Internet Engineering Task Force (IETF)
Request for Comments: 8277
Obsoletes: 3107
Category: Standards Track
ISSN: 2070-1721

E. Rosen
Juniper Networks, Inc.

October 2017

Using BGP to Bind MPLS Labels to Address Prefixes

Abstract

This document specifies a set of procedures for using BGP to advertise that a specified router has bound a specified MPLS label (or a specified sequence of MPLS labels organized as a contiguous part of a label stack) to a specified address prefix. This can be done by sending a BGP UPDATE message whose Network Layer Reachability Information field contains both the prefix and the MPLS label(s) and whose Next Hop field identifies the node at which said prefix is bound to said label(s). This document obsoletes RFC 3107.

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/rfc8277.

Table of Contents

1. Introduction

[RFC3107] specifies encodings and procedures for using BGP to indicate that a particular router has bound either a single MPLS label or a sequence of MPLS labels to a particular address prefix. (A sequence of labels would be organized as a contiguous part of an MPLS label stack as specified in [RFC3031] and [RFC3032].) This is done by sending a BGP UPDATE message whose Network Layer Reachability Information field contains both the prefix and the MPLS label(s) and whose Next Hop field identifies the node at which said prefix is bound to said label(s). Each such UPDATE also advertises a path to the specified prefix via the specified next hop.

Although there are many implementations and deployments of [RFC3107], there are a number of issues with it that have impeded interoperability in the past and may potentially impede interoperability in the future:

- Although [RFC3107] specifies an encoding that allows a sequence of MPLS labels (rather than just a single label) to be bound to a prefix, it does not specify the semantics of binding a sequence of labels to a prefix.

- Many implementations of [RFC3107] assume that only one label will be bound to a prefix, and cannot properly process a BGP UPDATE message that binds a sequence of labels to a prefix. Thus, an implementation attempting to provide this feature is likely to experience problems interoperating with other implementations.

- The procedures in [RFC3107] for withdrawing the binding of a label or sequence of labels to a prefix are not specified clearly and correctly.

- [RFC3107] specifies an optional feature, known as "Advertising Multiple Routes to a Destination", that, to the best of the author's knowledge, has never been implemented as specified. The functionality that this feature was intended to provide can and has been implemented in a different way using the procedures of [RFC7911], which were not available at the time that [RFC3107] was written. In [RFC3107], this feature was controlled by a BGP Capability Code that has never been implemented and is now deprecated; see Section 6.

- It is possible for a BGP speaker to receive two BGP UPDATEs that advertise paths to the same address prefix, where one UPDATE binds a label (or sequence of labels) to the prefix and the other does not. [RFC3107] is silent on the issue of how the presence of two such UPDATEs impacts the BGP decision process and does not say explicitly whether one or the other or both of these UPDATEs should be propagated. This has led different implementations to handle this situation in different ways.

- Much of [RFC3107] applies to the VPN-IPv4 ([RFC4364]) and VPN-IPv6 ([RFC4659]) address families, but those address families are not mentioned in it.

This document replaces and obsoletes [RFC3107]. It defines a new BGP Capability to be used when binding a sequence of labels to a prefix; by using this Capability, the interoperability problems alluded to above can be avoided. This document also removes the unimplemented

"Advertising Multiple Routes to a Destination" feature (see Section 4 of [RFC3107]), while specifying how to use [RFC7911] to provide the same functionality. This document also addresses the issue of the how UPDATEs that bind labels to a given prefix interact with UPDATEs that advertise paths to that prefix but do not bind labels to it. However, for backwards compatibility, it declares most of these interactions to be matters of local policy.

The places where this specification differs from [RFC3107] are indicated in the text. It is believed that implementations that conform to the current document will interoperate correctly with existing deployed implementations of [RFC3107].

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. Using BGP to Bind an Address Prefix to One or More MPLS Labels

BGP may be used to advertise that a particular node (call it N) has bound a particular MPLS label, or a particular sequence of MPLS labels (organized as a contiguous part of an MPLS label stack), to a particular address prefix. This is done by sending a Multiprotocol BGP UPDATE message, i.e., an UPDATE message with an MP_REACH_NLRI attribute as specified in [RFC4760]. The Network Address of Next Hop field of that attribute contains an IP address of node N. The label(s) and the prefix are encoded in the Network Layer Reachability Information (NLRI) field of the MP_REACH_NLRI. The encoding of the NLRI field is specified in Sections 2.2 and 2.3.

If the prefix is an IPv4 address prefix or a VPN-IPv4 ([RFC4364]) address prefix, the Address Family Identifier (AFI) of the MP_REACH_NLRI attribute is set to 1. If the prefix is an IPv6 address prefix or a VPN-IPv6 prefix ([RFC4659]), the AFI is set to 2.

If the prefix is an IPv4 address prefix or an IPv6 address prefix, the Subsequent Address Family Identifier (SAFI) field is set to 4. If the prefix is a VPN-IPv4 address prefix or a VPN-IPv6 address prefix, the SAFI is set to 128.

The use of SAFI 4 or SAFI 128 when the AFI is other than 1 or 2 is outside the scope of this document.

