     Shortest Path First (SPF) Back-Off Delay Algorithm for Link-State IGPs

Abstract

   This document defines a standard algorithm to temporarily postpone or
   "back off" link-state IGP Shortest Path First (SPF) computations.
   This reduces the computational load and churn on IGP nodes when
   multiple temporally close network events trigger multiple 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.

Status of This Memo

   This is an Internet Standards Track document.

   This document is a product of the Internet Engineering Task Force
   (IETF). It represents the consensus of the IETF community. It has
   received public review and has been approved for publication by the
   Internet Engineering Steering Group (IESG). Further information on
   Internet Standards is available in Section 2 of RFC 7841.

   Information about the current status of this document, any errata,
   and how to provide feedback on it may be obtained at
   https://www.rfc-editor.org/info/rfc8405.

Copyright Notice

Table of Contents

1. Introduction

   Link-state IGPs, such as IS-IS [ISO10589], OSPF [RFC2328], and OSPFv3
   [RFC5340], 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 essentially at the same time.

   In general, when the network is stable, there is a desire to trigger
   a new Shortest Path First (SPF) computation as soon as a failure is
   detected in order to quickly route around the failure. However, when
   the network is experiencing multiple failures over a short period of
   time, there is a conflicting desire to limit the frequency of SPF
   computations, which would allow a reduction in control plane
   resources used by IGPs and all protocols/subsystems reacting to the
   attendant route change, such as LDP [RFC5036], RSVP-TE [RFC3209], BGP
   [RFC4271], Fast Reroute computations (e.g., Loop-Free Alternates
   (LFAs) [RFC5286]), FIB updates, etc. This also reduces network churn
   and, in particular, reduces side effects (such as micro-loops
   [RFC5715]) that ensue during IGP convergence.

   To allow for this, IGPs usually implement an SPF Back-Off Delay
   algorithm that postpones or backs off the SPF computation. However,
   different implementations chose 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 and maximum 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 as discussed in Section 8.

   This document specifies a standard algorithm to allow multi-vendor
   networks to have all routers delay their SPF computations for the
   same duration.

1.1. 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.

2. High-Level Goals

   The high-level goals of this algorithm are the following:

   o very fast convergence for a single event (e.g., link failure),

   o paced fast convergence for multiple temporally close IGP events
      while IGP stability is considered acceptable,

   o delayed convergence when IGP stability is problematic (this will
      allow the IGP and related processes to conserve resources during
      the period of instability), and

   o avoidance of having different SPF_DELAY timer values (Section 3)
      across different routers in the area/level. This requires
      specific consideration as different routers may receive IGP
      messages at different intervals, or even in different orders, due
      to differences both in the distance from the originator of the IGP
      event and in flooding implementations.

3. Definitions and Parameters

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

   Routing table computation, in this document, is scoped to the IGP;
   so, this is the computation of the IGP RIB, performed by the IGP,
   using the IGP LSDB. No distinction is made between the type of
   computation performed, e.g., full SPF, incremental SPF, or 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 a single
    isolated link failure, e.g., 0 milliseconds.

    SHORT_SPF_DELAY: A small delay to provide fast convergence in the
    case of a single component failure (such as a node failure or Shared
    Risk Link Group (SRLG) failure) that leads to multiple IGP events,
    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
(such as router failure or 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/parameters across the
network and their relation to the origination and flooding of link
state advertisements.

HOLDDOWN_INTERVAL: The time required with no received IGP event
before considering the IGP to be stable again and allowing the
SPF_DELAY to be restored to INITIAL_SPF_DELAY, e.g., a
HOLDDOWN_INTERVAL of 3 seconds. The HOLDDOWN_INTERVAL MUST be
defaulted or configured to be longer than the TIME_TO_LEARN_INTERVAL.

