Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 10, 2015

B. Decraene
Orange

March 9, 2015

Back-off SPF algorithm for link state IGP
draft-decraene-rtgwg-backoff-algo-01

Abstract

This document defines a standard algorithm to back-off link-state IGP SPF computations.

Having one standardized algorithm improves interoperability by reducing the probability and/or duration of transient forwarding loops during the IGP convergence in the area/level when the network reacts to multiple consecutive events.

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

Copyright Notice

1. Introduction

Link state IGP, such as IS-IS [ISO10589-Second-Edition] and OSPF [RFC2328], performs distributed computation on all nodes of the area/level. In order to have consistent routing tables across the network, such distributed computation requires that all routers have the same vision of the network (Link State DataBase (LSDB)) and perform their computation at the same time.

In general, when the network is stable, there is a desire to compute the new SPF as soon as the failure is known, in order to quickly route around the failure. However, when the network is experiencing multiple consecutive failures over a short period of time, there is a desire to limit the frequency of SPF computations. Indeed, this allow reducing the control plane resources used by IGP and all protocols/sub system reacting on it such as LDP, RSVP-TE, BGP, Fast ReRoute computations, FIB updates..., reduce the churn on nodes and in the network, in particular reduce side effects such as micro-loops which may happen during each IGP convergence.

To allow for this, some back-off algorithm have been implemented. Different implementations choose different algorithms, hence in a multi-vendor network, it's not possible to enforce that all routers triggers their SPF computation after the same waiting delay. This situation increases the average differential delay between routers end of RIB computation. It also increases the probability that different routers compute their RIB based on a different LSDB. Both increases the probability and/or duration of micro-loops.

To allow for multi-vendors networks having all the routers delaying their SPF for the same duration, this document specifies a standardized algorithm. Implementations may offer alternative optional algorithms.

2. High level goals

The high level goals of this algorithm are the following:

- Very fast convergence for single simple events (link failure).
- Fast convergence in general while the IGP stability is considered under control.
- A long delay when the IGP stability is considered out of control, in order to let all related process calm down.
- At any time, try to avoid using different SPF_TIMERS values for nodes in the area/level. Even though not all nodes will receive IGP message at the same time (due to difference in distance from the source and due to different flooding implementations on the path from the source).

3. Definitions and parameters

IGP events: An LSDB change requiring a new RIB computation (topology change, prefix change, metric change). No distinction is done between the type of computation performed (e.g. full SPF, incremental SPF, PRC). The type of computation is a local consideration.

The SPF_DELAY timer can take the following values:

INITIAL_WAIT: a very small delay to quickly handle link failure. e.g. 0 millisecond.

FAST_WAIT: a small delay to have a fast convergence. e.g. 50-100 millisecond. Note: we want to be fast, but as this failure requires multiple IGP events, being too fast increase the probability to receive additional IGP events just after the RIB computation.

LONG_WAIT: a long delay as IGP is unstable. e.g. 2 seconds. Note: let’s bring calm in the IGP.

The TIME_TO_CONVERGE timer is the time to learn all the IGP events related to a single failure (e.g. node failure, SRLG failure). e.g. 1 second. It’s mostly dependent on variation of failure detection times between all nodes which are neighbour to the failure, and then may depend on different flooding algorithms of nodes in the network.

The HOLD_DOWN timer is the time needed with no IGP events received, before considering that the IGP is quiet again and we can set the SPF_DELAY back to INITAL_WAIT. e.g. 5 seconds.

4. Principle of SPF delay algorithm

The first IGP event is handled very quickly (INITIAL_WAIT) in order to be very reactive for the first event if it only needs one IGP event (e.g. link failure, prefix change).

If more IGP events are received quickly after, we consider that they are related to the same single failure, and handle the IGP events relatively quickly (FAST_WAIT) during the time needed to receive all the IGP events related to the failure (TIME_TO_CONVERGE).

If IGP events are still received after this time, then the network is presumably experiencing multiple independent failures and the while waiting for its stability, the computations are delayed for a longer time (LONG_WAIT).

Note: previous SPF delay algorithms used to count the number of RIB computations. However, as all nodes may receive the LSP events in a different way we cannot assume that all nodes will perform the same number of SPF computations or that they will schedule them at the same time. For example, assuming that the SPF delay is 50 ms, node R1 may receive 3 IGP events (E1, E2, E3) in those 50 ms and hence will perform a single routing computation. While another node R2 may only receive 2 events (E1, E2) in those 50ms and hence will schedule another routing computation when further receiving E3. That’s why this document prefers to define a time limit (TIME_TO_CONVERGE) since the first event, rather than a number of routing computations.

5. Specification of SPF delay algorithm

When the previous IGP events is more than HOLD_DOWN ago:

- The IGP is set to the QUIET state.

When the IGP is in the QUIET state and an IGP event is received:

- The time of this first IGP event is stored in FIRST_EVENT_TIME.
- The next RIB computation time is set to LSP receive time + INITIAL_WAIT.
- The IGP is set to the FAST_WAIT state.

When the IGP is in the FAST_WAIT state and an IGP event is received:

- If more than TIME_TO_CONVERGE has passed since FIRST_EVENT_TIME, then the IGP is set to the HOLD_DOWN state.

- If the next RIB_computation time is in the past, set the next RIB computation time to LSP receive time + FAST_WAIT.

When the IGP is in the HOLD_DOWN state and an IGP event is received:

- If the next RIB_computation time is in the past, set the next RIB computation time to LSP receive time + LONG_WAIT.

6. Impact on micro-loops

Micro-loops during IGP convergence are due to a non synchronized or non ordered update of the forwarding information tables (FIB) [RFC5715] [RFC6976] [I-D.litkowski-rtgwg-spf-uloop-pb-statement]. FIB are installed after multiple steps such as SPF wait time, SPF computation, FIB distribution and FIB update. This document only address the first contribution. This standardized procedure reduces the probability and/or duration of micro-loops when the IGP experience multiple consecutive events. It does not remove all micro-loops. However, it is beneficial and its cost seems limited compared to full solutions such as [RFC5715] or [RFC6976].

7. IANA Considerations

No IANA actions required.

8. Security considerations

This document has no impact on the security of the IGP.

9. Acknowledgements

We would like to acknowledge Hannes Gredler, Les Ginsberg and Pierre Francois for the discussions related to this document.

10. References

10.1. Normative References

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

10.2. Informative References

[I-D.litkowski-rtgwg-spf-uloop-pb-statement] Litkowski, S., "Link State protocols SPF trigger and delay algorithm impact on IGP microloops", draft-litkowski-rtgwg-spf-uloop-pb-statement-02 (work in progress), March 2015.

[ISO10589-Second-Edition]
International Organization for Standardization, "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.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

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

[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.

Author’s Address

Bruno Decraene
Orange
38 rue du General Leclerc
Issy Moulineaux cedex 9 92794
France

Email: bruno.decraene@orange.com