This document does not specify the format of the Network Address of Next Hop field of the MP_REACH_NLRI attribute. The format of the Next Hop field depends upon a number of factors and is discussed in a number of other RFCs: see [RFC4364], [RFC4659], [RFC4798], and [RFC5549].

There are a variety of applications that make use of alternative methods of using BGP to advertise MPLS label bindings: see, e.g., [RFC7432], [RFC6514], or [TUNNEL-ENCAPS]. The method described in the current document is not claimed to be the only way of using BGP to advertise MPLS label bindings. Discussion of which method to use for which application is outside the scope of the current document.

In the remainder of this document, we will use the term "SAFI-x UPDATE" to refer to a BGP UPDATE message containing an MP_REACH_NLRI attribute or an MP_UNREACH_NLRI attribute ([RFC4760]) whose SAFI field contains the value x.

This document defines a BGP Optional Capabilities parameter ([RFC5492]) known as the Multiple Labels Capability.

- Unless this Capability is sent on a given BGP session by both of that session's BGP speakers, a SAFI-4 or SAFI-128 UPDATE message sent on that session from either speaker MUST bind a prefix to only a single label and MUST use the encoding of Section 2.2.

- If this Capability is sent by both BGP speakers on a given session, an UPDATE message on that session, from either speaker, MUST use the encoding of Section 2.3 and MAY bind a prefix to a sequence of more than one label.

The encoding of the Multiple Labels Capability is specified in Section 2.1.

Procedures for explicitly withdrawing a label binding are given in Section 2.4. Procedures for changing the label(s) bound to a given prefix by a given node are given in Section 2.5.

Procedures for propagating SAFI-4 and SAFI-128 UPDATEs are discussed in Section 3.

When a BGP speaker installs and propagates a SAFI-4 or SAFI-128 UPDATE, and if it changes the value of the Network Address of Next Hop field, it must program its data plane appropriately. This is discussed in Section 4.

2.1. Multiple Labels Capability

[RFC5492] defines the "Capabilities Optional Parameter". A BGP speaker can include a Capabilities Optional Parameter in a BGP OPEN message. The Capabilities Optional Parameter is a triple that includes a one-octet Capability Code, a one-octet Capability length, and a variable-length Capability Value.

This document defines a Capability Code known as the Multiple Labels Capability code. IANA has assigned value 8 to this Capability Code. (This Capability Code is new to this document and does not appear in [RFC3107].)

If a BGP speaker has not sent the Multiple Labels Capability in its BGP OPEN message on a particular BGP session, or if it has not received the Multiple Labels Capability in the BGP OPEN message from its peer on that BGP session, that BGP speaker MUST NOT send on that session any UPDATE message that binds more than one MPLS label to any given prefix. Further, when advertising the binding of a single label to a prefix, the BGP speaker MUST use the encoding specified in Section 2.2.

The value field of the Multiple Labels Capability (shown in Figure 1) consists of one or more triples, where each triple consists of four octets. The first two octets of a triple specify an AFI value, the third octet specifies a SAFI value, and the fourth specifies a Count. If one of the triples is <AFI, SAFI, Count>, the Count is the maximum number of labels that the BGP speaker sending the Capability can process in a received UPDATE of the specified AFI/SAFI. If the Count is 255, then no limit has been placed on the number of labels that can be processed in a received UPDATE of the specified AFI/SAFI.

Any implementation that sends a Multiple Labels Capability MUST be able to support at least two labels in the NLRI. However, there may be deployment scenarios in which a larger number of labels is needed.

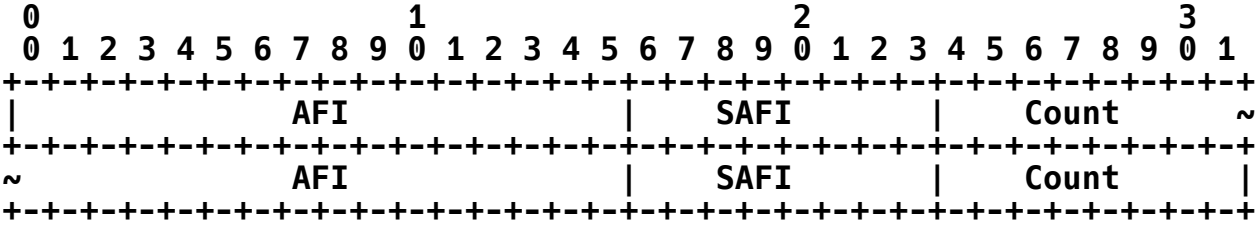

Figure 1: Value Field of Multiple Labels Capability

If the Capability contains more than one triple with a given AFI/SAFI, all but the first MUST be ignored.

A triple of the form <AFI=x, SAFI=y, Count=0> or <AFI=x, SAFI=y, Count=1> MUST NOT be sent. If such a triple is received, it MUST be ignored.

A Multiple Labels Capability whose length is not a multiple of four MUST be considered to be malformed.

"Graceful Restart Mechanism for BGP" [RFC4724] describes a procedure that allows routes learned over a given BGP session to be maintained when the session fails and then restarts. This procedure requires the entire RIB to be transmitted when the session restarts. If the Multiple Labels Capability for a given AFI/SAFI was exchanged on the failed session but has not been exchanged on the restarted session, then any prefixes advertised in that AFI/SAFI with multiple labels MUST be explicitly withdrawn. Similarly, if the maximum label count (specified in the Capability for a given AFI/SAFI) is reduced, any prefixes advertised with more labels than are valid for the current session MUST be explicitly withdrawn.