4. Principles of the SPF Delay Algorithm

For the 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 by delaying the
initial routing computation for a small interval, 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: the IGP can then compute the post-
failure routing table in a single additional 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 (see Section 5.1), 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, which is represented by
LONG_SPF_DELAY. This SPF delay is used until no IGP events are
received for HOLDDOWN_INTERVAL.

   Note that in order to increase the consistency network wide, the
   algorithm uses a delay (TIME_TO_LEARN_INTERVAL) from the initial IGP
   event rather than the number of SPF computations performed. Indeed,
   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. For example, assuming that the SPF delay is 50
   milliseconds, router R1 may receive three IGP events (E1, E2, E3) in
   those 50 milliseconds and hence will perform a single routing
   computation, while another router R2 may only receive two events (E1,
   E2) in those 50 milliseconds and hence will schedule another routing
   computation when receiving E3.

5. Specification of the SPF Delay State Machine

   This section specifies the Finite State Machine (FSM) intended to
   control the timing of the execution of SPF calculations in response
   to IGP events.

5.1. State Machine

   The FSM is initialized to the QUIET state with all three timers
   (SPF_TIMER, HOLDDOWN_TIMER, and LEARN_TIMER) deactivated.

   The events that may change the FSM states are an IGP event or the
   expiration of one timer (SPF_TIMER, HOLDDOWN_TIMER, or LEARN_TIMER).

   The following diagram briefly describes the state transitions.

```
                     +-------------------+
             +---->|                   |<-------------------+
             |     |       QUIET       |                    |
             +-----|                   |<---------+         |
          7:        +-------------------+          |         |
          SPF_TIMER          |                     |         |
          expiration         |                     |         |
                             | 1: IGP event        |         |
                             |                     |         |
                             v                     |         |
                     +-------------------+          |         |
             +---->|                   |          |         |
             |     |    SHORT_WAIT     |----->----+         |
             +-----|                   |                    |
          2:        +-------------------+  6: HOLDDOWN_TIMER |
          IGP event          |               expiration      |
          8: SPF_TIMER       |                               |
             expiration      |                               |
                             | 3: LEARN_TIMER                |
                             |    expiration                 |
                             |                               |
                             v                               |
                     +-------------------+                   |
             +---->|                   |                   |
             |     |     LONG_WAIT     |----------->-------+
             +-----|                   |
          4:        +-------------------+  5: HOLDDOWN_TIMER
          IGP event                         expiration
          9: SPF_TIMER expiration
```
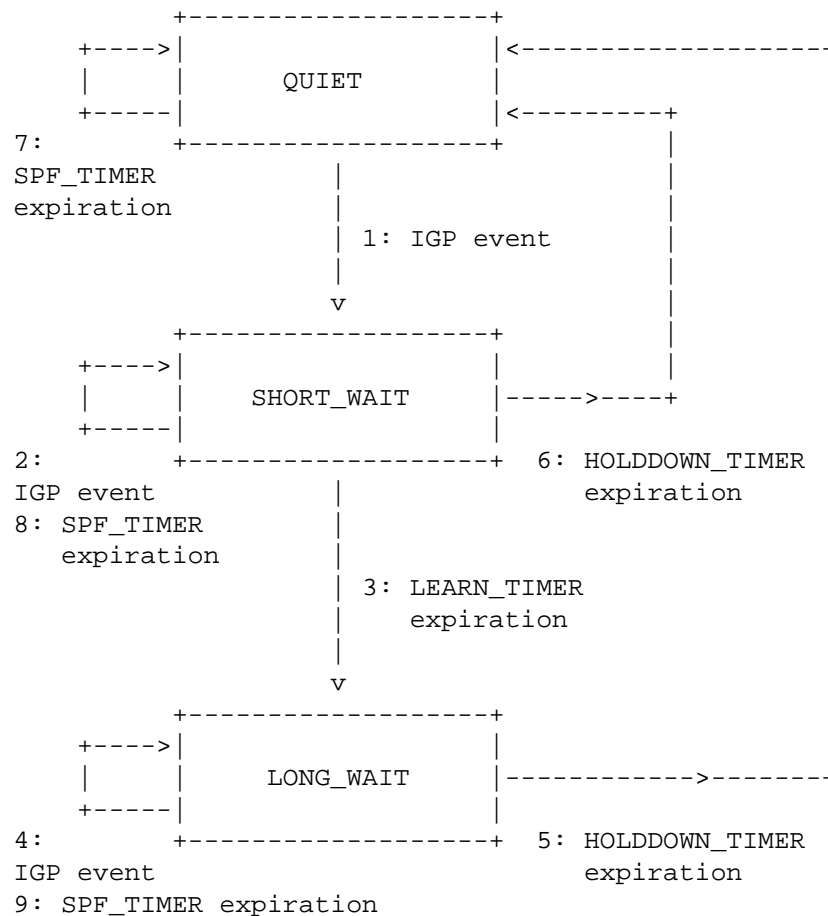
                        Figure 1: State Machine

