Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 28 May 2024

K. Patel
Arrcus, Inc.
A. Lindem
LabN Consulting, LLC
S. Zandi
LinkedIn
W. Henderickx
Nokia

25 November 2023

BGP Link-State Shortest Path First (SPF) Routing
draft-ietf-lsvr-bgp-spf-29

Abstract

Many Massively Scaled Data Centers (MSDCs) have converged on simplified layer 3 routing. Furthermore, requirements for operational simplicity has led many of these MSDCs to converge on BGP as their single routing protocol for both their fabric routing and their Data Center Interconnect (DCI) routing. This document describes extensions to BGP to use BGP Link-State distribution and the Shortest Path First (SPF) algorithm. In doing this, it allows BGP to be efficiently used as both the underlay protocol and the overlay protocol in MSDCs.

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 28 May 2024.

Table of Contents

1. Introduction

Many Massively Scaled Data Centers (MSDCs) have converged on simplified layer 3 routing. Furthermore, requirements for operational simplicity has led many of these MSDCs to converge on BGP [RFC4271] as their single routing protocol for both their fabric routing and their Data Center Interconnect (DCI) routing [RFC7938]. This document describes an alternative solution which leverages BGP-LS [I-D.ietf-idr-rfc7752bis] and the Shortest Path First algorithm used by Internal Gateway Protocols (IGPs).

This document leverages both the BGP protocol [RFC4271] and the BGP-LS [I-D.ietf-idr-rfc7752bis] protocols. The relationship, as well as the scope of changes is described respectively in Section 2 and Section 3. The modifications to [RFC4271] for BGP SPF described herein only apply to IPv4 and IPv6 as underlay unicast Subsequent Address Families Identifiers (SAFIs). Operations for any other BGP SAFIs are outside the scope of this document.

This solution avails the benefits of both BGP and SPF-based IGPs. These include TCP-based flow-control, no periodic link-state refresh, and completely incremental NLRI advertisement. These advantages can reduce the overhead in MSDCs where there is a high degree of Equal Cost Multi-Path (ECMPs) and the topology is very stable. Additionally, using an SPF-based computation can support fast convergence and the computation of Loop-Free Alternatives (LFAs). The SPF LFA extensions defined in [RFC5286] can be similarly applied to BGP SPF calculations. However, the details are a matter of implementation detail. Furthermore, a BGP-based solution lends itself to multiple peering models including those incorporating route-reflectors [RFC4456] or controllers.

1.1. Terminology

This specification reuses terms defined in section 1.1 of [RFC4271] including BGP speaker, NLRI, and Route.

Additionally, this document introduces the following terms:

BGP SPF Routing Domain: A set of BGP routers that are under a single administrative domain and exchange link-state information using the BGP-LS-SPF SAFI and compute routes using BGP SPF as described herein.

BGP-LS-SPF NLRI: This refers to BGP-LS Network Layer Reachability Information (NLRI) that is being advertised in the BGP-LS-SPF SAFI (Section 5.1) and is being used for BGP SPF route computation.

Dijkstra Algorithm: An algorithm for computing the shortest path from a given node in a graph to every other node in the graph.

1.2. BGP Shortest Path First (SPF) Motivation

Given that [RFC7938] already describes how BGP could be used as the sole routing protocol in an MSDC, one might question the motivation for defining an alternate BGP deployment model when a mature solution exists. For both alternatives, BGP offers the operational benefits of a single routing protocol as opposed to the combination of an IGP for the underlay and BGP as an overlay. However, BGP SPF offers some

unique advantages above and beyond standard BGP distance-vector routing. With BGP SPF, the standard hop-by-hop peering model is relaxed.

A primary advantage is that all BGP SPF speakers in the BGP SPF routing domain have a complete view of the topology. This allows support for ECMP, IP fast-reroute (e.g., Loop-Free Alternatives), Shared Risk Link Groups (SRLGs), and other routing enhancements without advertisement of additional BGP paths [RFC7911] or other extensions.

With the BGP SPF decision process as defined in Section 6, NLRI changes can be disseminated throughout the BGP routing domain much more rapidly. The added advantage of BGP using TCP for reliable transport leverages TCP's inherent flow-control and guaranteed in-order delivery.

Another primary advantage is a potential reduction in NLRI advertisement. With standard BGP distance-vector routing, a single link failure may impact 100s or 1000s prefixes and result in the withdrawal or re-advertisement of the attendant NLRI. With BGP SPF, only the BGP SPF speakers corresponding to the link NLRI need to withdraw the corresponding BGP-LS-SPF Link NLRI. Additionally, the changed NLRI is advertised immediately as opposed to normal BGP where it is only advertised after the best route selection. These advantages provide NLRI dissemination throughout the BGP SPF routing domain with efficiencies similar to link-state protocols.

With controller and route-reflector peering models, BGP SPF advertisement and distributed computation require a minimal number of sessions and copies of the NLRI since only the latest version of the NLRI from the originator is required. Given that verification of the adjacencies is done outside of BGP (see Section 4), each BGP SPF speaker only needs as many sessions and copies of the NLRI as required for redundancy. Additionally, a controller could inject topology that is learned outside the BGP SPF routing domain.

Given BGP-LS NLRI is already consumed [I-D.ietf-idr-rfc7752bis], this functionality can be reused for BGP-LS-SPF NLRI.

Another advantage of BGP SPF is that both IPv6 and IPv4 can be supported using the BGP-LS-SPF SAFI with the same BGP-LS-SPF NLRIs. In many MSDC fabrics, the IPv4 and IPv6 topologies are congruent (refer to Section 5.2.2 and Section 5.2.3). Although beyond the scope of this document, multi-topology extensions could be used to support separate IPv4, IPv6, unicast, and multicast topologies while sharing the same NLRI.

Finally, the BGP SPF topology can be used as an underlay for other BGP SAFIs (using the existing model) and realize all the above advantages.

1.3. Document Overview

The document begins with sections defining the precise relationship that BGP SPF has with both the base BGP protocol [RFC4271] (Section 2) and the BGP Link-State (BGP-LS) extensions [I-D.ietf-idr-rfc7752bis] (Section 3). The BGP peering models, as well as their respective trade-offs are then discussed in Section 4. The remaining sections, which make up the bulk of the document, define the protocol enhancements necessary to support BGP SPF including BGP-LS Extensions (Section 5), replacement of the base BGP decision process with the SPF computation (Section 6), and BGP SPF error handling (Section 7).

1.4. 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. Base BGP Protocol Relationship

With the exception of the decision process, the BGP SPF extensions leverage the BGP protocol [RFC4271] without change. This includes the BGP protocol Finite State Machine, BGP messages and their encodings, processing of BGP messages, BGP attributes and path attributes, BGP NLRI encodings, and any error handling defined in [RFC4271] and [RFC7606].

Due to the changes to the decision process, there are mechanisms and encodings that are no longer applicable. Unless explicitly specified in the context of BGP SPF, all optional path attributes SHOULD NOT be advertised. If received, all path attributes MUST be accepted, validated, and propagated consistent with the BGP protocol [RFC4271], even if not needed by BGP SPF.

Section 9 of [RFC4271] defines the decision process that is used to select routes for subsequent advertisement by applying the policies in the local Policy Information Base (PIB) to the routes stored in its Adj-RIBs-In. The output of the Decision Process is the set of routes that are announced by a BGP speaker to its peers. These selected routes are stored by a BGP speaker in the speaker’s Adj-RIBs-Out according to policy.

The BGP SPF extension fundamentally changes the decision process, as described herein. Specifically:

1. BGP advertisements are readvertised to neighbors immediately without waiting or dependence on the route computation as specified in phase 3 of the base BGP decision process. Multiple peering models are supported as specified in Section 4.

2. Determining the degree of preference for BGP routes for the SPF calculation as described in phase 1 of the base BGP decision process is replaced with the mechanisms in Section 6.1.

3. Phase 2 of the base BGP protocol decision process is replaced with the Shortest Path First (SPF) algorithm, also known as the Dijkstra algorithm.

3. BGP Link-State (BGP-LS) Relationship

[I-D.ietf-idr-rfc7752bis] describes a mechanism by which link-state and Traffic Engineering (TE) information can be collected from networks and shared with external entities using BGP. This is achieved by defining NLRI advertised using the BGP-LS AFI. The BGP-LS extensions defined in [I-D.ietf-idr-rfc7752bis] make use of the decision process defined in [RFC4271]. Rather than reusing the BGP-LS SAFI, the BGP-LS-SPF SAFI (Section 5.1) is introduced to insure backward compatibility for the BGP-LS SAFI usage.

