Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 25, 2018

B. Decraene
Orange
S. Litkowski
Orange Business Service
H. Gredler
RtBrick Inc
A. Lindem
Cisco Systems
P. Francois

C. Bowers
Juniper Networks, Inc.
October 22, 2017

SPF Back-off algorithm for link state IGPs
draft-ietf-rtgwg-backoff-algo-06

Abstract

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

Having one standard algorithm improves interoperability by reducing the probability and/or duration of transient forwarding loops during the IGP convergence when the IGP reacts to multiple temporally close IGP 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

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 April 25, 2018.

Table of Contents

1. Introduction

Link state IGPs, such as IS-IS [ISO10589-Second-Edition] and OSPF [RFC2328], perform distributed route computation on all routers in the area/level. In order to have consistent routing tables across the network, such distributed computation requires that all routers have the same version of the network topology (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 a new SPF as soon as a failure is detected in order to quickly route around the failure. However, when the network is experiencing multiple temporally close failures over a short period of time, there is a conflicting desire to limit the frequency of SPF computations. Indeed, this allows a reduction in control plane resources used by IGPs and all protocols/subsystems reacting on the attendant route change, such as LDP, RSVP-TE, BGP, Fast ReRoute computations, FIB updates... This also reduces the churn on routers and in the network and, in particular, reduces the side effects such as micro-loops that ensue during IGP convergence.

To allow for this, IGPs implement an SPF back-off algorithm. However, different implementations have choosen different algorithms. Hence, in a multi-vendor network, it's not possible to ensure that all routers trigger their SPF computation after the same delay. This situation increases the average differential delay between routers completing their SPF computation. It also increases the probability that different routers compute their FIBs based on different LSDB versions. Both factors increase the probability and/or duration of micro-loops.

To allow multi-vendor networks to have all routers delay their SPF computations for the same duration, this document specifies a standard algorithm. Optionally, implementations may offer alternative algorithms.

2. High level goals

The high level goals of this algorithm are the following:

- Very fast convergence for a single event (e.g., link failure).
- Paced fast convergence for multiple temporally close IGP events while IGP stability is considered acceptable.
- Delayed convergence when IGP stability is problematic. This will allow the IGP and related processes to conserve resources during the period of instability.
- Always try to avoid different SPF_DELAY timers values across different routers in the area/level. Even though not all routers will receive IGP messages at the same time, due to differences both in the distance from the originator of the IGP event and in flooding implementations.

3. Definitions and parameters

IGP events: The reception or origination of an IGP LSDB change requiring a new routing table computation. Examples are a topology change, a prefix change, a metric change on a link or prefix... Note that locally triggering a routing table computation is not considered as an IGP event since other IGP routers are unaware of this occurrence.

Routing table computation: Computation of the routing table, by the IGP, using the IGP LSDB. No distinction is made between the type of computation performed. e.g., full SPF, incremental SPF, Partial Route Computation (PRC). The type of computation is a local consideration. This document may interchangeably use the terms routing table computation and SPF computation.

SPF_DELAY: The delay between the first IGP event triggering a new routing table computation and the start of that routing table computation. It can take the following values:

INITIAL_SPF_DELAY: A very small delay to quickly handle link failure, e.g., 0 milliseconds.

SHORT_SPF_DELAY: A small delay to have a fast convergence in case of a single component failure (node, SRLG..), e.g., 50-100 milliseconds.

LONG_SPF_DELAY: A long delay when the IGP is unstable, e.g., 2 seconds. Note that this allows the IGP network to stabilize.

TIME_TO_LEARN_INTERVAL: This is the maximum duration typically needed to learn all the IGP events related to a single component failure (e.g., router failure, SRLG failure), e.g., 1 second. It's mostly dependent on failure detection time variation between all routers that are adjacent to the failure. Additionally, it may depend on the different IGP implementations across the network, related to origination and flooding of their link state advertisements.