"Accelerated Routing Convergence for BGP Graceful Restart" [Enhanced-GR] describes another procedure that allows the routes learned over a given BGP session to be maintained when the session fails and then restarts. These procedures MUST NOT be applied if either of the following conditions hold:

- The Multiple Labels Capability for a given AFI/SAFI had been exchanged prior to the restart but has not been exchanged on the restarted session.

- The Multiple Labels Capability for a given AFI/SAFI had been exchanged with a given Count prior to the restart but have been exchanged with a smaller count on the restarted session.

If either of these conditions hold, the complete set of routes for the given AFI/SAFI MUST be exchanged.

If a BGP OPEN message contains multiple copies of the Multiple Labels Capability, only the first copy is significant; subsequent copies MUST be ignored.

If (a) a BGP speaker has sent the Multiple Labels Capability in its BGP OPEN message for a particular BGP session, (b) it has received the Multiple Labels Capability in its peer's BGP OPEN message for that session, and (c) both Capabilities specify AFI/SAFI x/y, then when using an UPDATE of AFI x and SAFI y to advertise the binding of a label or sequence of labels to a given prefix, the BGP speaker MUST use the encoding of Section 2.3. This encoding MUST be used even if only one label is being bound to a given prefix.

If both BGP speakers of a given BGP session have sent the Multiple Labels Capability, but AFI/SAFI x/y has not been specified in both Capabilities, then UPDATEs of AFI/SAFI x/y on that session MUST use the encoding of Section 2.2, and such UPDATEs can only bind one label to a prefix.

A BGP speaker SHOULD NOT send an UPDATE that binds more labels to a given prefix than its peer is capable of receiving, as specified in the Multiple Labels Capability sent by that peer. If a BGP speaker receives an UPDATE that binds more labels to a given prefix than the number of labels the BGP speaker is prepared to receive (as announced in its Multiple Labels Capability), the BGP speaker MUST apply the "treat-as-withdraw" strategy of [RFC7606] to that UPDATE.

Notwithstanding the number of labels that a BGP speaker has claimed to be able to receive, its peer MUST NOT attempt to send more labels than can be properly encoded in the NLRI field of the MP_REACH_NLRI attribute. Please note that there is only a limited amount of space in the NLRI field for labels:

- per [RFC4760], the size of this field is limited to 255 bits (not 255 octets), including the number of bits in the prefix;
- in a SAFI-128 UPDATE, the prefix is at least 64 bits long and may be as long as 192 bits (e.g., in a VPN-IPv6 host route).

2.2. NLRI Encoding When the Multiple Labels Capability Is Not Used

If the Multiple Labels Capability has not been both sent and received on a given BGP session, then in a BGP UPDATE on that session whose MP_REACH_NLRI attribute contains one of the AFI/SAFI combinations specified in Section 2, the NLRI field is encoded as shown in Figure 2:

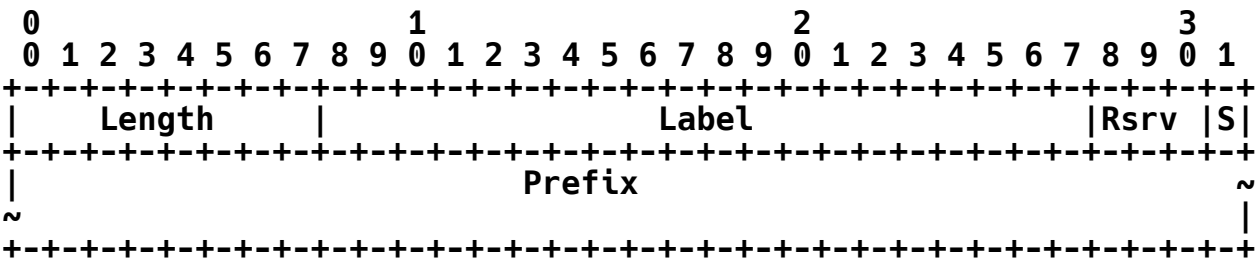

Figure 2: NLRI with One Label

- Length:

  The Length field consists of a single octet. It specifies the length in bits of the remainder of the NLRI field.

  Note that the length will always be the sum of 20 (number of bits in Label field), plus 3 (number of bits in Rsrv field), plus 1 (number of bits in S field), plus the length in bits of the prefix.

  In an MP_REACH_NLRI attribute whose AFI/SAFI is 1/4, the prefix length will be 32 bits or less. In an MP_REACH_NLRI attribute whose AFI/SAFI is 2/4, the prefix length will be 128 bits or less. In an MP_REACH_NLRI attribute whose SAFI is 128, the prefix will be 96 bits or less if the AFI is 1 and will be 192 bits or less if the AFI is 2.

  As specified in [RFC4760], the actual length of the NLRI field will be the number of bits specified in the Length field, rounded up to the nearest integral number of octets.

- Label:

  The Label field is a 20-bit field containing an MPLS label value (see [RFC3032]).

- Rsrv:

  This 3-bit field SHOULD be set to zero on transmission and MUST be ignored on reception.