The BGP SPF extensions reuse the format of the Link-State NLRI, the BGP-LS Attribute, and the TLVs defined in [I-D.ietf-idr-rfc7752bis]. The usage of is described in Section 5.2. The usage of other BGP-LS TLVs or extensions is not precluded and is, in fact, expected. However, the details are beyond the scope of this document and may be specified in future documents.

4. BGP SPF Peering Models

Depending on the topology, scaling, capabilities of the BGP SPF speakers, and redundancy requirements, various peering models are supported. The only requirement is that all BGP SPF speakers in the BGP SPF routing domain adhere to this specification.

4.1. BGP Single-Hop Peering on Network Node Connections

The simplest peering model is the one where EBGP single-hop sessions are established over direct point-to-point links interconnecting the nodes in the BGP SPF routing domain. Once the single-hop BGP session has been established and the Multi-Protocol Extensions Capability with the BGP-LS-SPF AFI/SAFI has been exchanged [RFC4760] for the corresponding session, then the link is considered up from a BGP SPF perspective and the corresponding BGP-LS-SPF Link NLRI is advertised.

An End-of-RIB (EoR) Marker [RFC4724] for the BGP-LS-SPF SAFI MAY be expected prior to advertising the BGP-LS Link NLRI for to peer.

A failure to consistently configure the use of the EoR marker can result in transient micro-loops and dropped traffic due to incomplete forwarding state.

If the session goes down, the corresponding Link NLRI are withdrawn. Topologically, this would be equivalent to the peering model in [RFC7938] where there is a BGP session on every link in the data center switch fabric. The content of the Link NLRI is described in Section 5.2.2.

4.2. BGP Peering Between Directly-Connected Nodes

In this model, BGP SPF speakers peer with all directly-connected nodes but the sessions may be between loopback addresses (i.e., two-hop sessions) and the direct connection discovery and liveliness detection for the interconnecting links are independent of the BGP protocol. However, the BFD protocol [RFC5880] is RECOMMENDED for liveliness detection. Usage of other liveliness connection mechanisms is outside the scope of this document. Consequently, there is a single BGP session even if there are multiple direct connections between BGP SPF speakers. The BGP-LS-SPF Link NLRI is advertised as long as a BGP session has been established, the BGP-LS-SPF AFI/SAFI capability has been exchanged [RFC4760], the link is operational as determined using liveliness detection mechanisms, and, optionally, the EoR Marker has been received as described in the Section 4.1. This is much like the previous peering model only peering is between loopback addresses and the interconnecting links can be unnumbered. However, since there are BGP sessions between every directly-connected node in the BGP SPF routing domain, there is a reduction in BGP sessions when there are parallel links between nodes. Hence, this peering model is RECOMMENDED over the single-hop peering model Section 4.1.

An End-of-RIB (EoR) Marker [RFC4724] for the BGP-LS-SPF SAFI MAY also be expected prior to advertising the BGP-LS Link NLRI for the link(s) to this peer.

4.3. BGP Peering in Route-Reflector or Controller Topology

In this model, BGP SPF speakers peer solely with one or more Route Reflectors [RFC4456] or controllers. As in the previous model, direct connection discovery and liveliness detection for those links in the BGP SPF routing domain are done outside of the BGP protocol. BGP-LS-SPF Link NLRI is advertised as long as the corresponding link is considered up as per the chosen liveness detection mechanism.

This peering model, known as sparse peering, allows for fewer BGP sessions and, consequently, fewer instances of the same NLRI received from multiple peers. Normally, the route-reflectors or controller BGP sessions would be on directly-connected links to avoid dependence on another routing protocol for session connectivity. However, multi-hop peering is not precluded. The number of BGP sessions is dependent on the redundancy requirements and the stability of the BGP sessions. This is discussed in greater detail in [I-D.ietf-lsvr-applicability].

The controller may use constraints to determine when to advertise BGP-LS-SPF NLRI for BGP-LS peers. For example, a controller may defer advertisement until the EoR marker has been received from both BGP peers and both have received each other's NLRI. These constraints are outside the scope of this document and, since they are internal to the controller, need not be standardized.

5. BGP Shortest Path Routing (SPF) Protocol Extensions

5.1. BGP-LS Shortest Path Routing (SPF) SAFI

This document introduces the BGP-LS-SPF SAFI with a value of 80. The SPF-based decision process (Section 6) applies only to the BGP-LS-SPF SAFI and MUST NOT be used with other combinations of the BGP-LS AFI (16388). In order for two BGP SPF speakers to exchange BGP-LS-SPF NLRI, they MUST exchange the Multiprotocol Extensions Capability [RFC4760] to ensure that they are both capable of properly processing such NLRI. This is done with AFI 16388 / SAFI 80. The BGP-LS-SPF SAFI is used to advertise IPv4 and IPv6 prefix information in a format facilitating an SPF-based decision process.

5.1.1. BGP-LS-SPF NLRI TLVs

All the TLVs defined for BGP-LS [I-D.ietf-idr-rfc7752bis] are applicable and can be used with the BGP-LS-SPF SAFI to describe links, nodes, and prefixes comprising IGP link-state information.

The NLRI and comprising TLVs MUST be processed as specified in section 5.1 [I-D.ietf-idr-rfc7752bis]. TLVs specified as mandatory in [I-D.ietf-idr-rfc7752bis] are considered mandatory for the BGP-LS-SPF SAFI as well. If a mandatory TLV is not present, the NLRI MUST NOT be used in the BGP SPF route calculation. All the other TLVs are considered as optional TLVs.

5.1.2. BGP-LS Attribute

The BGP-LS attribute of the BGP-LS-SPF SAFI uses exactly same format of the BGP-LS AFI [I-D.ietf-idr-rfc7752bis]. In other words, all the TLVs used in the BGP-LS attribute of the BGP-LS AFI are applicable and used for the BGP-LS attribute of the BGP-LS-SPF SAFI. This attribute is an optional, non-transitive BGP attribute that is used to carry link, node, and prefix properties and attributes. The BGP-LS attribute is a set of TLVs.

All the TLVs defined for the BGP-LS Attribute [I-D.ietf-idr-rfc7752bis] are applicable and can be used with the BGP-LS-SPF SAFI to carry link, node, and prefix properties and attributes.

The BGP-LS attribute may potentially be quite large depending on the amount of link-state information associated with a single Link- State NLRI. The BGP specification [RFC4271] mandates a maximum BGP message size of 4096 octets. It is RECOMMENDED that an implementation support [RFC8654] in order to accommodate a greater amount of information within the BGP-LS Attribute. BGP SPF speakers MUST ensure that they limit the TLVs included in the BGP-LS Attribute to ensure that a BGP update message for a single Link-State NLRI does not cross the maximum limit for a BGP message. The determination of the types of TLVs to be included by the BGP SPF speaker originating the attribute is outside the scope of this document. When a BGP SPF speaker finds that it is exceeding the maximum BGP message size due to addition or update of some other BGP Attribute (e.g., AS_PATH), it MUST consider the BGP-LS Attribute to be malformed and the attribute discard handling of [RFC7606] applies.

5.2. Extensions to BGP-LS

[I-D.ietf-idr-rfc7752bis] describes a mechanism by which link-state and TE information can be collected from IGPs and shared with external components using the BGP protocol. It describes both the definition of the BGP-LS NLRI that advertise links, nodes, and prefixes comprising IGP link-state information and the definition of a BGP path attribute (BGP-LS attribute) that carries link, node, and prefix properties and attributes, such as the link and prefix metric or auxiliary Router-IDs of nodes, etc. This document extends the usage of BGP-LS NLRI for the purpose of BGP SPF calculation via advertisement in the BGP-LS-SPF SAFI.

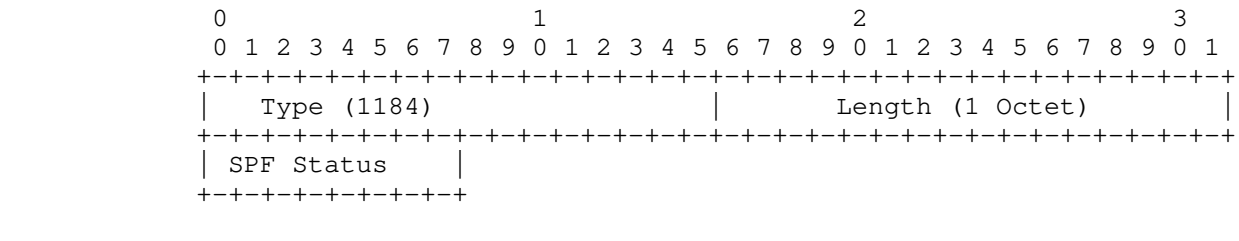
The protocol identifier specified in the Protocol-ID field [I-D.ietf-idr-rfc7752bis] represents the origin of the advertised NLRI. For Node NLRI and Link NLRI, the specified Protocol-ID MUST be the direct protocol (4). Node or Link NLRI with a Protocol-ID other than the direct protocol is considered malformed. For Prefix NLRI, the specified Protocol-ID MUST be the origin of the prefix. The local and remote node descriptors for all NLRI MUST include the BGP Identifier (TLV 516) [RFC9086] and the AS Number (TLV 512) [I-D.ietf-idr-rfc7752bis]. The BGP Confederation Member (TLV 517) [RFC9086] is currently not applicable.