5.2. States

   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 occurred
   for at least HOLDDOWN_INTERVAL since the last routing table
   computation. The state is meant to handle link failures very
   quickly.

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

LONG_WAIT: This is the 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.

5.3. Timers

SPF_TIMER: This is the FSM timer that uses the computed SPF delay.
Upon expiration, the routing table computation (as defined in
Section 3) is performed.

HOLDDOWN_TIMER: This is the FSM timer that is (re)started when an IGP
event is received and set to HOLDDOWN_INTERVAL. Upon expiration, the
FSM is moved to the QUIET state.

LEARN_TIMER: This is the FSM timer that is started when an IGP event
is received while the FSM is in the QUIET state. Upon expiration,
the FSM is moved to the LONG_WAIT state.

5.4. FSM Events

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

Transition 1: IGP event while in QUIET state

Actions on event 1:

o If SPF_TIMER is not already running, start it with value
   INITIAL_SPF_DELAY.

o Start LEARN_TIMER with TIME_TO_LEARN_INTERVAL.

o Start HOLDDOWN_TIMER with HOLDDOWN_INTERVAL.

o Transition to SHORT_WAIT state.

Transition 2: IGP event while in SHORT_WAIT

Actions on event 2:

o Reset HOLDDOWN_TIMER to HOLDDOWN_INTERVAL.

o If SPF_TIMER is not already running, start it with value
   SHORT_SPF_DELAY.

o Remain in current state.

   Transition 3: LEARN_TIMER expiration

   Actions on event 3:

   o Transition to LONG_WAIT state.

   Transition 4: IGP event while in LONG_WAIT

   Actions on event 4:

   o Reset HOLDDOWN_TIMER to HOLDDOWN_INTERVAL.

   o If SPF_TIMER is not already running, start it with value
      LONG_SPF_DELAY.

   o Remain in current state.

   Transition 5: HOLDDOWN_TIMER expiration while in LONG_WAIT

   Actions on event 5:

   o Transition to QUIET state.

   Transition 6: HOLDDOWN_TIMER expiration while in SHORT_WAIT

   Actions on event 6:

   o Deactivate LEARN_TIMER.

   o Transition to QUIET state.

   Transition 7: SPF_TIMER expiration while in QUIET

   Actions on event 7:

   o Compute SPF.

   o Remain in current state.

   Transition 8: SPF_TIMER expiration while in SHORT_WAIT

   Actions on event 8:

   o Compute SPF.

   o Remain in current state.

   Transition 9: SPF_TIMER expiration while in LONG_WAIT

   Actions on event 9:

   o Compute SPF.

   o Remain in current state.