- S:

  This 1-bit field MUST be set to one on transmission and MUST be ignored on reception.

Note that the UPDATE message not only advertises the binding between the prefix and the label, it also advertises a path to the prefix via the node identified in the Network Address of Next Hop field of the MP_REACH_NLRI attribute.

[RFC3107] requires that if only a single label is bound to a prefix, the S bit must be set. If the S bit is not set, [RFC3107] specifies that additional labels will appear in the NLRI. However, some implementations assume that the NLRI will contain only a single label and thus do not check the setting of the S bit. The procedures specified in the current document will interwork with such implementations. As long as the Multiple Labels Capability is not

sent and received by both BGP speakers on a given BGP session, this document REQUIRES that only one label be specified in the NLRI, that the S bit be set on transmission, and that it be ignored on reception.

If the procedures of [RFC7911] are being used, a four-octet "path identifier" (as defined in Section 3 of [RFC7911]) is part of the NLRI and precedes the Length field.

2.3. NLRI Encoding When the Multiple Labels Capability Is Used

If the Multiple Labels Capability has been both sent and received on a given BGP session, then in a BGP UPDATE on that session whose MP_REACH_NLRI attribute contains one of the AFI/SAFI combinations specified in Section 2, the NLRI field is encoded as shown in Figure 3:

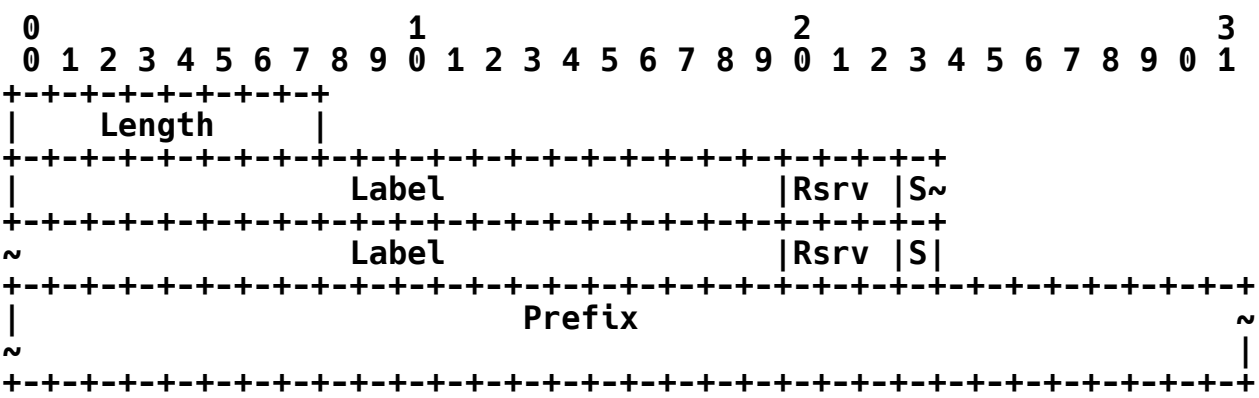

Figure 3: NLRI with Multiple Labels

- Length:

  The Length field consists of a single octet. It specifies the length in bits of the remainder of the NLRI field.

  Note that for each label, the length is increased by 24 bits (20 bits in the Label field, plus 3 bits in the Rsrv field, plus 1 S bit).

  In an MP_REACH_NLRI attribute whose AFI/SAFI is 1/4, the prefix length will be 32 bits or less. In an MP_REACH_NLRI attribute whose AFI/SAFI is 2/4, the prefix length will be 128 bits or less. In an MP_REACH_NLRI attribute whose SAFI is 128, the prefix will be 96 bits or less if the AFI is 1 and will be 192 bits or less if the AFI is 2.

As specified in [RFC4760], the actual length of the NLRI field will be the number of bits specified in the Length field rounded up to the nearest integral number of octets.

- Label:

  The Label field is a 20-bit field containing an MPLS label value (see [RFC3032]).

- Rsrv:

  This 3-bit field SHOULD be set to zero on transmission and MUST be ignored on reception.

- S:

  In all labels except the last (i.e., in all labels except the one immediately preceding the prefix), the S bit MUST be 0. In the last label, the S bit MUST be 1.

  Note that failure to set the S bit in the last label will make it impossible to parse the NLRI correctly. See Section 3, paragraph j of [RFC7606] for a discussion of error handling when the NLRI cannot be parsed.

Note that the UPDATE message not only advertises the binding between the prefix and the labels, it also advertises a path to the prefix via the node identified in the Next Hop field of the MP_REACH_NLRI attribute.

If the procedures of [RFC7911] are being used, a four-octet "path identifier" (as defined in Section 3 of [RFC7911]) is part of the NLRI and precedes the Length field.