5.2.1. Node NLRI Usage

The Node NLRI MUST be advertised unconditionally by all routers in the BGP SPF routing domain.

5.2.1.1. BGP-LS-SPF Node NLRI Attribute SPF Status TLV

A BGP-LS Attribute TLV of the BGP-LS-SPF Node NLRI is defined to indicate the status of the node with respect to the BGP SPF calculation. This is used to rapidly take a node out of service (refer to Section 6.5.2) or to indicate the node is not to be used for transit (i.e., non-local) traffic (refer to Section 6.3). If the SPF Status TLV is not included with the Node NLRI, the node is considered to be up and is available for transit traffic. The SPF status is acted upon with the execution of the next SPF calculation (refer to Section 6.3).



SPF Status Values: 0 - Reserved
1 - Node unreachable with respect to BGP SPF
2 - Node does not support transit with respect to BGP SPF
3-254 - Undefined
255 - Reserved

The BGP-LS-SPF Node Attribute SPF Status TLV, Link Attribute SPF Status TLV, and Prefix Attribute SPF Status TLV use the same TLV Type (1184).

If a BGP SPF speaker received the Node NLRI but the SPF Status TLV is not received, then any previously received information is considered as implicitly withdrawn and the update is propagated to other BGP SPF speakers. A BGP SPF speaker receiving a BGP Update containing a SPF Status TLV in the BGP-LS attribute [I-D.ietf-idr-rfc7752bis] with a value that is undefined SHOULD be advertised to other BGP SPF speakers. However, a BGP SPF speaker MUST NOT use the Status TLV in its SPF computation. An implementation MAY log this condition for further analysis.

5.2.2. Link NLRI Usage

The criteria for advertisement of Link NLRI are discussed in Section 4.

Link NLRI is advertised with unique local and remote node descriptors dependent on the IP addressing. For IPv4 links, the link's local IPv4 (TLV 259) and remote IPv4 (TLV 260) addresses are used. For IPv6 links, the local IPv6 (TLV 261) and remote IPv6 (TLV 262) addresses are used. For links supporting having both IPv4 and IPv6 addresses, both sets of descriptors MAY be included in the same Link NLRI.

For unnumbered links, the Link Local/Remote Identifiers (TLV 258) are used. The Link Remote Identifier isn't normally exchanged in BGP and discovering the Link Remote Identifier is beyond the scope of this document. If the Link Remote Identifier is unknown, a Link Remote Identifier of 0 MUST be advertised. When 0 is advertised and there parallel unnumbered links between a pair of BGP SPF speakers, there

may be transient intervals where the BGP SPF speakers don't agree on which of the parallel unnumbered links are operational. For this reason, it is RECOMMENDED that the Link Remote Identifiers be known (e.g., discovered using alternate mechanisms or configured) in the presence of parallel unnumbered links.

The link descriptors are described in table 4 of [I-D.ietf-idr-rfc7752bis]. Additionally, an address family link descriptor is defined to determine whether an unnumbered link can be used in the IPv4 SPF, the IPv6, or both (refer to section Section 5.2.2.1).

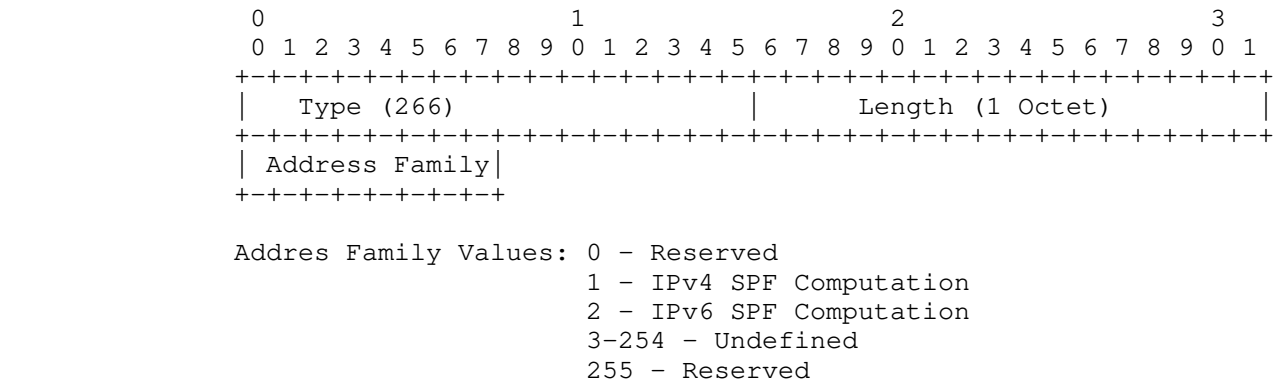
For a link to be used in SPF computation for a given address family, i.e., IPv4 or IPv6, both routers connecting the link MUST have matching addresses (i.e., interface addresses must match the neighbor addresses).

The IGP metric attribute TLV (TLV 1095) MUST be advertised. If a BGP SPF speaker receives a Link NLRI without an IGP metric attribute TLV, then it MUST consider the received NLRI as a malformed and the receiving BGP SPF speaker MUST handle such malformed NLRI as 'Treat-as-withdraw' [RFC7606]. The BGP SPF metric length is 4 octets. A metric is associated with the output side of each router interface. This metric is configurable by the system administrator. The lower the metric, the more likely the interface is to be used to forward data traffic. One possible default for metric would be to give each interface a metric of 1 making it effectively a hop count.

The usage of other link attribute TLVs is beyond the scope of this document.

5.2.2.1. BGP-LS Link NLRI Adddress Family Link Descriptor TLV

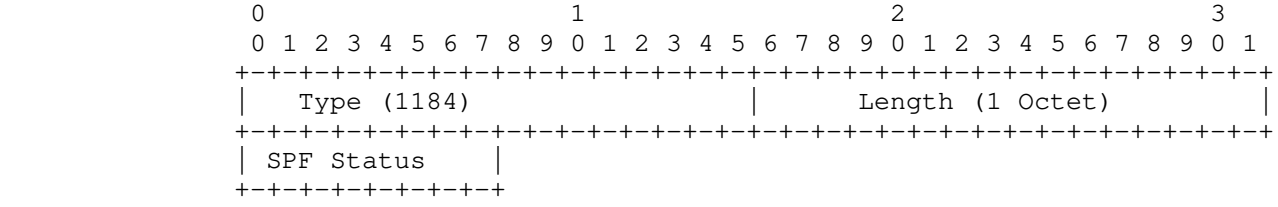
For unnumberred links, the address family cannot be ascertained from the endpoint link descriptors. Hence, the Address Family (AF) Link Descriptor SHOULD be included with the Link Local/Remote Identifiers TLV so that the link can be used in the respective address family SPF. If the Address Family Link Descriptor is not present for an unnumbered link, the link will not be used in the SPF computation for either address family. If the Address Family Link Descriptor is present for a numbered link, the link descriptor will be ignored.



Addres Family Values: 0 - Reserved
1 - IPv4 SPF Computation
2 - IPv6 SPF Computation
3-254 - Undefined
255 - Reserved

5.2.2.2. BGP-LS-SPF Link NLRI Attribute SPF Status TLV

This BGP-LS-SPF Attribute TLV of the BGP-LS-SPF Link NLRI is defined to indicate the status of the link with respect to the BGP SPF calculation. This is used to expedite convergence for link failures as discussed in Section 6.5.1. If the SPF Status TLV is not included with the Link NLRI, the link is considered up and available. The SPF status is acted upon with the execution of the next SPF calculation Section 6.3. A single TLV type is shared by the Node, Link, and Prefix NLRI. The TLV type is 1184.



BGP Status Values: 0 - Reserved
1 - Link Unreachable with respect to BGP SPF
2-254 - Undefined
255 - Reserved

The BGP-LS-SPF Node Attribute SPF Status TLV, Link Attribute SPF Status TLV, and Prefix Attribute SPF Status TLV use the same TLV Type (1184). This implies that a BGP Update cannot contain multiple NLRI.