HOLDDOWN_INTERVAL: The time required with no received IGP events before considering the IGP to be stable again and allowing the SPF_DELAY to be restored to INITIAL_SPF_DELAY. e.g., 3 seconds.

SPF_TIMER: The Finite State Machine (FSM) abstract timer that uses the computed SPF delay. Upon expiration, the Route Table Computation (as defined above) is performed.

4. Principles of SPF delay algorithm

For this first IGP event, we assume that there has been a single simple change in the network which can be taken into account using a single routing computation (e.g., link failure, prefix (metric) change) and we optimize for very fast convergence, delaying the routing computation by INITIAL_SPF_DELAY. Under this assumption, there is no benefit in delaying the routing computation. In a typical network, this is the most common type of IGP event. Hence, it makes sense to optimize this case.

If subsequent IGP events are received in a short period of time (TIME_TO_LEARN_INTERVAL), we then assume that a single component failed, but that this failure requires the knowledge of multiple IGP events in order for IGP routing to converge. Under this assumption, we want fast convergence since this is a normal network situation. However, there is a benefit in waiting for all IGP events related to this single component failure so that the IGP can compute the post-failure routing table in a single route computation. In this situation, we delay the routing computation by SHORT_SPF_DELAY.

If IGP events are still received after TIME_TO_LEARN_INTERVAL from the initial IGP event received in QUIET state, then the network is presumably experiencing multiple independent failures. In this case, while waiting for network stability, the computations are delayed for a longer time represented by LONG_SPF_DELAY. This SPF delay is kept until no IGP events are received for HOLDDOWN_INTERVAL.

Note that previous SPF delay algorithms used to count the number of SPF computations. However, as all routers may receive the IGP events at different times, we cannot assume that all routers 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, router R1 may receive 3 IGP events (E1, E2, E3) in those 50 ms and hence will perform a single routing computation. While another router R2 may only receive 2 events (E1, E2) in those 50 ms and hence will schedule another routing computation when receiving E3. That's why this document uses a time (TIME_TO_LEARN) from the initial event detection/reception as opposed to counting the number of SPF computations to determine when the IGP is unstable.

5. Specification of the SPF delay state machine

5.1. States

This section describes the state machine. The naming and semantics of each state corresponds directly to the SPF delay used for IGP events received in that state. Three states are defined:

QUIET: This is the initial state, when no IGP events have occured for at least HOLDDOWN_INTERVAL since the previous routing table computation. The state is meant to handle link failures very quickly.

SHORT_WAIT: State entered when an IGP event has been received in QUIET state. This state is meant to handle single component failure requiring multiple IGP events (e.g., node, SRLG).

LONG_WAIT: State reached after TIME_TO_LEARN_INTERVAL. In other words, state reached after TIME_TO_LEARN_INTERVAL in state SHORT_WAIT. This state is meant to handle multiple independent component failures during periods of IGP instability.

5.2. States Transitions

The FSM is initialized to the QUIET_STATE with all three timers deactivated. The following diagram describes briefly the state transitions.