2.4. How to Explicitly Withdraw the Binding of a Label to a Prefix

Suppose a BGP speaker has announced, on a given BGP session, the binding of a given label or sequence of labels to a given prefix. Suppose it now wishes to withdraw that binding. To do so, it may send a BGP UPDATE message with an MP_UNREACH_NLRI attribute. The NLRI field of this attribute is encoded as follows:

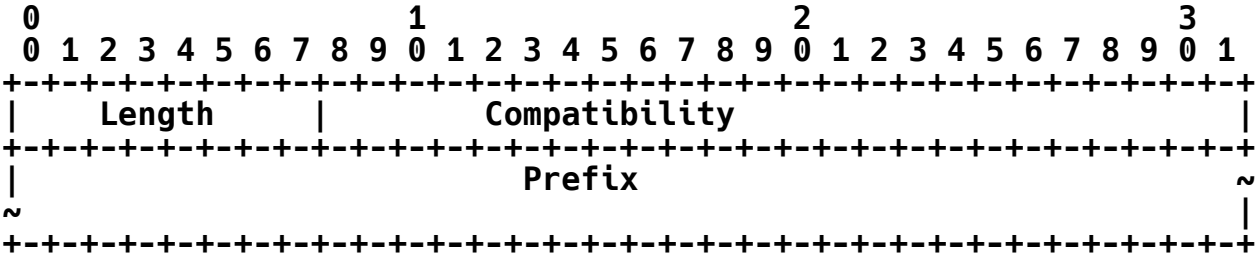

Figure 4: NLRI for Withdrawal

Upon transmission, the Compatibility field SHOULD be set to 0x800000. Upon reception, the value of the Compatibility field MUST be ignored.

This encoding is used for explicitly withdrawing the binding (on a given BGP session) between the specified prefix and whatever label or sequence of labels had previously been bound by the procedures of this document to that prefix on the given session. This encoding is used whether or not the Multiple Labels Capability has been sent or received on the session. Note that label/prefix bindings that were not advertised on the given session cannot be withdrawn by this method. (However, if the bindings were advertised on a previous session with the same peer, and the current session is the result of a "graceful restart" ([RFC4724]) of the previous session, then this withdrawal method may be used.)

When using an MP_UNREACH_NLRI attribute to withdraw a route whose NLRI was previously specified in an MP_REACH_NLRI attribute, the lengths and values of the respective prefixes must match, and the respective AFI/SAFIs must match. If the procedures of [RFC7911] are being used, the respective values of the "path identifier" fields must match as well. Note that the prefix length is not the same as the NLRI length; to determine the prefix length of a prefix in an MP_UNREACH_NLRI, the length of the Compatibility field must be subtracted from the length of the NLRI.

An explicit withdrawal in a SAFI-x UPDATE on a given BGP session not only withdraws the binding between the prefix and the label(s), it also withdraws the path to that prefix that was previously advertised in a SAFI-x UPDATE on that session.

[RFC3107] made it possible to specify a particular label value in the Compatibility field. However, the functionality that required the presence of a particular label value (or sequence of label values) was never implemented, and that functionality is not present in the current document. Hence, the value of this field is of no significance; there is never any reason for this field to contain a label value or a sequence of label values.

[RFC3107] also made it possible to withdraw a binding without specifying the label explicitly, by setting the Compatibility field to 0x800000. However, some implementations set it to 0x000000. In order to ensure backwards compatibility, it is RECOMMENDED by this document that the Compatibility field be set to 0x800000, but it is REQUIRED that it be ignored upon reception.

2.5. Changing the Label That Is Bound to a Prefix

Suppose a BGP speaker, S1, has received on a given BGP session, a SAFI-4 or SAFI-128 UPDATE, U1, that specifies label (or sequence of labels) L1, prefix P, and next hop N1. As specified above, this indicates that label (or sequence of labels) L1 is bound to prefix P at node N1. Suppose that S1 now receives, on the same session, an UPDATE, U2, of the same AFI/SAFI, that specifies label (or sequence of labels) L2, prefix P, and the same next hop, N1.

- If [RFC7911] is not being used, UPDATE U2 MUST be interpreted as meaning that L2 is now bound to P at N1 and that L1 is no longer bound to P at N1. That is, the UPDATE U1 is implicitly withdrawn and is replaced by UPDATE U2.

- Suppose that [RFC7911] is being used, that UPDATE U1 has Path Identifier I1, and that UPDATE U2 has Path Identifier I2.

  * If I1 is the same as I2, UPDATE U2 MUST be interpreted as meaning that L2 is now bound to P at N1 and that L1 is no longer bound to P at N1. UPDATE U1 is implicitly withdrawn.

  * If I1 is not the same as I2, U2 MUST be interpreted as meaning that L2 is now bound to P at N1, but U2 MUST NOT be interpreted as meaning that L1 is no longer bound to P at N1. Under certain conditions (specification of which is outside the scope of this document), S1 may choose to load-balance traffic between the path represented by U1 and the path represented by U2. To send traffic on the path represented by U1, S1 uses the label(s) advertised in U1; to send traffic on the path represented by U2, S1 uses the label(s) advertised in U2. (Although these two paths have the same next hop, one must suppose that they diverge further downstream.)

Suppose a BGP speaker, S1, has received, on a given BGP session, a SAFI-4 or SAFI-128 UPDATE that specifies label L1, prefix P, and next hop N1. Suppose that S1 now receives, on a different BGP session, an UPDATE of the same AFI/SAFI, that specifies label L2, prefix P, and the same next hop, N1. BGP speaker S1 SHOULD treat this as an indication that N1 has at least two paths to P, and S1 MAY use this fact to do load-balancing of any traffic that it has to send to P.

Note that this section discusses only the case where two UPDATEs have the same next hop. Procedures for the case where two UPDATEs have different next hops are adequately described in [RFC4271].