If a BGP SPF speaker received the Link NLRI but the SPF Status TLV is not received, then any previously received information is considered as implicitly withdrawn and the update is propagated to other BGP SPF speakers. A BGP SPF speaker receiving a BGP Update containing an SPF

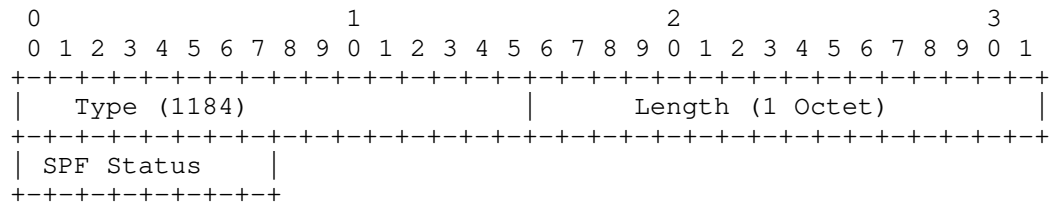
Status TLV in the BGP-LS attribute [I-D.ietf-idr-rfc7752bis] with a value that is undefined SHOULD be advertised to other BGP SPF speakers. However, a BGP SPF speaker MUST NOT use the Status TLV in its SPF computation. An implementation MAY log this information for further analysis.

5.2.3. IPv4/IPv6 Prefix NLRI Usage

IPv4/IPv6 Prefix NLRI is advertised with a Local Node Descriptor and the prefix and length. The Prefix Descriptors field includes the IP Reachability Information TLV (TLV 265) as described in [I-D.ietf-idr-rfc7752bis]. The Prefix Metric TLV (TLV 1155) MUST be advertised. The IGP Route Tag TLV (TLV 1153) MAY be advertised. The usage of other BGP-LS attribute TLVs is beyond the scope of this document.

5.2.3.1. BGP-LS-SPF Prefix NLRI Attribute SPF Status TLV

A BGP-LS Attribute TLV to BGP-LS-SPF Prefix NLRI is defined to indicate the status of the prefix with respect to the BGP SPF calculation. This is used to expedite convergence for prefix unreachability as discussed in Section 6.5.1. If the SPF Status TLV is not included with the Prefix NLRI, the prefix is considered reachable. A single TLV type is shared by the Node, Link, and Prefix NLRI. The TLV type is 1184.



BGP Status Values: 0 - Reserved
1 - Prefix Unreachable with respect to SPF
2-254 - Undefined
255 - Reserved

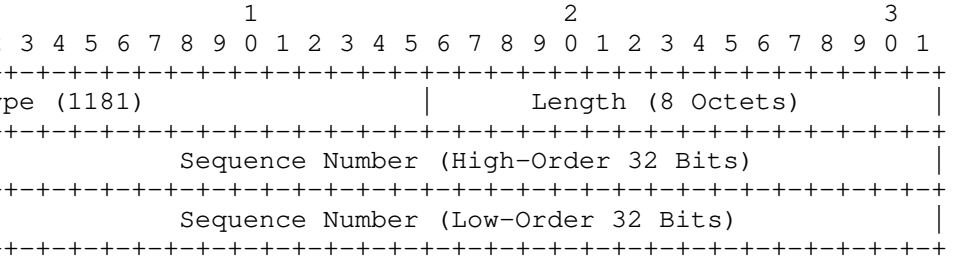
The BGP-LS-SPF Node Attribute SPF Status TLV, Link Attribute SPF Status TLV, and Prefix Attribute SPF Status TLV use the same TLV Type (1184). This implies that a BGP Update cannot contain multiple NLRI.

If a BGP SPF speaker received the Prefix NLRI but the SPF Status TLV is not received, then any previously received information is considered as implicitly withdrawn and the update is propagated to other BGP SPF speakers. A BGP SPF speaker receiving a BGP Update containing an SPF Status TLV in the BGP-LS attribute

[I-D.ietf-idr-rfc7752bis] with a value that is undefined SHOULD be advertised to other BGP SPF speakers. However, a BGP SPF speaker MUST NOT use the Status TLV in its SPF computation. An implementation MAY log this information for further analysis.

5.2.4. BGP-LS Attribute Sequence-Number TLV

A BGP-LS Attribute TLV of the BGP-LS-SPF NLRI types is defined to assure the most recent version of a given NLRI is used in the SPF computation. The Sequence-Number TLV is mandatory for BGP-LS-SPF NLRI. The TLV type 1181 has been assigned by IANA. The BGP-LS Attribute TLV contains an 8-octet sequence number. The usage of the Sequence Number TLV is described in Section 6.1.



Sequence Number: The 64-bit strictly-increasing sequence number MUST be incremented for every self-originated version of a BGP-LS-SPF NLRI. BGP SPF speakers implementing this specification MUST use available mechanisms to preserve the sequence number's strictly increasing property for the deployed life of the BGP SPF speaker (including cold restarts). One mechanism for accomplishing this would be to use the high-order 32 bits of the sequence number as a wrap/boot count that is incremented any time the BGP router loses its sequence number state or the low-order 32 bits wrap.

When incrementing the sequence number for each self-originated NLRI, the sequence number should be treated as an unsigned 64-bit value. If the lower-order 32-bit value wraps, the higher-order 32-bit value should be incremented and saved in non-volatile storage. If a BGP SPF speaker completely loses its sequence number state (e.g., the BGP SPF speaker hardware is replaced or experiences a cold-start), the BGP NLRI selection rules (see Section 6.1) insure convergence, albeit not immediately.

If the Sequence-Number TLV is not received, then the corresponding NLRI is considered as malformed and MUST be handled as 'Treat-as-withdraw'. An implementation MAY log an error for further analysis.

5.3. NEXT_HOP Attribute Manipulation

The rules for setting the next hop information for the BGP-LS-SPF SAFI follow the specification in section 5.5 of [I-D.ietf-idr-rfc7752bis]. All BGP peers that support SPF extensions will locally compute the Local-RIB Next-Hop as a result of the SPF process.

6. Decision Process with SPF Algorithm

The Decision Process described in [RFC4271] takes place in three distinct phases. The Phase 1 decision function of the Decision Process is responsible for calculating the degree of preference for each route received from a BGP SPF speaker's peer. The Phase 2 decision function is invoked on completion of the Phase 1 decision function and is responsible for choosing the best route out of all those available for each distinct destination, and for installing each chosen route into the Local-RIB. The combination of the Phase 1 and 2 decision functions is characterized as a Path Vector algorithm.

The SPF-based Decision process replaces the BGP Decision process described in [RFC4271]. Since Link-State NLRI always contains the local node descriptor as described in Section 5.2, each NLRI is uniquely originated by a single BGP SPF speaker in the BGP SPF routing domain (the BGP node matching the NLRI's Node Descriptors). Instances of the same NLRI originated by multiple BGP SPF speakers would be indicative of a configuration error or a masquerading attack (refer to Section 9). These selected Node NLRI and their Link/Prefix NLRI are used to build a directed graph during the SPF computation as described below. The best routes for BGP prefixes are installed in the RIB as a result of the SPF process.

When BGP-LS-SPF NLRI is received, all that is required is to determine whether it is the most recent by examining the Node-ID and sequence number as described in Section 6.1. If the received NLRI has changed, it is advertised to other BGP-LS-SPF peers. If the attributes have changed (other than the sequence number), a BGP SPF calculation is triggered. However, a changed NLRI MAY be advertised immediately to other peers and prior to any SPF calculation. Note that the BGP MinRouteAdvertisementIntervalTimer and MinASOriginationIntervalTimer [RFC4271] timers are not applicable to the BGP-LS-SPF SAFI. The scheduling of the SPF calculation, as described in Section 6.3, is an implementation issue. Scheduling MAY be dampened consistent with the SPF back-off algorithm specified in [RFC8405].

The Phase 3 decision function of the Decision Process [RFC4271] is also simplified since under normal SPF operation, a BGP SPF speaker MUST advertise the changed NLRIs to all BGP peers with the BGP-LS-SPF AFI/SAFI and install the changed routes in the GLOBAL-RIB. The only exception are unchanged NLRIs or stale NLRIs, i.e., NLRI received with a less recent (numerically smaller) sequence number.

6.1. BGP SPF NLRI Selection

The rules for all BGP-LS-SPF NLRIs selection for phase 1 of the BGP decision process, section 9.1.1 [RFC4271], no longer apply.

1. NLRI originated by directly connected BGP SPF peers are preferred. This condition can be determined by comparing the BGP Identifiers in the received Local Node Descriptor and the BGP OPEN message for an active BGP session. This rule assures that stale NLRI is updated even if a BGP-LS router loses its sequence number state due to a cold-start. Note that once the BGP session goes down, the NLRI received is no longer considered as being from a directly connected BGP SPF peer.

