Internet Engineering Task Force (IETF)
Request for Comments: 9069
Updates: 7854
Category: Standards Track
ISSN: 2070-1721

T. Evens
Cisco Systems
S. Bayraktar
Menlo Security
M. Bhardwaj
Cisco Systems
P. Lucente
NTT Communications

February 2022

Support for Local RIB in the BGP Monitoring Protocol (BMP)

Abstract

The BGP Monitoring Protocol (BMP) defines access to local Routing Information Bases (RIBs). This document updates BMP (RFC 7854) by adding access to the Local Routing Information Base (Loc-RIB), as defined in RFC 4271. The Loc-RIB contains the routes that have been selected by the local BGP speaker's Decision Process.

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.

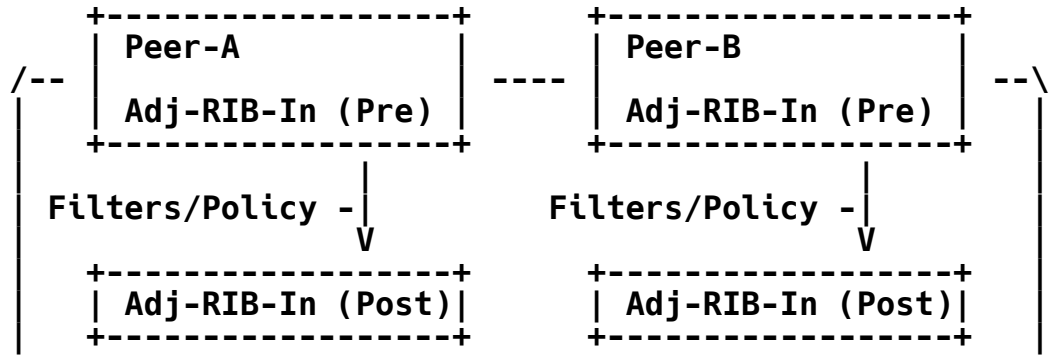
Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc9069.

Table of Contents

1. Introduction

This document defines a mechanism to monitor the BGP Loc-RIB state of remote BGP instances without the need to establish BGP peering sessions. BMP [RFC7854] does not define a method to send the BGP instance Loc-RIB. It does define locally originated routes in Section 8.2 of [RFC7854], but these routes are defined as the routes that originated into BGP (e.g., Section 9.4 of [RFC4271]). Loc-RIB includes all selected received routes from BGP peers in addition to locally originated routes.

Figure 1 shows the flow of received routes from one or more BGP peers into the Loc-RIB.



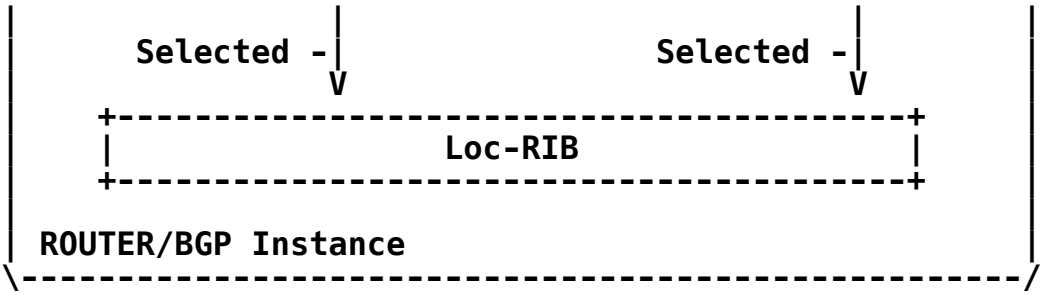


Figure 1: BGP Peering Adj-RIBs-In into Loc-RIB

The following are some use cases for Loc-RIB access:

* The Adj-RIB-In for a given peer post-policy may contain hundreds of thousands of routes, with only a handful of routes selected and installed in the Loc-RIB after best-path selection. Some monitoring applications, such as those that need only to correlate flow records to Loc-RIB entries, only need to collect and monitor the routes that are actually selected and used.

  Requiring the applications to collect all Adj-RIB-In post-policy data forces the applications to receive a potentially large unwanted data set and to perform the BGP decision process selection, which includes having access to the interior gateway protocol (IGP) next-hop metrics. While it is possible to obtain the IGP topology information using BGP - Link State (BGP-LS), it requires the application to implement Shortest Path First (SPF) and possibly Constrained Shortest Path First (CSPF) based on additional policies. This is overly complex for such a simple application that only needs to have access to the Loc-RIB.