3. Installing and/or Propagating SAFI-4 or SAFI-128 Routes

3.1. Comparability of Routes

Suppose a BGP speaker has received two SAFI-4 UPDATEs specifying the same Prefix and that either:

- the two UPDATEs are received on different BGP sessions; or
- the two UPDATEs are received on the same session, add-paths is used on that session, and the NLRIs of the two UPDATEs have different path identifiers.

These two routes MUST be considered to be comparable, even if they specify different labels. Thus, the BGP best-path selection procedures (see Section 9.1 of [RFC4271]) are applied to select one of them as the better path. If the procedures of [RFC7911] are not being used on a particular BGP session, only the best path is propagated on that session. If the procedures of [RFC7911] are being used on a particular BGP session, then both paths may be propagated on that session, though with different path identifiers.

The same applies to SAFI-128 routes.

3.2. Modification of Label(s) Field When Propagating

3.2.1. When the Next Hop Field Is Unchanged

When a SAFI-4 or SAFI-128 route is propagated, if the Network Address of Next Hop field is left unchanged, the Label field(s) MUST also be left unchanged.

Note that a given route MUST NOT be propagated to a given peer if the route's NLRI has multiple labels, but the Multiple Labels Capability was not negotiated with the peer. Similarly, a given route MUST NOT be propagated to a given peer if the route's NLRI has more labels

than the peer has announced (through its Multiple Labels Capability) that it can handle. In either case, if a previous route with the same AFI, SAFI, and prefix (but with fewer labels) has already been propagated to the peer, that route MUST be withdrawn from that peer using the procedure specified in Section 2.4.

3.2.2. When the Next Hop Field Is Changed

If the Network Address of Next Hop field is changed before a SAFI-4 or SAFI-128 route is propagated, the Label field(s) of the propagated route MUST contain the label(s) that is (are) bound to the prefix at the new next hop.

Suppose BGP speaker S1 has received an UPDATE that binds a particular sequence of one or more labels to a particular prefix. If S1 chooses to propagate this route after changing its next hop, S1 may change the label in any of the following ways, depending upon local policy:

- A single label may be replaced by a single label of the same or different value.
- A sequence of multiple labels may be replaced by a single label.
- A single label may be replaced by a sequence of multiple labels.
- A sequence of multiple labels may be replaced by a sequence of multiple labels; the number of labels may be left the same or may be changed.

Of course, when deciding whether to propagate, to a given BGP peer, an UPDATE binding a sequence of more than one label, a BGP speaker must attend to the information provided by the Multiple Labels Capability (see Section 2.1). A BGP speaker MUST NOT send multiple labels to a peer with which it has not exchanged the Multiple Labels Capability and MUST NOT send more labels to a given peer than the peer has announced (via the Multiple Labels Capability) than it can handle.

It is possible that a BGP speaker's local policy will tell it to encode N labels in a given route's NLRI before propagating the route, but that one of the BGP speaker's peers cannot handle N labels in the NLRI. In this case, the BGP speaker has two choices:

- It can propagate the route to the given peer with fewer than N labels; however, whether this makes sense, and if so, how to choose the labels, is also a matter of local policy.

- It can decide not to propagate the route to the given peer. In that case, if a previous route with the same AFI, SAFI, and prefix (but with fewer labels) has already been propagated to that peer, that route MUST be withdrawn from that peer using the procedure of Section 2.4.

4. Data Plane

In the following, we will use the phrase "node S tunnels packet P to node N", where packet P is an MPLS packet. By this phrase, we mean that node S encapsulates packet P and causes packet P to be delivered to node N in such a way that P's label stack before encapsulation will be seen unchanged by N but will not be seen by the nodes (if any) between S and N.

If the tunnel is a Label Switched Path (LSP), encapsulating the packet may be as simple as pushing on another MPLS label. If node N is a Layer 2 adjacency of node S, a Layer 2 encapsulation may be all that is needed. Other sorts of tunnels (e.g., IP tunnels, GRE tunnels, UDP tunnels) may also be used, depending upon the particular deployment scenario.

Suppose BGP speaker S1 receives a SAFI-4 or SAFI-128 BGP UPDATE with an MP_REACH_NLRI specifying label L1, prefix P, and next hop N1, and suppose S1 installs this route as its (or one of its) best path(s) towards P. And suppose S1 propagates this route after changing the next hop to itself and changing the label to L2. Suppose further that S1 receives an MPLS data packet and, in the process of forwarding that MPLS data packet, S1 sees label L2 rise to the top of the packet's label stack. Then, to forward the packet further, S1 must replace L2 with L1 as the top entry in the packet's label stack, and S1 must then tunnel the packet to N1.

Suppose that the route received by S1 specified not a single label, but a sequence of k labels <L11, L12, ..., L1k> where L11 is the first label appearing in the NLRI, and L1k is the last. And suppose again that S1 propagates this route after changing the next hop to itself and changing the Label field to the single label L2. Suppose further that S1 receives an MPLS data packet, and in the process of forwarding that MPLS data packet, S1 sees label L2 rise to the top of the packet's label stack. In this case, instead of simply replacing L2 with L1, S1 removes L2 from the top of the label stack and then pushes labels L1k through L11 onto the label stack such that L11 is now at the top of the label stack. Then, S1 must tunnel the packet to N1. (Note that L1k will not be at the bottom of the packet's label stack and hence will not have the "bottom of stack" bit set unless L2 had previously been at the bottom of the packet's label stack.)