2. The NLRI with the most recent Sequence Number TLV, i.e., highest sequence number is selected.

3. The NLRI received from the BGP SPF speaker with the numerically larger BGP Identifier is preferred.

When a BGP SPF speaker completely loses its sequence number state, i.e., due to a cold start, or in the unlikely possibility that 64-bit sequence number wraps, the BGP routing domain will still converge. This is due to the fact that BGP SPF speakers adjacent to the router always accept self-originated NLRI from the associated speaker as more recent (rule # 1). When a BGP SPF speaker reestablishes a connection with its peers, any existing sessions are taken down and stale NLRI are replaced. The adjacent BGP SPF speakers update their NLRI advertisements and advertise to their neighbors until the BGP routing domain has converged.

The modified SPF Decision Process performs an SPF calculation rooted at the local BGP SPF speaker using the metrics from the Link Attribute IGP Metric TLV (1095) and the Prefix Attribute Prefix Metric TLV (1155) [I-D.ietf-idr-rfc7752bis]. As a result, any other BGP attributes that would influence the BGP decision process defined in [RFC4271] including ORIGIN, MULTI_EXIT_DISC, and LOCAL_PREF attributes are ignored by the SPF algorithm. The NEXT_HOP attribute is discussed in Section 5.3. The AS_PATH and AS4_PATH [RFC6793] attributes are preserved and used for loop detection [RFC4271]. They are ignored during the SPF computation for BGP-LS-SPF NLRIs.

6.1.1. BGP Self-Originated NLRI

Node, Link, or Prefix NLRI with Node Descriptors matching the local BGP SPF speaker are considered self-originated. When self-originated NLRI is received and it doesn't match the local node's NLRI content (including sequence number), special processing is required.

* If self-originated NLRI is received and the sequence number is more recent (i.e., greater than the local node's sequence number for the NLRI), the NLRI sequence number is advanced to one greater than the received sequence number and the NLRI is readvertised to all peers.

* If self-originated NLRI is received and the sequence number is the same as the local node's sequence number but the attributes differ, the NLRI sequence number is advanced to one greater than the received sequence number and the NLRI is readvertised to all peers.

* If self-originated Link or Prefix NLRI is received and the Link or Prefix NLRI is no longer being advertised by the local node, the NLRI is withdrawn.

The above actions are performed immediately when the first instance of a newer self-originated NLRI is received. In this case, the newer instance is considered to be a stale instance that was advertised by the local node prior to a restart where the NLRI state was lost. However, if subsequent newer self-originated NLRI is received for the same Node, Link, or Prefix NLRI, the readvertisement or withdrawal is delayed by 5 seconds since it is likely being advertised by a misconfigured or rogue BGP SPF speaker (refer to Section 9).

6.2. Dual Stack Support

The SPF-based decision process operates on Node, Link, and Prefix NLRIs that support both IPv4 and IPv6 addresses. Whether to run a single SPF computation or multiple SPF computations for separate AFs is an implementation matter. Normally, IPv4 next-hops are calculated for IPv4 prefixes and IPv6 next-hops are calculated for IPv6 prefixes.

6.3. SPF Calculation based on BGP-LS-SPF NLRI

This section details the BGP-LS-SPF local routing information base (RIB) calculation. The router uses BGP-LS-SPF Node, Link, and Prefix NLRI to compute routes using the following algorithm. This calculation yields the set of routes associated with the BGP SPF Routing Domain. A router calculates the shortest-path tree using itself as the root. Optimizations to the BGP-LS-SPF algorithm are possible but MUST yield the same set of routes. The algorithm below supports Equal Cost Multi-Path (ECMP) routes. Weighted Unequal Cost Multi-Path routes are out of scope.

The following abstract data structures are defined in order to specify the algorithm.

* Local Route Information Base (Local-RIB) - This routing table contains reachability information (i.e., next hops) for all prefixes (both IPv4 and IPv6) as well as BGP-LS-SPF node reachability. Implementations may choose to implement this with separate RIBs for each address family and/or Prefix versus Node reachability.

* Global Routing Information Base (GLOBAL-RIB) - This is the Routing Information Base (RIB) containing the current routes that are installed in the router's forwarding plane. This is commonly referred to in networking parlance as "the RIB".

* Link State NLRI Database (LSNDB) - Database of BGP-LS-SPF NLRI that facilitates access to all Node, Link, and Prefix NLRI.

* Candidate List (CAN-LIST) - This is a list of candidate Node NLRIs used during the BGP SPF calculation. The list is sorted by the cost to reach the Node NLRI with the Node NLRI with the lowest reachability cost at the head of the list. This facilitates execution of the Dijkstra algorithm where the shortest paths between the local node and other nodes in graph are computed. The CAN-LIST is typically implemented as a heap but other data structures have been used.

The algorithm is comprised of the steps below:

1. The current Local-RIB is invalidated, and the CAN-LIST is initialized to empty. The Local-RIB is rebuilt during the course of the SPF computation. The existing routing entries are preserved for comparison to determine changes that need to be made to the GLOBAL-RIB in step 6. These routes are referred to as stale routes.

2. The computing router's Node NLRI is updated in the Local-RIB with a cost of 0 and the Node NLRI is also added to the CAN-LIST. The next-hop list is set to the internal loopback next-hop.

3. The Node NLRI with the lowest cost is removed from the CAN-LIST for processing. If the BGP-LS Node attribute includes an SPF Status TLV (refer to Section 5.2.1.1) indicating the node is unreachable, the Node NLRI is ignored and the next lowest cost Node NLRI is selected from the CAN-LIST. The Node corresponding to this NLRI is referred to as the Current-Node. If the CAN-LIST list is empty, the SPF calculation has completed and the algorithm proceeds to step 6.

4. All the Prefix NLRI with the same Local Node Descriptors as the Current-Node are considered for installation. The next-hop(s) for these Prefix NLRI are inherited from the Current-Node. If the Current-Node is for the local BGP Router, the next-hop for the prefix is a direct next-hop. The cost for each prefix is the metric advertised in the Prefix Attribute Prefix Metric TLV (1155) added to the cost to reach the Current-Node. The following is done for each Prefix NLRI (referred to as the Current-Prefix):

   * If the BGP-LS Prefix attribute includes an SPF Status TLV indicating the prefix is unreachable, the Current-Prefix is considered unreachable and the next Prefix NLRI is examined in Step 4.

   * If the Current-Prefix's corresponding prefix is in the Local-RIB and the Local-RIB metric is less than the Current-Prefix's metric, the Current-Prefix does not contribute to the route and the next Prefix NLRI is examined in Step 4.

   * If the Current-Prefix's corresponding prefix is not in the Local-RIB, the prefix is installed with the Current-Node's next-hops installed as the Local-RIB route's next-hops and the metric being updated. If the IGP Route Tag TLV (1153) is included in the Current-Prefix's NLRI Attribute, the tag(s) are installed in the current Local-RIB route's tag(s).

   * If the Current-Prefix's corresponding prefix is in the Local-RIB and the cost is less than the Local-RIB route's metric, the prefix is installed with the Current-Node's next-hops replacing the Local-RIB route's next-hops and the metric being updated and any route tags removed. If the IGP Route Tag TLV (1153) is included in the Current-Prefix's NLRI Attribute, the tag(s) are installed in the current Local-RIB route's tag(s).

* If the Current-Prefix's corresponding prefix is in the Local-RIB and the cost is the same as the Local-RIB route's metric, the Current-Node's next-hops are merged with Local-RIB route's next-hops. The algorithm below supports Equal Cost Multi-Path (ECMP) routes. Some platforms or implementations may have limits on the number of ECMP routes that can be supported. The setting or identification of any limitations is outside the scope if this document. Nonetheless, step 4 (below) includes a set of recommendations in case such a limit is encountered. Weighted Unequal Cost Multi-Path routes are out of scope as well.

5. All the Link NLRI with the same Node Identifiers as the Current-Node are considered for installation. Each link is examined and is referred to in the following text as the Current-Link. The cost of the Current-Link is the advertised IGP Metric TLV (1095) from the Link NLRI BGP-LS attribute added to the cost to reach the Current-Node. If the Current-Node is for the local BGP Router, the next-hop for the link is a direct next-hop pointing to the corresponding local interface. For any other Current-Node, the next-hop(s) for the Current-Link are inherited from the Current-Node. The following is done for each link:

   a. If the Current-Link's NLRI attribute includes an SPF Status TLV indicating the link is down, the BGP-LS-SPF Link NLRI is considered down and the next link for the Current-Node is examined in Step 5.

   b. If the Current-Node NLRI attributes includes the SPF Status TLV (refer to Section 5.2.1.1) and the status indicates that the Node doesn't support transit, the next link for the Current-Node is processed in Step 5.

   c. The Current-Link's Remote Node NLRI is accessed (i.e., the Node NLRI with the same Node identifiers as the Current-Link's Remote Node Descriptors). If it exists, it is referred to as the Remote-Node and the algorithm proceeds as follows:

      * If the Remote-Node's NLRI attribute includes an SPF Status TLV indicating the node is unreachable, the next link for the Current-Node is examined in Step 5.