* It is common to see frequent changes over many BGP peers, but those changes do not always result in the router's Loc-RIB changing. The change in the Loc-RIB can have a direct impact on the forwarding state. It can greatly reduce the time to troubleshoot and resolve issues if operators have the history of Loc-RIB changes. For example, a performance issue might have been seen for only a duration of 5 minutes. Post-facto troubleshooting this issue without Loc-RIB history hides any decision-based routing changes that might have happened during those 5 minutes.

* Operators may wish to validate the impact of policies applied to the Adj-RIB-In by analyzing the final decision made by the router when installing into the Loc-RIB. For example, in order to validate if multipath prefixes are installed as expected for all advertising peers, the Adj-RIB-In post-policy and Loc-RIB need to be compared. This is only possible if the Loc-RIB is available. Monitoring the Adj-RIB-In for this router from another router to derive the Loc-RIB is likely to not show the same installed prefixes. For example, the received Adj-RIB-In will be different if ADD-PATH [RFC7911] is not enabled or if the maximum supported number of equal paths is different between Loc-RIB and advertised routes.

This document adds Loc-RIB to the BGP Monitoring Protocol and

replaces Section 8.2 of [RFC7854] ("Locally Originated Routes").

1.1. Alternative Method to Monitor Loc-RIB

Loc-RIB is used to build Adj-RIB-Out when advertising routes to a peer. It is therefore possible to derive the Loc-RIB of a router by monitoring the Adj-RIB-In pre-policy from another router. This becomes overly complex and error prone when considering the number of peers being monitored per router.

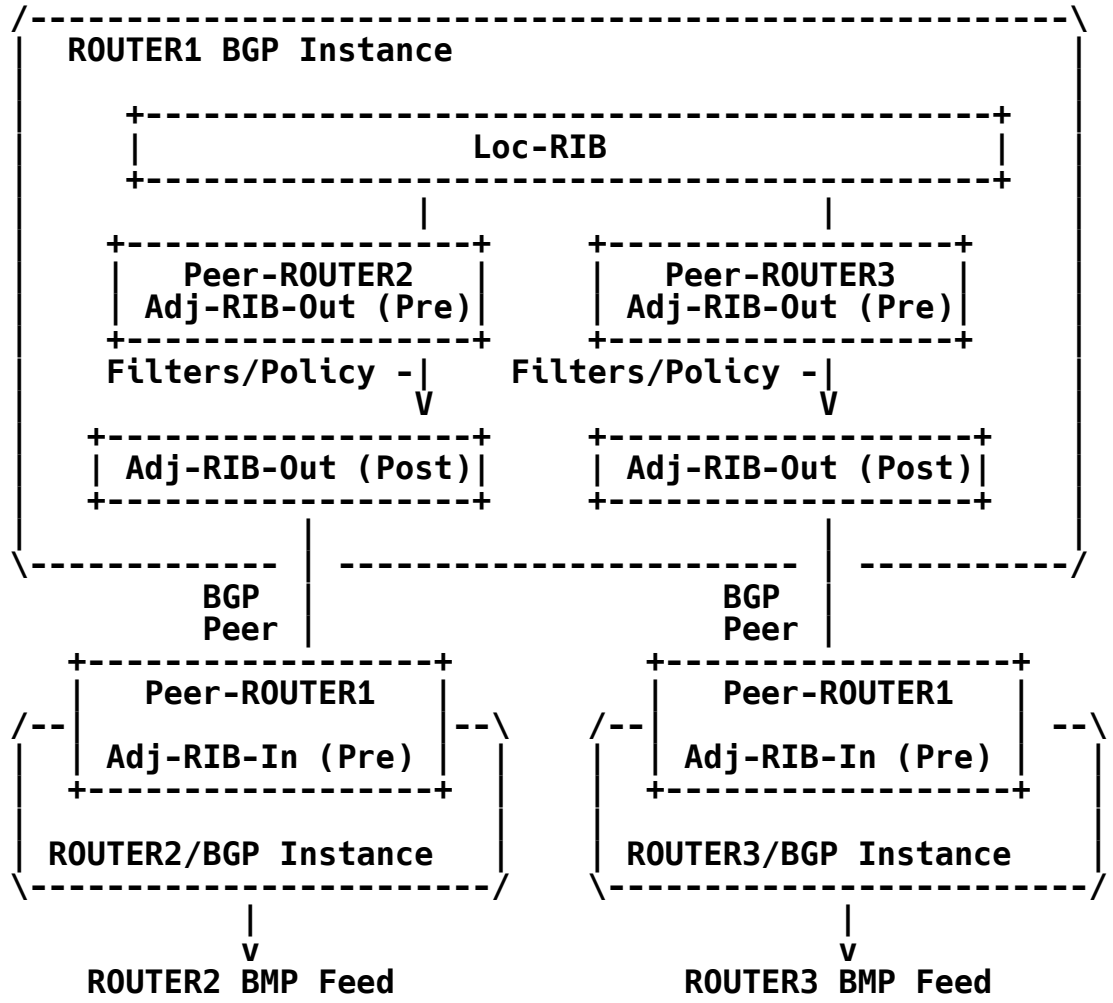


Figure 2: Alternative Method to Monitor Loc-RIB

The setup needed to monitor the Loc-RIB of a router requires another router with a peering session to the target router that is to be monitored. As shown in Figure 2, the target router Loc-RIB is advertised via the Adj-RIB-Out to the BMP router over a standard BGP peering session. The BMP router then forwards the Adj-RIB-In pre-policy to the BMP receiver.