The above paragraph assumes that when S1 propagates a SAFI-4 or SAFI-128 route after setting the next hop to itself, it replaces the label or labels specified in the NLRI of that route with a single label. However, it is also possible, as determined by local policy, for a BGP speaker to specify multiple labels when it propagates a SAFI-4 or SAFI-128 route after setting the next hop to itself.

Suppose, for example, that S1 supports context labels ([RFC5331]). Let L21 be a context label supported by S1, and let L22 be a label that is in the label space identified (at S1) by L21. Suppose S1 receives a SAFI-4 or SAFI-128 UPDATE whose prefix is P, whose Label field is <L11, L12, ..., L1k> and whose next hop is N1. Before propagating the UPDATE, S1 may set the next hop to itself (by replacing N1 with S1) and may replace the label stack <L11, L12, ..., L1k> with the pair of labels <L21, L22>.

In this case, if S1 receives an MPLS data packet whose top label is L21 and whose second label is L22, S1 will remove both L21 and L22 from the label stack and replace them with <L11, L12, ..., L1k>. Note that the fact that L21 is a context label is known only to S1; other BGP speakers do not know how S1 will interpret L21 (or L22).

The ability to replace one or more labels by one or more labels can provide great flexibility, but it must be done carefully. Let's suppose again that S1 receives an UPDATE that specifies prefix P, label stack <L11, L12, ..., L1k>, and next hop N1. And suppose that S1 propagates this UPDATE to BGP speaker S2 after setting next hop self and after replacing the Label field with <L21, L22, ..., L2k>. Finally, suppose that S1 programs its data plane so that when it processes a received MPLS packet whose top label is L21, it replaces L21 with <L11, L12, ..., L1k> and then tunnels the packet to N1.

In this case, BGP speaker S2 will have received a route with prefix P, Label field <L21, L22, ..., L2k>, and next hop S1. If S2 decides to forward an IP packet according to this route, it will push <L21, L22, ..., L2k> onto the packet's label stack and tunnel the packet to S1. S1 will replace L21 with <L11, L12, ..., L1k> and will tunnel the packet to N1. N1 will receive the packet with the following label stack: <L11, L12, ..., L1k, L22, ..., L2k>. While this may be useful in certain scenarios, it may provide unintended results in other scenarios.

Procedures for choosing, setting up, maintaining, or determining the liveness of a particular tunnel or type of tunnel are outside the scope of this document.

When pushing labels onto a packet's label stack, the Time-to-Live (TTL) field ([RFC3032], [RFC3443]) and the Traffic Class (TC) field ([RFC3032], [RFC5462]) of each label stack entry must, of course, be set. This document does not specify any set of rules for setting these fields; that is a matter of local policy.

This document does not specify any new rules for processing the label stack of an incoming data packet.

It is a matter of local policy whether SAFI-4 routes can be used as the basis for forwarding IP packets or whether SAFI-4 routes can only be used for forwarding MPLS packets. If BGP speaker S1 is forwarding IP packets according to SAFI-4 routes, then consider an IP packet with destination address D, such that P is the "longest prefix match" for D from among the routes that are being used to forward IP packets. And suppose the packet is being forwarded according to a SAFI-4 route whose prefix is P, whose next hop is N1 and whose sequence of labels is L1. To forward the packet according to this route, S1 must create a label stack for the packet, push on the sequence of labels L1, and then tunnel the packet to N1.

5. Relationship between SAFI-4 and SAFI-1 Routes

It is possible that a BGP speaker will receive both a SAFI-1 route for prefix P and a SAFI-4 route for prefix P. Different implementations treat this situation in different ways.

For example, some implementations may regard SAFI-1 routes and SAFI-4 routes as completely independent and may treat them in a "ships in the night" fashion. In this case, best-path selection for the two SAFIs is independent, and there will be a best SAFI-1 route to P as well as a best SAFI-4 route to P. Which packets get forwarded according to the routes of which SAFI is then a matter of local policy.

Other implementations may treat the SAFI-1 and SAFI-4 routes for a given prefix as comparable, such that the best route to prefix P is either a SAFI-1 route or a SAFI-4 route but not both. In such implementations, if load-balancing is done among a set of equal cost routes, some of the equal cost routes may be SAFI-1 routes and some may be SAFI-4 routes. Whether this is allowed is, again, a matter of local policy.

Some implementations may allow a single BGP session to carry UPDATEs of both SAFI-1 and SAFI-4; other implementations may disallow this. Some implementations that allow both SAFIs on the same session may treat the receipt of a SAFI-1 route for prefix P on a given session

as an implicit withdrawal of a previous SAFI-4 route for prefix P on that session, and vice versa. Other implementations may have different behavior.

A BGP speaker may receive a SAFI-4 route over a given BGP session but may have other BGP sessions for which SAFI-4 is not enabled. In this case, the BGP speaker MAY convert the SAFI-4 route to a SAFI-1 route and then propagate the result over the session on which SAFI-4 is not enabled. Whether this is done is a matter of local policy.