* All the Link NLRI corresponding the Remote-Node are searched for a Link NLRI pointing to the Current-Node. Each Remote-Node's Link NLRI (referred to as the Remote-Link) is examined for Remote Node Descriptors matching the Current-Node and Link Descriptors matching the Current-Link.

  - For IPv4/IPv6 numbered Link Descriptors to match during the IPv4 SPF computation, the Current-Link's IP4/IPv6 interface address link descriptor MUST match the Remote-Link IPv4/IPv6 neighbor address link descriptor and the Current-Link's IPv4/IPv6 neighbor address MUST match the Remode-Link's IPv4/IPv6 interface address.

  - For unnumbered links to match during the IPv4 or IPv6 SPF computation, Current-Link and Remote-Link's Address Family link descriptor must match address family of the IPv4 or IPv6 SPF computation, the Current--Link's Local Identifier MUST match the Remote-Link's Remote Identifier, and the Current-Link's Remote Identifier MUST the Remote-Link's Local Identifier. Since the Link's Remote Identifier may not be known, a value of 0 is considered a wildcard and will match any Current or Remote Link's Local Identifier (see TLV 258 [I-D.ietf-idr-rfc7752bis]).

  If these conditions are satisfied for one of the Remote-Node's links, the bi-directional connectivity check succeeds and the Remote-Node may be processed further. The Remote-Node's Link NLRI providing bi-directional connectivity is referred to as the Remote-Link. If no Remote-Link is found, the next link for the Current-Node is examined in Step 5.

* If the Remote-Link NLRI attribute includes an SPF Status TLV indicating the link is down, the Remote-Link NLRI is considered down and the next link for the Current-Node is examined in Step 5.

* If the Remote-Node is not on the CAN-LIST, it is inserted based on the cost. The Remote Node's cost is the cost of Current-Node added the Current-Link's IGP Metric TLV (1095). The next-hop(s) for the Remote-Node are inherited from the Current-Link.

* If the Remote-Node NLRI is already on the CAN-LIST with a higher cost, it must be removed and reinserted with the Remote-Node cost based on the Current-Link (as calculated in the previous step). The next-hop(s) for the Remote-Node are inherited from the Current-Link.

* If the Remote-Node NLRI is already on the CAN-LIST with the same cost, it need not be reinserted on the CAN-LIST. However, the Current-Link's next-hop(s) must be merged into the current set of next-hops for the Remote-Node.

* If the Remote-Node NLRI is already on the CAN-LIST with a lower cost, it need not be reinserted on the CAN-LIST.

d. Return to step 3 to process the next lowest cost Node NLRI on the CAN-LIST.

6. The Local-RIB is examined and changes (adds, deletes, modifications) are installed into the GLOBAL-RIB. For each route in the Local-RIB:

* If the route was added during the current BGP SPF computation, install the route into the GLOBAL-RIB.

* If the route modified during the current BGP SPF computation (e.g., metric, tags, or next-hops), update the route in the GLOBAL-RIB.

* If the route was not installed during the current BGP SPF computation, remove the route from the GLOBAL-RIB.

6.4. IPv4/IPv6 Unicast Address Family Interaction

While the BGP-LS-SPF address family and the BGP unicast address families may install routes into the same device routing tables, they operate independently (i.e., "Ships-in-the-Night" mode). There is no implicit route redistribution between the BGP-LS-SPF address family and the BGP unicast address families.

It is RECOMMENDED that BGP-LS-SPF IPv4/IPv6 route computation and installation be given scheduling priority by default over other BGP address families as these address families are considered as underlay SAFIs.

6.5. NLRI Advertisement

6.5.1. Link/Prefix Failure Convergence

A local failure prevents a link from being used in the SPF calculation due to the IGP bi-directional connectivity requirement. Consequently, local link failures SHOULD always be given priority over updates (e.g., withdrawing all routes learned on a session) in order to ensure the highest priority propagation and optimal convergence.

With a BGP advertisement, the link would continue to be used until the last copy of the BGP-LS-SPF Link NLRI is withdrawn. In order to avoid this delay, the originator of the Link NLRI SHOULD advertise a more recent version with an increased Sequence Number TLV for the BGP-LS-SPF Link NLRI including the SPF Status TLV (refer to Section 5.2.2.2) indicating the link is down with respect to BGP SPF. The configurable LinkStatusDownAdvertise timer controls the interval that the BGP-LS-LINK NLRI is advertised with SPF Status indicating the link is down prior to withdrawal. If BGP-LS-SPF Link NLRI has been advertised with the SPF Status TLV and the link becomes available in that period, the originator of the BGP-LS-SPF LINK NLRI MUST advertise a more recent version of the BGP-LS-SPF Link NLRI without the SPF Status TLV in the BGP-LS Link Attributes. The suggested default value for the LinkStatusDownAdvertise timer is 2 seconds.

Similarly, when a prefix becomes unreachable, a more recent version of the BGP-LS-SPF Prefix NLRI SHOULD be advertised with the SPF Status TLV (refer to Section 5.2.3.1) indicating the prefix is unreachable in the BGP-LS Prefix Attributes and the prefix will be considered unreachable with respect to BGP SPF. The configurable PrefixStatusDownAdvertise timer controls the interval that the BGP-LS-Prefix NLRI is advertised with SPF Status indicating the prefix is unreachable prior to withdrawal. If the BGP-LS-SPF Prefix has been advertised with the SPF Status TLV and the prefix becomes reachable in that period, the originator of the BGP-LS-SPF Prefix NLRI MUST advertise a more recent version of the BGP-LS-SPF Prefix NLRI without the SPF Status TLV in the BGP-LS Prefix Attributes. The suggested default value for the PrefixStatusDownAdvertise timer is 2 seconds.

6.5.2. Node Failure Convergence

By default [RFC4271], all the NLRI advertised by a node are withdrawn when a session failure is detected. If fast failure detection such as BFD is utilized, and the node is on the fastest converging path, the most recent versions of BGP-LS-SPF NLRI may be withdrawn. This results in an older version of the NLRI received on a different path being used until the new versions arrive and, potentially, unnecessary route flaps. For the BGP-LS-SPF SAFI, NLRI received from

the failing node SHOULD NOT be implicitly withdrawn immediately to prevent such unnecessary route flaps. The configurable NLRIImplicitWithdrawalDelay timer controls the interval that NLRI from the failed node is retained prior to implicit withdrawal after a BGP SPF speaker has transitioned out of Established state. This does not delay convergence since the adjacent nodes detect the link failure and advertise a more recent NLRI indicating the link is down with respect to BGP SPF (refer to Section 6.5.1) and the bi-directional connectivity check fails during the BGP SPF calculation (refer to Section 6.3). The suggested default value for the NLRIImplicitWithdrawalDelay timer is 2 seconds.

7. Error Handling

This section describes the Error Handling actions, as described in [RFC7606], that are specific to SAFI BGP-LS-SPF BGP Update message processing.

7.1. Processing of BGP-LS-SPF TLVs

When a BGP SPF speaker receives a BGP Update containing a malformed Node NLRI SPF Status TLV in the BGP-LS Attribute [I-D.ietf-idr-rfc7752bis], the corresponding Node NLRI is considered as malformed and MUST be handled as 'Treat-as-withdraw'. An implementation SHOULD log an error (subject to rate-limiting) for further analysis.

When a BGP SPF speaker receives a BGP Update containing a malformed Link NLRI SPF Status TLV in the BGP-LS Attribute [I-D.ietf-idr-rfc7752bis], the corresponding Link NLRI is considered as malformed and MUST be handled as 'Treat-as-withdraw'. An implementation SHOULD log an error (subject to rate-limiting) for further analysis.

When a BGP SPF speaker receives a BGP Update containing a malformed Prefix NLRI SPF Status TLV in the BGP-LS Attribute [I-D.ietf-idr-rfc7752bis], the corresponding Prefix NLRI is considered as malformed and MUST be handled as 'Treat-as-withdraw'. An implementation SHOULD log an error (subject to rate-limiting) for further analysis.

