may have a chance to appear.

- the SPF delay: most implementations support a delay for the SPF computation to try to catch as many events as possible. If S uses an SPF delay timer of x msec and E uses an SPF delay timer of y msec and x < y, E would start converging after S leading to a potential micro-loop.

- the SPF computation time: mostly a matter of CPU power and optimizations like incremental SPF. If S computes its SPF faster than E, there is a chance for a micro-loop to appear. CPUs are today fast enough to consider SPF computation time as negligible (on the order of milliseconds in a large network).

- the SPF computation ordering: an SPF trigger can be common to multiple IGP areas or levels (e.g., IS-IS Level1/Level2) or for multiple address families with multi-topologies. There is no specified order for SPF computation today and it is implementation dependent. In such scenarios, if the order of SPF computation

done in S and E for each area/level/topology/SPF-algorithm is different, there is a possibility for a micro-loop to appear.

- the RIB and FIB prefix insertion speed or ordering. This is highly dependent on the implementation.

Even if all of these factors may increase the probability for a micro-loop to appear, the SPF delay, especially in case of churn, plays a significant role. As the number of IGP events increase, the delta between SPF delay values used by routers becomes significant and the dominating factor (especially when one router increases its timer exponentially while another one increases it in a more smoother way). Another important factor is the time to update the FIB. As of today, total FIB update time is the major factor for IGP convergence. However, for micro-loops, what matters is not the total time, but the difference to install the same prefix between nodes. The time to update the FIB may be the main part for the first iteration but is not for subsequent IGP events. In addition, the time to update the FIB is very implementation specific and difficult/impossible to standardize, while the SPF delay algorithm may be standardized.

As a consequence, this document will focus on the analysis of the SPF delay behavior and associated triggers.

3. SPF trigger strategies

Depending on the change advertised in LSPDU (Link State Protocol Data Unit) or LSA (Link State Advertisement), the topology may be affected or not. An implementation may avoid running the SPF computation (and may only run an IP reachability computation instead) if the advertised change does not affect the topology.

Different strategies exists to trigger the SPF computation:

1. An implementation may always run a full SPF for any type of change.

2. An implementation may run a full SPF only when required. For example, if a link fails, a local node will run an SPF for its local LSP update. If the LSP from the neighbor (describing the same failure) is received after SPF has started, the local node can decide that a new full SPF is not required as the topology has not changed.

3. If the topology does not change, an implementation may only recompute the IP reachability.

As noted in Section 1, SPF optimizations are not mandatory in specifications. This has led to the implementation of different strategies.

4. SPF delay strategies

Implementations of link state routing protocols use different strategies to delay the SPF computation. The two most common SPF delay behaviors are the following:

1. Two phase SPF delay.

2. Exponential backoff delay.

These behaviors will be explained in the next sections.

4.1. Two steps SPF delay

The SPF delay is managed by four parameters:

- Rapid delay: amount of time to wait before running SPF, after the initial SPF trigger event.

- Rapid runs: the number of consecutive SPF runs that can use the rapid delay. When the number is exceeded, the delay moves to the slow delay value.

- Slow delay: amount of time to wait before running SPF.

- Wait time: amount of time to wait without receiving SPF trigger events before going back to the rapid delay.

Example: Rapid delay (RD) = 50msec, Rapid runs = 3, Slow delay (SD) = 1sec, Wait time = 2sec

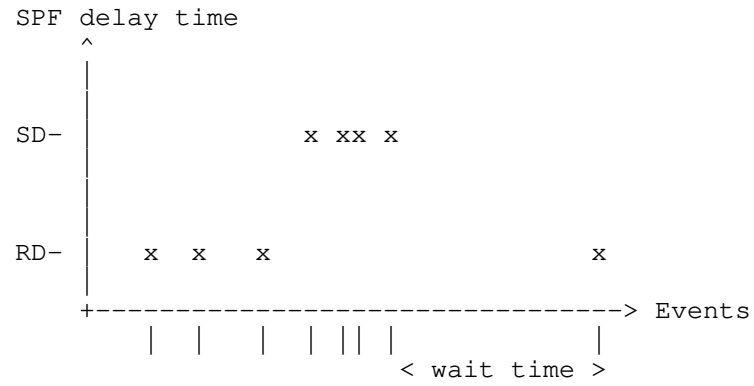


Figure 2 - Two phase delay algorithm

4.2. Exponential backoff

The algorithm has two modes: the fast mode and the backoff mode. In the fast mode, the SPF delay is usually delayed by a very small amount of time (fast reaction). When an SPF computation has run in the fast mode, the algorithm automatically moves to the backoff mode (a single SPF run is authorized in the fast mode). In the backoff mode, the SPF delay is increasing exponentially at each run. When the network becomes stable, the algorithm moves back to the fast mode. The SPF delay is managed by four parameters:

- First delay: amount of time to wait before running SPF. This delay is used only when SPF is in fast mode.
- Incremental delay: amount of time to wait before running SPF. This delay is used only when SPF is in backoff mode and increments exponentially at each SPF run.
- Maximum delay: maximum amount of time to wait before running SPF.
- Wait time: amount of time to wait without events before going back to the fast mode.