6. Parameters

   All the parameters MUST be configurable at the protocol instance
   level. They MAY be configurable on a per IGP LSDB basis (e.g., IS-IS
   level, OSPF area, or IS-IS Level 1 area). All the delays
   (INITIAL_SPF_DELAY, SHORT_SPF_DELAY, LONG_SPF_DELAY,
   TIME_TO_LEARN_INTERVAL, and HOLDDOWN_INTERVAL) SHOULD be configurable
   with a granularity of a millisecond. They MUST be configurable with
   a granularity of at least a tenth of a second. The configurable
   range for all the parameters SHOULD be from 0 milliseconds to at
   least 6000 milliseconds. The HOLDDOWN_INTERVAL MUST be defaulted or
   configured to be longer than the TIME_TO_LEARN_INTERVAL.

   If this SPF Back-Off algorithm is enabled by default, then in order
   to have consistent SPF delays between implementations with default
   configuration, the following default values SHOULD be implemented:

       INITIAL_SPF_DELAY   50 ms
       SHORT_SPF_DELAY   200 ms
       LONG_SPF_DELAY   5000 ms
       TIME_TO_LEARN_INTERVAL   500 ms
       HOLDDOWN_INTERVAL   10000 ms

   In order to satisfy the goals stated in Section 2, operators are
   RECOMMENDED to configure delay intervals such that INITIAL_SPF_DELAY
   <= SHORT_SPF_DELAY and SHORT_SPF_DELAY <= LONG_SPF_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 as determined primarily by the
   number of IP prefixes advertised in the IGP, the frequency and number

of IGP events, and the number of protocol reactions/computations
triggered by IGP SPF computation (e.g., BGP, Path Computation Element
Communication Protocol (PCEP), Traffic Engineering Constrained SPF
(CSPF), and 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 that timer intervals should be chosen
conservatively (i.e., longer timer values).

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 Back-Off Delay algorithm is only effective in
mitigating micro-loops if it is deployed with the same parameters on
all routers in the IGP domain or, at least, all routers in an IGP
area/level. The impact of partial deployment is dependent on the
particular event, the topology, and the algorithm(s) used on other
routers in the IGP area/level. In cases where the previous SPF Back-
Off Delay 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; over time, 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 using a given SPF Back-Off Delay algorithm with
another implementation using a different one.

8. Impact on Micro-loops

Micro-loops during IGP convergence are due to a non-synchronized or
non-ordered update of FIBs [RFC5715] [RFC6976] [SPF-MICRO]. FIBs are
installed after multiple steps, such as flooding of the IGP event
across the network, SPF wait time, SPF computation, FIB distribution
across line cards, and FIB update. This document only addresses the
contribution from the SPF wait time. 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 Distributed
Tunnels [RFC5715], Synchronized FIB Update [RFC5715], or the ordered
FIB approach [RFC6976].

9. IANA Considerations

   This document has no IANA actions.

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. References

11.1. 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>.

   [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>.

11.2. Informative References

   [ISO10589]
              International Organization for Standardization,
              "Information technology -- Telecommunications and
              information exchange between systems -- 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,
              November 2002.

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

   [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V.,
              and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP
              Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001,
              <https://www.rfc-editor.org/info/rfc3209>.

   [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A
              Border Gateway Protocol 4 (BGP-4)", RFC 4271,
              DOI 10.17487/RFC4271, January 2006,
              <https://www.rfc-editor.org/info/rfc4271>.

   [RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed.,
              "LDP Specification", RFC 5036, DOI 10.17487/RFC5036,
              October 2007, <https://www.rfc-editor.org/info/rfc5036>.

   [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,
              <https://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,
              <https://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, <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>.

   [SPF-MICRO]
              Litkowski, S., Decraene, B., and M. Horneffer, "Link State
              protocols SPF trigger and delay algorithm impact on IGP
              micro-loops", Work in Progress, draft-ietf-rtgwg-spf-
              uloop-pb-statement-07, May 2018.

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
   United States of America

   Email: acee@cisco.com

   Pierre Francois

   Email: pfrpfr@gmail.com

   Chris Bowers
   Juniper Networks, Inc.
   1194 N. Mathilda Ave.
   Sunnyvale, CA 94089
   United States of America

   Email: cbowers@juniper.net