When a BGP SPF speaker receives a BGP Update containing any malformed BGP-LS Attribute TE and IGP Metric TLV, the corresponding NLRI is considered as malformed and MUST be handled as 'Treat-as-withdraw' [RFC7606]. An implementation SHOULD log an error (subject to rate-limiting) for further analysis.

The BGP-LS Attribute consists of Node attribute TLVs, Link attribute TLVs, and the Prefix attribute TLVs. Node attribute TLVs and their error handling rules are either defined in [I-D.ietf-idr-rfc7752bis] or derived from [RFC5305] and [RFC6119]. If a BGP SPF speaker receives a BGP-LS Attribute which is considered malformed based on these error handling rules, then it MUST consider the received NLRI as malformed and the receiving BGP SPF speaker MUST handle such malformed NLRI as 'Treat-as-withdraw' [RFC7606].

Node Descriptor TLVs and their error handling rules are either defined in section 5.2.1 of [I-D.ietf-idr-rfc7752bis]. Node Attribute TLVs and their error handling rules are either defined in [I-D.ietf-idr-rfc7752bis] or derived from [RFC5305] and [RFC6119].

Link Descriptor TLVs and their error handling rules are either defined in section 5.2.2 of [I-D.ietf-idr-rfc7752bis]. Link Attribute TLVs and their error handling rules are either defined in [I-D.ietf-idr-rfc7752bis] or derived from [RFC5305] and [RFC6119].

Prefix Descriptor TLVs and their error handling rules are either defined in section 5.2.3 of [I-D.ietf-idr-rfc7752bis]. Prefix Attribute TLVs and their error handling rules are either defined in [I-D.ietf-idr-rfc7752bis] or derived from [RFC5130] and [RFC2328].

If a BGP SPF speaker receives NLRI with a Node Descriptor TLV, Link Descriptor TLV, or Prefix Descriptor TLV that is considered malformed based on error handling rules defined in the above references, then it MUST consider the received NLRI as malformed and the receiving BGP SPF speaker MUST handle such malformed NLRI as 'Treat-as-withdraw' [RFC7606].

When a BGP SPF speaker receives a BGP Update that does not contain any BGP-LS Attribute, then a BGP SPF speaker MUST consider the corresponding NLRI as malformed and MUST handle it as 'Treat-as-withdraw' [RFC7606]. An implementation SHOULD log an error (subject to rate-limiting) for further analysis.

7.2. Processing of BGP-LS-SPF NLRIs

A BGP-LS-SPF Speaker MUST perform the syntactic validation checks of the BGP-LS-SPF NLRI listed in Section 8.2.2 of [I-D.ietf-idr-rfc7752bis] to determine if it is malformed.

In common deployment scenarios, the unicast routes installed during BGP-LS-SPF AFI/SAFI SPF computation serve as the underlay for other BGP AFI/SAFIs. To avoid errors encountered in other AFI/SAFIs from impacting the BGP-LS-SPF AFI/SAFI or vice-versa, isolation mechanisms such as separate BGP instances or separate BGP sessions (e.g., using different addresses for peering) for BGP SPF Link-State information distribution SHOULD be used.

7.3. Processing of BGP-LS Attribute

A BGP-LS-SPF Speaker MUST perform the syntactic validation checks of the BGP-LS Attribute listed in Section 8.2.2 of [I-D.ietf-idr-rfc7752bis] to determine if it is malformed.

An implementation SHOULD log an error for further analysis for problems detected during syntax validation.

7.4. BGP-LS-SPF Link State NLRI Database Synchronization

While uncommon, there may be situations where the LSNDBs of two BGP-LS-SPF speakers lose synchronization. In these situations, the BGP session MUST be reset. The mechanisms to detect loss of synchronization are beyond the scope of this document.

8. IANA Considerations

8.1. BGP-LS-SPF Allocation in SAFI Parameters Registry

IANA has assigned value 80 for BGP-LS-SPF from the First Come First Served range in the "Subsequent Address Family Identifiers (SAFI) Parameters" registry. IANA is requested to update the registration to reference only to this document.

8.2. BGP-LS Address Family Link Descriptor

IANA is requested to assign the value TBD (266 suggested) to the BGP-LS BGP-LS TLVs Registry for the BGP-LS Address Family Link Descriptor TLV (refer to section Section 5.2.2.1.

8.3. BGP-LS-SPF Assignments to BGP-LS NLRI and Attribute TLV Registry

IANA has assigned four TLVs for BGP-LS-SPF NLRI in the "BGP-LS NLRI and Attribute TLV" registry. These TLV types include the SPF Status TLV and Sequence Number TLV.

TLV Code Point	Description	Reference
1184	SPF Status	Section 5.2.1.1, RFCXXXX ([this document]), Section 5.2.2.2 and Section 5.2.3.1
1181	Sequence Number	RFCXXXX ([this document]), Section 5.2.4

Table 1: NLRI Attribute TLVs

8.4. BGP-LS-SPF Node NLRI Attribute SPF Status TLV Status Registry

IANA is requested to create the "BGP-LS-SPF Node NLRI Attribute SPF Status TLV Status" Registry for status values in a new BGP SPF group. Initial values for this registry are provided below. Future assignments are to be made using the IETF Review registration policy [RFC8126].

Values	Description
0	Reserved
1	Node unreachable with respect to BGP SPF
2	Node does not support transit traffic with respect to BGP SPF
3-254	Unassigned
255	Reserved

Table 2: BGP-LS-SPF Node NLRI Attribute SPF Status TLV Status Registry Assignments

8.5. BGP-LS-SPF Link NLRI Attribute SPF Status TLV Status Registry

IANA is requested to create the "BGP-LS-SPF Link NLRI Attribute SPF Status TLV Status" Registry for status values in a new BGP SPF group. Initial values for this registry are provided below. Future assignments are to be made using the IETF Review registration policy [RFC8126].

Value	Description
0	Reserved
1	Link unreachable with respect to BGP SPF
3-254	Unassigned
255	Reserved

Table 3: BGP-LS-SPF Link NLRI Attribute SPF Status TLV Status Registry Assignments

8.6. BGP-LS-SPF Prefix NLRI Attribute SPF Status TLV Status Registry

IANA is requested to create the "BGP-LS-SPF Prefix NLRI Attribute SPF Status TLV Status" Registry for status values in a new BGP SPF group. Initial values for this registry are provided below. Future assignments are to be made using the IETF Review registration policy [RFC8126].

Value	Description
0	Reserved
1	Prefix unreachable with respect to BGP SPF
3-254	Unassigned
255	Reserved

Table 4: BGP-LS-SPF Prefix NLRI Attribute SPF Status TLV Status Registry Assignments

9. Security Considerations

This document defines a BGP SAFI, i.e., the BGP-LS-SPF SAFI. This document does not change the underlying security issues inherent in the BGP protocol [RFC4271]. The Security Considerations discussed in [RFC4271] apply to the BGP SPF functionality as well. The analysis of the security issues for BGP mentioned in [RFC4272] and [RFC6952] also applies to this document. The analysis of Generic Threats to Routing Protocols done in [RFC4593] is also worth noting.

As the modifications described in this document for BGP SPF apply to IPv4 Unicast and IPv6 Unicast as underlay SAFIs in a single BGP SPF Routing Domain, the BGP security solutions described in [RFC6811] and [RFC8205] are out of scope as they are meant to apply for inter-domain BGP where multiple BGP Routing Domains are typically involved. The BGP-LS-SPF SAFI NLRI described in this document are typically advertised between EBGP or IBGP speakers under a single administrative domain.

The BGP SPF protocol and the BGP-LS-SPF SAFI inherit the encoding from BGP-LS [I-D.ietf-idr-rfc7752bis], and consequently, inherit the security considerations for BGP-LS associated with encoding. Additionally, given that the BGP SPF protocol is used to install IPv4 and IPv6 Unicast routes, the BGP SPF protocol is vulnerable to attacks to the routing control plane that aren't applicable to BGP-LS. One notable Denial-of-Service attack, would be to include malformed BGP attributes in a replicated BGP Update, causing the receiving peer to treat the advertised BGP-LS-SPF to a withdrawal [RFC7606].

In order to mitigate the risk of peering with BGP speakers masquerading as legitimate authorized BGP speakers, it is recommended that the TCP Authentication Option (TCP-AO) [RFC5925] be used to authenticate BGP sessions. If an authorized BGP peer is compromised, that BGP peer could advertise modified Node, Link, or Prefix NLRI which result in misrouting, repeating origination of NLRI, and/or excessive SPF calculations. When a BGP speaker detects that its self-originated NLRI is being originated by another BGP speaker, an appropriate error should be logged so that the operator can take corrective action. This exposure is similar to other BGP AFI/SAFIs.