Example: First delay (FD) = 50msec, Incremental delay (ID) = 50msec, Maximum delay (MD) = 1sec, Wait time = 2sec

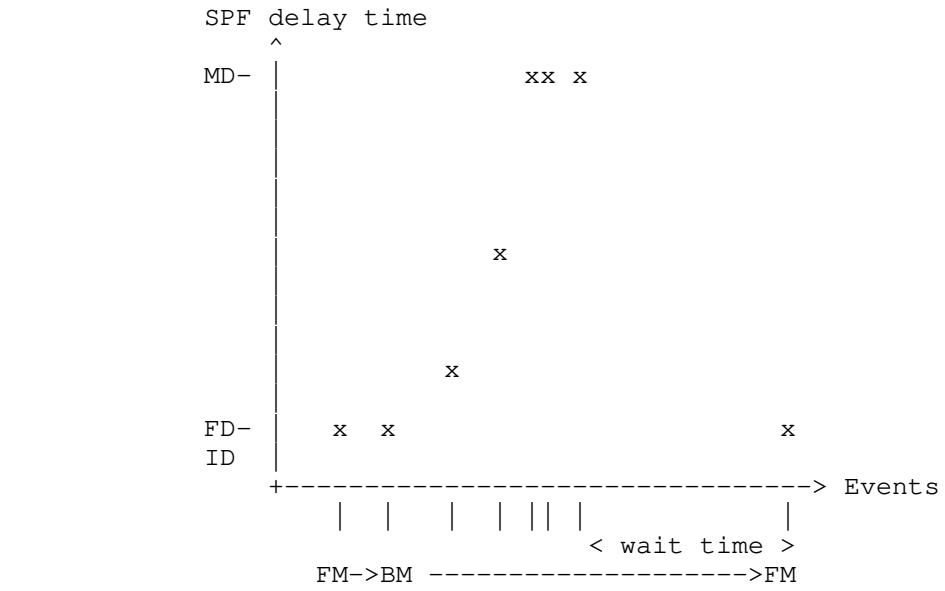


Figure 3 - Exponential delay algorithm

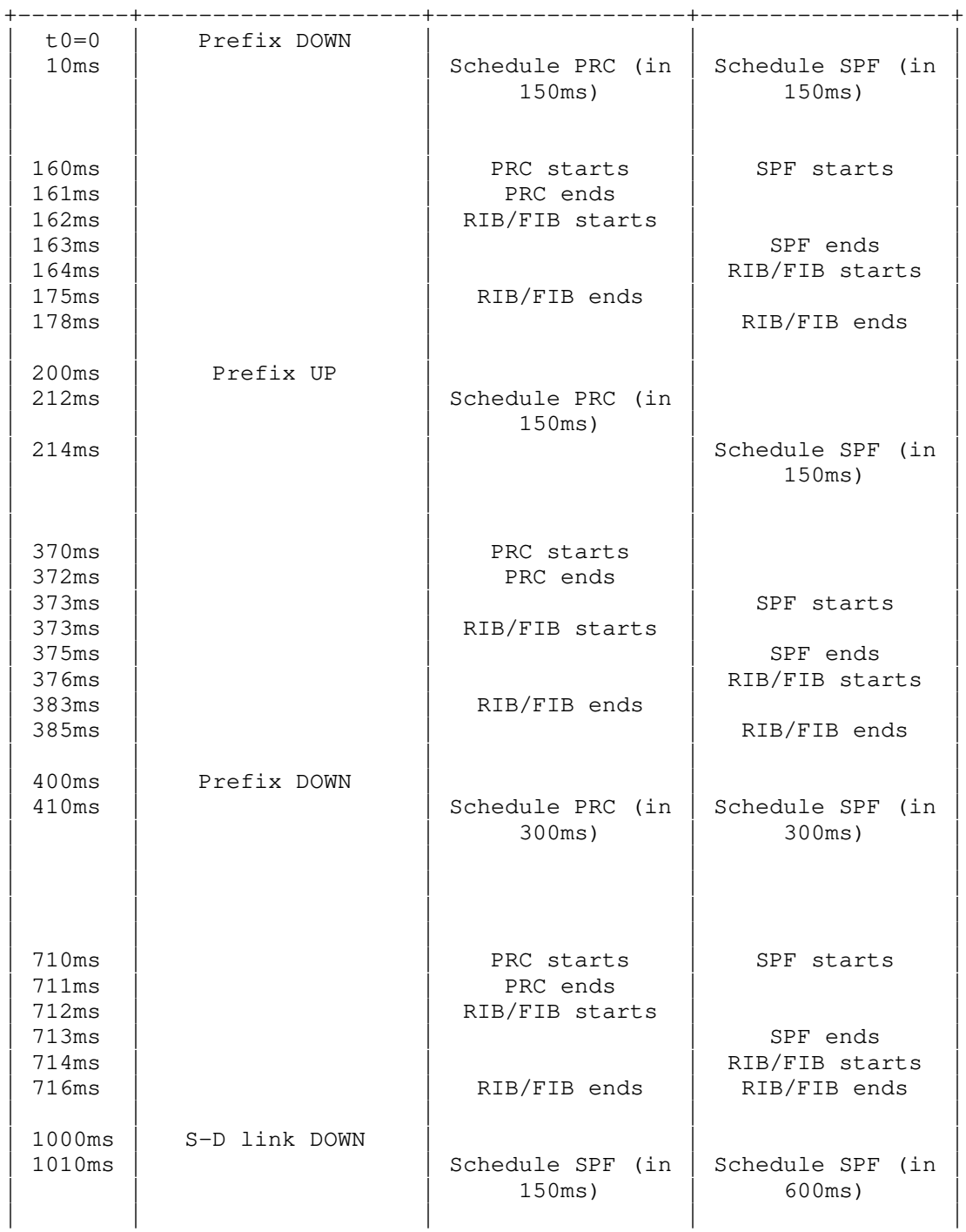
5. Mixing strategies

In Figure 1, we consider a flow of packet from S to D. We consider that S is using optimized SPF triggering (Full SPF is triggered only when necessary), and two steps SPF delay (rapid=150ms,rapid-runs=3, slow=1s). As implementation of S is optimized, Partial Reachability Computation (PRC) is available. We consider the same timers as SPF for delaying PRC. We consider that E is using a SPF trigger strategy that always compute a Full SPF for any change, and uses the exponential backoff strategy for SPF delay (start=150ms, inc=150ms, max=1s)

We also consider the following sequence of events:

- t0=0 ms: a prefix is declared down in the network. We consider this event to happen at time=0.
- 200ms: the prefix is declared as up.
- 400ms: a prefix is declared down in the network.
- 1000ms: S-D link fails.

Time	Network Event	Router S events	Router E events


t0=0
Prefix DOWN
10ms
Schedule PRC (in 150ms)
Schedule SPF (in 150ms)