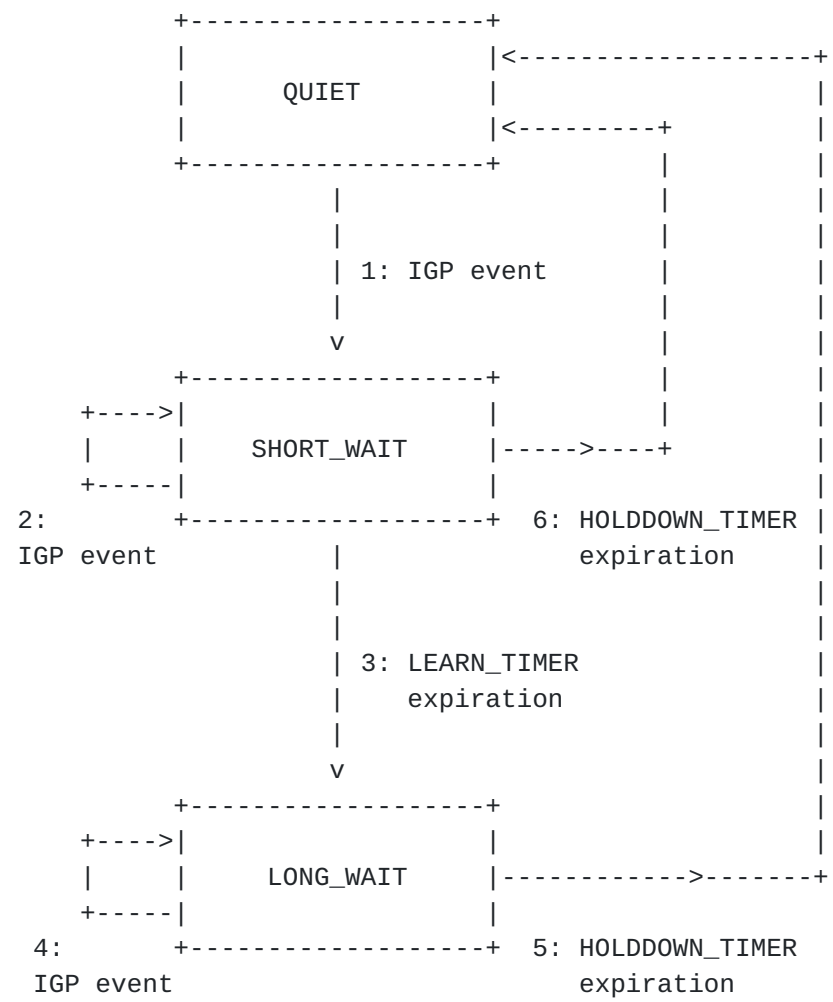

Figure 1: State Machine

5.3. FSM Events

This section describes the events and the actions performed in response.

Event 1: IGP event, while in QUIET_STATE.

Actions on event 1:

- If SPF_TIMER is not already running, start it with value INITIAL_SPF_DELAY.
- Start LEARN_TIMER with TIME_TO_LEARN_INTERVAL.
- Start HOLDDOWN_TIMER with HOLDDOWN_INTERVAL.

- Transition to SHORT_WAIT state.

Event 2: IGP event, while in SHORT_WAIT.

Actions on event 2:

- Reset HOLDDOWN_TIMER to HOLDDOWN_INTERVAL.
- If SPF_TIMER is not already running, start it with value SHORT_SPF_DELAY.
- Remain in current state.

Event 3: LEARN_TIMER expiration.

Actions on event 3:

- Transition to LONG_WAIT state.

Event 4: IGP event, while in LONG_WAIT.

Actions on event 4:

- Reset HOLDDOWN_TIMER to HOLDDOWN_INTERVAL.
- If SPF_TIMER is not already running, start it with value LONG_SPF_DELAY.
- Remain in current state.

Event 5: HOLDDOWN_TIMER expiration, while in LONG_WAIT.

Actions on event 5:

- Transition to QUIET state.

Event 6: HOLDDOWN_TIMER expiration, while in SHORT_WAIT.

Actions on event 6:

- Deactivate LEARN_TIMER.
- Transition to QUIET state.

6. Parameters

All the parameters MUST be configurable [I-D.ietf-isis-yang-isis-cfg] [I-D.ietf-ospf-yang] at the protocol instance granularity. They MAY be configurable at the area/level granularity. All the delays (INITIAL_SPF_DELAY, SHORT_SPF_DELAY, LONG_SPF_DELAY, TIME_TO_LEARN_INTERVAL, HOLDDOWN_INTERVAL) SHOULD be configurable at the millisecond granularity. They MUST be configurable at least at the tenth of second granularity. The configurable range for all the parameters SHOULD at least be from 0 milliseconds to 60 seconds.

This document does not propose default values for the parameters because these values are expected to be context dependent. Implementations are free to propose their own default values.