10. Management Considerations

This section includes unique management considerations for the BGP-LS-SPF address family.

10.1. Configuration

All routers in BGP SPF Routing Domain are under a single administrative domain allowing for consistent configuration.

10.2. Link Metric Configuration

For loopback prefixes, it is RECOMMMENDED that the metric be 0. For non-loopback prefixes, the setting of the metric is a local matter and beyond the scope of this document.

Algorithms such as setting the metric inversely to the link speed as supported in some IGP implementations MAY be supported. However, the details of how the metric is computed are beyond the scope of this document.

Within a BGP SPF Routing Domain, the IGP metrics for all advertised links SHOULD be configured or defaulted consistently. For example, if a default metric is used for one router's links, then a similar metric should be used for all router's links. Similarly, if the link metric is derived from using the inverse of the link bandwidth on one router, then this SHOULD be done for all routers and the same reference bandwidth SHOULD be used to derive the inversely proportional metric. Failure to do so will result in incorrect routing based on link metric.

10.3. Unnumbered Link Configuration

When parallel unnumbered links between BGP-SPF routers are included in the BGP SPF routing domain and the Remote Link Identifiers aren't readily discovered, it is RECOMMENDED that these the Remote Link Identifiers be configured so that precise NLRI Link matching can be done.

10.4. Adjacency End-of-RIB (EOR) Marker Requirement

Depending on the peering model, topology, and convergence requirements, an End-of-RIB (EoR) Marker [RFC4724] for the BGP-LS-SPF SAFI MAY be required from the peer prior to advertising a BGP-LS Link NLRI for the peer. If configuration is supported, this SHOULD be configurable at the BGP SPF instance level and SHOULD be configured consistently throughout the BGP SPF routing domain.

10.5. backoff-config

In addition to configuration of the BGP-LS-SPF address family, implementations SHOULD support the "Shortest Path First (SPF) Back-Off Delay Algorithm for Link-State IGPs" [RFC8405]. If supported, configuration of the INITIAL_SPF_DELAY, SHORT_SPF_DELAY, LONG_SPF_DELAY, TIME_TO_LEARN, and HOLDDOWN_INTERVAL MUST be supported [RFC8405]. Section 6 of [RFC8405] recommends consistent configuration of these values throughout the IGP routing domain and this also applies to the BGP SPF Routing Domain.

10.6. Operational Data

In order to troubleshoot SPF issues, implementations SHOULD support an SPF log including entries for previous SPF computations. Each SPF log entry would include the BGP-LS-SPF NLRI SPF triggering the SPF, SPF scheduled time, SPF start time, SPF end time, and SPF type if different types of SPF are supported. Since the size of the log is finite, implementations SHOULD also maintain counters for the total number of SPF computations and the total number of SPF triggering events. Additionally, to troubleshoot SPF scheduling and back-off [RFC8405], the current SPF back-off state, remaining time-to-learn, remaining hold-down interval, last trigger event time, last SPF time, and next SPF time should be available.

11. Implementation Status

Note RFC Editor: Please remove this section and the associated references prior to publication.

This section records the status of known implementations of the protocol defined by this specification at the time of posting of this Internet-Draft and is based on a proposal described in [RFC7942]. The description of implementations in this section is intended to assist the IETF in its decision processes in progressing drafts to RFCs. Please note that the listing of any individual implementation here does not imply endorsement by the IETF. Furthermore, no effort has been spent to verify the information presented here that was supplied by IETF contributors. This is not intended as, and must not be construed to be, a catalog of available implementations or their features. Readers are advised to note that other implementations may exist.

According to RFC 7942, "this will allow reviewers and working groups to assign due consideration to documents that have the benefit of running code, which may serve as evidence of valuable experimentation and feedback that have made the implemented protocols more mature. It is up to the individual working groups to use this information as they see fit".

The BGP-LS-SPF implementation status is documented in [I-D.psarkar-lsvr-bgp-spf-impl].

12. Acknowledgements

The authors would like to thank Sue Hares, Jorge Rabadan, Boris Hassanov, Dan Frost, Matt Anderson, Fred Baker, Lukas Krattiger, Yingzhen Qu, and Haibo Wang for their review and comments. Thanks to Pushpasis Sarkar for discussions on preventing a BGP SPF Router from being used for non-local traffic (i.e., transit traffic).

The authors extend special thanks to Eric Rosen for fruitful discussions on BGP-LS-SPF convergence as compared to IGPs.

13. Contributors

In addition to the authors listed on the front page, the following co-authors have contributed to the document.

Derek Yeung
Arrcus, Inc.
derek@arrcus.com

Gunter Van De Velde
Nokia
gunter.van_de_velde@nokia.com

Abhay Roy
Arrcus, Inc.
abhay@arrcus.com

Venu Venugopal
Cisco Systems
venuv@cisco.com

Chaitanya Yadlapalli
AT&T
cy098d@att.com

14. References

14.1. Normative References

[I-D.ietf-idr-rfc7752bis]
Talaulikar, K., "Distribution of Link-State and Traffic Engineering Information Using BGP", Work in Progress, Internet-Draft, draft-ietf-idr-rfc7752bis-17, 25 August 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-rfc7752bis-17>.

[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>.

[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>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC5130] Previdi, S., Shand, M., Ed., and C. Martin, "A Policy Control Mechanism in IS-IS Using Administrative Tags", RFC 5130, DOI 10.17487/RFC5130, February 2008, <https://www.rfc-editor.org/info/rfc5130>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <https://www.rfc-editor.org/info/rfc5880>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <https://www.rfc-editor.org/info/rfc5925>.

[RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, DOI 10.17487/RFC6119, February 2011, <https://www.rfc-editor.org/info/rfc6119>.

[RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, DOI 10.17487/RFC6793, December 2012, <https://www.rfc-editor.org/info/rfc6793>.

[RFC6811] Mohapatra, P., Scudder, J., Ward, D., Bush, R., and R. Austein, "BGP Prefix Origin Validation", RFC 6811, DOI 10.17487/RFC6811, January 2013, <https://www.rfc-editor.org/info/rfc6811>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[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>.

[RFC8205] Lepinski, M., Ed. and K. Sriram, Ed., "BGPsec Protocol Specification", RFC 8205, DOI 10.17487/RFC8205, September 2017, <https://www.rfc-editor.org/info/rfc8205>.

[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>.

[RFC8654] Bush, R., Patel, K., and D. Ward, "Extended Message Support for BGP", RFC 8654, DOI 10.17487/RFC8654, October 2019, <https://www.rfc-editor.org/info/rfc8654>.

[RFC9086] Previdi, S., Talaulikar, K., Ed., Filsfils, C., Patel, K., Ray, S., and J. Dong, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing BGP Egress Peer Engineering", RFC 9086, DOI 10.17487/RFC9086, August 2021, <https://www.rfc-editor.org/info/rfc9086>.

14.2. Informational References

[I-D.ietf-lsvr-applicability]
Patel, K., Lindem, A., Zandi, S., and G. Dawra, "Usage and Applicability of Link State Vector Routing in Data Centers", Work in Progress, Internet-Draft, draft-ietf-lsvr-applicability-10, 21 August 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-lsvr-applicability-10>.

[I-D.psarkar-lsvr-bgp-spf-impl]
Sarkar, P., Patel, K., Pallagatti, S., and sajibasil@gmail.com, "BGP Shortest Path Routing Extension Implementation Report", Work in Progress, Internet-Draft, draft-psarkar-lsvr-bgp-spf-impl-01, 6 June 2023, <https://datatracker.ietf.org/doc/html/draft-psarkar-lsvr-bgp-spf-impl-01>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <https://www.rfc-editor.org/info/rfc4272>.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[RFC4593] Barbir, A., Murphy, S., and Y. Yang, "Generic Threats to Routing Protocols", RFC 4593, DOI 10.17487/RFC4593, October 2006, <https://www.rfc-editor.org/info/rfc4593>.

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <https://www.rfc-editor.org/info/rfc4724>.

[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>.

[RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, DOI 10.17487/RFC6952, May 2013, <https://www.rfc-editor.org/info/rfc6952>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

[RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <https://www.rfc-editor.org/info/rfc7942>.

Authors' Addresses

Keyur Patel
Arrcus, Inc.
Email: keyur@arrcus.com

Acee Lindem
LabN Consulting, LLC
301 Midenhall Way
Cary, NC 27513
United States of America
Email: acee.ietf@gmail.com

Shawn Zandi
LinkedIn
222 2nd Street
San Francisco, CA 94105
United States of America
Email: szandi@linkedin.com

Wim Henderickx
Nokia
copernicuslaan 50
2018 Antwerp
Belgium
Email: wim.henderickx@nokia.com