160ms
PRC starts
SPF starts
161ms
PRC ends
162ms
RIB/FIB starts
163ms
SPF ends
164ms
RIB/FIB starts
175ms
RIB/FIB ends
178ms
RIB/FIB ends
200ms
Prefix UP
212ms
Schedule PRC (in 150ms)
214ms
Schedule SPF (in 150ms)
370ms
PRC starts
372ms
PRC ends
373ms
SPF starts
373ms
RIB/FIB starts
375ms
SPF ends
376ms
RIB/FIB starts
383ms
RIB/FIB ends
385ms
RIB/FIB ends
400ms
Prefix DOWN
410ms
Schedule PRC (in 300ms)
Schedule SPF (in 300ms)
710ms
PRC starts
SPF starts
711ms
PRC ends
712ms
RIB/FIB starts
713ms
SPF ends
714ms
RIB/FIB starts
716ms
RIB/FIB ends
RIB/FIB ends
1000ms
S-D link DOWN
1010ms
Schedule SPF (in 150ms)
Schedule SPF (in 600ms)

1160ms			SPF starts	
1161ms			SPF ends	
1162ms	Micro-loop may start from here		RIB/FIB starts	
1175ms			RIB/FIB ends	
1612ms				SPF starts
1615ms				SPF ends
1616ms				RIB/FIB starts
1626ms	Micro-loop ends			RIB/FIB ends

Table 1 - Route computation when S and E use the different behaviors and multiple events appear

In the Table 1, we can see that due to discrepancies in the SPF management, after multiple events of a different type, the values of the SPF delay are completely misaligned between node S and node E, leading to the creation of micro-loops.