In order to satisfy the goals stated in Section 2, operators are RECOMMENDED to configure delay intervals such that SPF_INITIAL_DELAY <= SPF_SHORT_DELAY and SPF_SHORT_DELAY <= SPF_LONG_DELAY.

When setting (default) values, one SHOULD consider the customers and their application requirements, the computational power of the routers, the size of the network, and, in particular, the number of IP prefixes advertised in the IGP, the frequency and number of IGP events, the number of protocols reactions/computations triggered by IGP SPF (e.g., BGP, PCEP, Traffic Engineering CSPF, Fast ReRoute computations).

Note that some or all of these factors may change over the life of the network. In case of doubt, it's RECOMMENDED to play it safe and start with safe, i.e., longer timers.

For the standard algorithm to be effective in mitigating micro-loops, it is RECOMMENDED that all routers in the IGP domain, or at least all the routers in the same area/level, have exactly the same configured values.

7. Partial Deployment

In general, the SPF delay algorithm is only effective in mitigating micro-loops if it is deployed on all routers in the IGP domain or, at least, all routers in an IGP area/level. The impact of partial deployment is based on the particular event, topology, and the SPF algorithm(s) used on other routers in the IGP area/level. In cases where the previous SPF algorithm was implemented uniformly, partial deployment will increase the frequency and duration of micro-loops. Hence, it is RECOMMENDED that all routers in the IGP domain or at least within the same area/level be migrated to the SPF algorithm described herein at roughly the same time.

Note that this is not a new consideration as over times, network operators have changed SPF delay parameters in order to accommodate new customer requirements for fast convergence, as permitted by new software and hardware. They may also have progressively replaced an implementation with a given SPF delay algorithm by another implementation with a different one.

8. 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.ietf-rtgwg-spf-uloop-pb-statement]. FIBs are installed after multiple steps such as SPF wait time, SPF computation, FIB distribution, and FIB update. This document only addresses the first contribution. This standardized procedure reduces the probability and/or duration of micro-loops when IGPs experience multiple temporally close events. It does not prevent all micro-loops. However, it is beneficial and is less complex and costly to implement when compared to full solutions such as [RFC5715] or [RFC6976].

9. IANA Considerations

No IANA actions required.

10. Security considerations

The algorithm presented in this document does not compromise IGP security. An attacker having the ability to generate IGP events would be able to delay the IGP convergence time. The LONG_SPF_DELAY state may help mitigate the effects of Denial-of-Service (DOS) attacks generating many IGP events.

11. Acknowledgements

We would like to acknowledge Les Ginsberg, Uma Chunduri, Mike Shand and Alexander Vainshtein for the discussions and comments related to this document.

12. References

12.1. 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, <https://www.rfc-editor.org/info/rfc2119>.

12.2. Informative References

[I-D.ietf-isis-yang-isis-cfg]
Litkowski, S., Yeung, D., Lindem, A., Zhang, Z., and L. Lhotka, "YANG Data Model for IS-IS protocol", draft-ietf-isis-yang-isis-cfg-18 (work in progress), July 2017.

[I-D.ietf-ospf-yang]
Yeung, D., Qu, Y., Zhang, Z., Chen, I., and A. Lindem, "Yang Data Model for OSPF Protocol", draft-ietf-ospf-yang-08 (work in progress), July 2017.

[I-D.ietf-rtgwg-spf-uloop-pb-statement]
Litkowski, S., Decraene, B., and M. Horneffer, "Link State protocols SPF trigger and delay algorithm impact on IGP micro-loops", draft-ietf-rtgwg-spf-uloop-pb-statement-04 (work in progress), May 2017.

[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, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

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

[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>.

Authors' Addresses

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

Stephane Litkowski
Orange Business Service

Email: stephane.litkowski@orange.com

Hannes Gredler
RtBrick Inc

Email: hannes@rtbrick.com

Acee Lindem
Cisco Systems
301 Midenhall Way
Cary, NC 27513
USA

Email: acee@cisco.com

Pierre Francois

Email: pfrpfr@gmail.com

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

Email: cbowers@juniper.net