These differences in the behavior of different implementations may result in unexpected behavior or lack of interoperability. In some cases, it may be difficult or impossible to achieve the desired policies with certain implementations or combinations of implementations.

6. IANA Considerations

IANA has assigned value 8 for Multiple Labels Capability in the BGP "Capability Codes" registry, with this document as the reference.

IANA has modified the BGP "Capability Codes" registry to mark value 4 ("Multiple routes to a destination capability") as deprecated, with this document as the reference.

IANA has changed the reference for SAFI 4 in the "Subsequent Address Family Identifiers (SAFI) Parameters" registry to this document.

Also, IANA has added this document as a reference for SAFI 128 in that same registry.

7. Security Considerations

The security considerations of BGP (as specified in [RFC4271]) apply.

If a BGP implementation that is not conformant with the current document encodes multiple labels in the NLRI but has not sent and received the Multiple Labels Capability, a BGP implementation that does conform with the current document will likely reset the BGP session.

This document specifies that certain data packets be "tunneled" from one BGP speaker to another. This requires that the packets be encapsulated while in flight. This document does not specify the encapsulation to be used. However, if a particular encapsulation is used, the security considerations of that encapsulation are applicable.

If a particular tunnel encapsulation does not provide integrity and authentication, it is possible that a data packet's label stack can be modified, through error or malfeasance, while the packet is in flight. This can result in misdelivery of the packet. It should be noted that the tunnel encapsulation (MPLS) most commonly used in deployments of this specification does not provide integrity or authentication; neither do the other tunnel encapsulations mentioned in Section 4.

There are various techniques one can use to constrain the distribution of BGP UPDATE messages. If a BGP UPDATE advertises the binding of a particular label or set of labels to a particular address prefix, such techniques can be used to control the set of BGP speakers that are intended to learn of that binding. However, if BGP sessions do not provide privacy, other routers may learn of that binding.

When a BGP speaker processes a received MPLS data packet whose top label it advertised, there is no guarantee that the label in question was put on the packet by a router that was intended to know about that label binding. If a BGP speaker is using the procedures of this document, it may be useful for that speaker to distinguish its "internal" interfaces from its "external" interfaces and to avoid advertising the same labels to BGP speakers reached on internal interfaces as to BGP speakers reached on external interfaces. Then, a data packet can be discarded if its top label was not advertised over the type of interface from which the packet was received. This reduces the likelihood of forwarding packets whose labels have been "spoofed" by untrusted sources.

8. References

8.1. 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>.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <https://www.rfc-editor.org/info/rfc3031>.

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, DOI 10.17487/RFC3032, January 2001, <https://www.rfc-editor.org/info/rfc3032>.

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, DOI 10.17487/RFC3107, May 2001, <https://www.rfc-editor.org/info/rfc3107>.

[RFC3443] Agarwal, P. and B. Akyol, "Time To Live (TTL) Processing in Multi-Protocol Label Switching (MPLS) Networks", RFC 3443, DOI 10.17487/RFC3443, January 2003, <https://www.rfc-editor.org/info/rfc3443>.

[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>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, DOI 10.17487/RFC4659, September 2006, <https://www.rfc-editor.org/info/rfc4659>.

[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>.

[RFC4798] De Clercq, J., Ooms, D., Prevost, S., and F. Le Faucheur, "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, DOI 10.17487/RFC4798, February 2007, <https://www.rfc-editor.org/info/rfc4798>.

[RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, DOI 10.17487/RFC5462, February 2009, <https://www.rfc-editor.org/info/rfc5462>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

[RFC5549] Le Faucheur, F. and E. Rosen, "Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop", RFC 5549, DOI 10.17487/RFC5549, May 2009, <https://www.rfc-editor.org/info/rfc5549>.

[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>.

[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>.

8.2. Informative References

[Enhanced-GR] Patel, K., Chen, E., Fernando, R., and J. Scudder, "Accelerated Routing Convergence for BGP Graceful Restart", Work in Progress, draft-ietf-idr-enhanced-gr-06, June 2016.

[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>.

[RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, DOI 10.17487/RFC5331, August 2008, <https://www.rfc-editor.org/info/rfc5331>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[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>.

[TUNNEL-ENCAPS] Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", Work in Progress, draft-ietf-idr-tunnel-encaps-07, July 2017.

Acknowledgements

This document obsoletes RFC 3107. We wish to thank Yakov Rekhter, co-author of RFC 3107, for his work on that document. We also wish to thank Ravi Chandra, Enke Chen, Srihari R. Sangli, Eric Gray, and Liam Casey for their review of and comments on that document.

We thank Alexander Okonnikov and David Lamparter for pointing out a number of the errors in RFC 3107.

We wish to thank Lili Wang and Kaliraj Vairavakkalai for their help and advice during the preparation of this document.

We also thank Mach Chen, Bruno Decraene, Jie Dong, Adrian Farrel, Jeff Haas, Jonathan Hardwick, Jakob Heitz, Alexander Okonnikov, Keyur Patel, Kevin Wang, and Lucy Yong for their review of and comments on this document.

Author's Address

Eric C. Rosen
Juniper Networks, Inc.
10 Technology Park Drive
Westford, Massachusetts 01886
United States of America

Email: erosen@juniper.net