The same issue can also appear with only a single type of event as shown below:

Time	Network Event	Router S events	Router E events
t0=0	Link DOWN		
10ms		Schedule SPF (in 150ms)	Schedule SPF (in 150ms)
160ms		SPF starts	SPF starts
161ms		SPF ends	
162ms		RIB/FIB starts	
163ms			SPF ends
164ms			RIB/FIB starts
175ms		RIB/FIB ends	
178ms			RIB/FIB ends
200ms	Link DOWN		
212ms		Schedule SPF (in 150ms)	
214ms			Schedule SPF (in 150ms)

370ms		SPF starts	
372ms		SPF ends	
373ms			SPF starts
373ms		RIB/FIB starts	
375ms			SPF ends
376ms			RIB/FIB starts
383ms		RIB/FIB ends	
385ms			RIB/FIB ends
400ms	Link DOWN		
410ms		Schedule SPF (in 150ms)	Schedule SPF (in 300ms)
560ms		SPF starts	
561ms		SPF ends	
562ms	Micro-loop may start from here	RIB/FIB starts	
568ms		RIB/FIB ends	
710ms			SPF starts
713ms			SPF ends
714ms			RIB/FIB starts
716ms	Micro-loop ends		RIB/FIB ends
1000ms	Link DOWN		
1010ms		Schedule SPF (in 1s)	Schedule SPF (in 600ms)
1612ms			SPF starts
1615ms			SPF ends
1616ms	Micro-loop may start from here		RIB/FIB starts
1626ms			RIB/FIB ends
2012ms		SPF starts	
2014ms		SPF ends	
2015ms		RIB/FIB starts	
2025ms	Micro-loop ends	RIB/FIB ends	

Table 2 - Route computation upon multiple link down events when S and E use the different behaviors

6. Benefits of standardized SPF delay behavior

Using the same event sequence as in Table 1, we may expect fewer and/or shorter micro-loops using a standardized SPF delay.

Time	Network Event	Router S events	Router E events
t0=0	Prefix DOWN		
10ms		Schedule PRC (in 150ms)	Schedule PRC (in 150ms)
160ms		PRC starts	PRC starts
161ms		PRC ends	
162ms		RIB/FIB starts	PRC ends
163ms			RIB/FIB starts
175ms		RIB/FIB ends	
176ms			RIB/FIB ends
200ms	Prefix UP		
212ms		Schedule PRC (in 150ms)	
213ms			Schedule PRC (in 150ms)
370ms		PRC starts	PRC starts
372ms		PRC ends	
373ms		RIB/FIB starts	PRC ends
374ms			RIB/FIB starts
383ms		RIB/FIB ends	
384ms			RIB/FIB ends
400ms	Prefix DOWN		
410ms		Schedule PRC (in 300ms)	Schedule PRC (in 300ms)

710ms		PRC starts	PRC starts
711ms		PRC ends	PRC ends
712ms		RIB/FIB starts	
713ms			RIB/FIB starts
716ms		RIB/FIB ends	RIB/FIB ends
1000ms	S-D link DOWN		
1010ms		Schedule SPF (in 150ms)	Schedule SPF (in 150ms)
1160ms		SPF starts	
1161ms		SPF ends	SPF starts
1162ms	Micro-loop may start from here	RIB/FIB starts	SPF ends
1163ms			RIB/FIB starts
1175ms		RIB/FIB ends	
1177ms	Micro-loop ends		RIB/FIB ends

Table 3 - Route computation when S and E use the same standardized behavior

As displayed above, there could be some other parameters like router computation power, flooding timers that may also influence micro-loops. In all the examples in this document comparing the SPF timer behavior of router S and router E, we have made router E a bit slower than router S. This can lead to micro-loops even when both S and E use a common standardized SPF behavior. However, we expect that by aligning implementations of the SPF delay, service providers may reduce the number and the duration of micro-loops.

7. Security Considerations

This document does not introduce any security consideration.

8. Acknowledgements

Authors would like to thank Mike Shand and Chris Bowers for their useful comments.

9. IANA Considerations

This document has no action for IANA.

10. References

10.1. Normative References

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

[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>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8405] Decraene, B., Litkowski, S., Gredler, H., Lindem, A., Francois, P., and C. Bowers, "Shortest Path First (SPF) Back-Off Delay Algorithm for Link-State IGPs", RFC 8405, DOI 10.17487/RFC8405, June 2018, <https://www.rfc-editor.org/info/rfc8405>.

10.2. Informative References

[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.

[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, DOI 10.17487/RFC6976, July 2013, <https://www.rfc-editor.org/info/rfc6976>.

[RFC8333] Litkowski, S., Decraene, B., Filsfils, C., and P. Francois, "Micro-loop Prevention by Introducing a Local Convergence Delay", RFC 8333, DOI 10.17487/RFC8333, March 2018, <https://www.rfc-editor.org/info/rfc8333>.

Authors' Addresses

Stephane Litkowski
Orange Business Service

Email: stephane.litkowski@orange.com

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

Martin Horneffer
Deutsche Telekom

Email: martin.horneffer@telekom.de