A BMP lacking access to Loc-RIB introduces the need for additional resources:

* Requires at least two routers when only one router was to be

monitored.

* Requires additional BGP peering to collect the received updates when peering may not have even been required in the first place. For example, virtual routing and forwarding (VRF) tables with no peers, redistributed BGP-LS with no peers, and segment routing egress peer engineering where no peers have link-state address family enabled are all situations with no preexisting BGP peers.

Many complexities are introduced when using a received Adj-RIB-In to infer a router Loc-RIB:

* Adj-RIB-Out received as Adj-RIB-In from another router may have a policy applied that generates aggregates, suppresses more specific prefixes, manipulates attributes, or filters routes. Not only does this invalidate the Loc-RIB view, it adds complexity when multiple BMP routers may have peering sessions to the same router. The BMP receiver user is left with the error-prone task of identifying which peering session is the best representative of the Loc-RIB.

* BGP peering is designed to work between administrative domains and therefore does not need to include internal system-level information of each peering router (e.g., the system name or version information). In order to derive the Loc-RIB of a router, the router name or other system information is needed. The BMP receiver and user are forced to do some type of correlation using whatever information is available in the peering session (e.g., peering addresses, autonomous system numbers, and BGP identifiers). This leads to error-prone correlations.

* Correlating BGP identifiers (BGP-ID) and session addresses to a router requires additional data, such as router inventory. This additional data provides the BMP receiver the ability to map and correlate the BGP-IDs and/or session addresses but requires the BMP receiver to somehow obtain this data outside of the BMP. How this data is obtained and the accuracy of the data directly affect the integrity of the correlation.

2. Terminology

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 RFC 2119 [RFC2119] RFC 8174 [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Definitions

BGP Instance: Refers to an instance of BGP-4 [RFC4271], and considerations in Section 8.1 of [RFC7854] apply to it.

Adj-RIB-In: As defined in [RFC4271], "The Adj-RIBs-In contains unprocessed routing information that has been advertised to the local BGP speaker by its peers." This is also referred to as the "pre-policy Adj-RIB-In" in this document.

Adj-RIB-Out: As defined in [RFC4271], "The Adj-RIBs-Out contains the routes for advertisement to specific peers by means of the local speaker's UPDATE messages."

Loc-RIB: As defined in Section 1.1 of [RFC4271], "The Loc-RIB contains the routes that have been selected by the local BGP speaker's Decision Process." Note that the Loc-RIB state as monitored through BMP might also contain routes imported from other routing protocols such as an IGP or local static routes.

Pre-Policy Adj-RIB-Out: The result before applying the outbound policy to an Adj-RIB-Out. This normally represents a similar view of the Loc-RIB but may contain additional routes based on BGP peering configuration.

Post-Policy Adj-RIB-Out: The result of applying the outbound policy to an Adj-RIB-Out. This MUST be what is actually sent to the peer.

4. Per-Peer Header

4.1. Peer Type

A new peer type is defined for Loc-RIB to indicate that it represents the router Loc-RIB, which may have a route distinguisher (RD). Section 4.2 of [RFC7854] defines a Local Instance Peer type, which is for the case of non-RD peers that have an instance identifier.

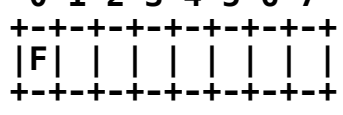
This document defines the following new peer type:

* Peer Type = 3: Loc-RIB Instance Peer

4.2. Peer Flags

If locally sourced routes are communicated using BMP, they MUST be conveyed using the Loc-RIB Instance Peer Type.

The per-peer header flags for the Loc-RIB Instance Peer Type are defined as follows:



* The F flag indicates that the Loc-RIB is filtered. This MUST be set when a filter is applied to Loc-RIB routes sent to the BMP collector.

  The unused bits are reserved for future use. They MUST be transmitted as 0, and their values MUST be ignored on receipt.

5. Loc-RIB Monitoring

The Loc-RIB contains all routes selected by the BGP Decision Process as described in Section 9.1 of [RFC4271]. These routes include those

learned from BGP peers via its Adj-RIBs-In post-policy, as well as routes learned by other means as per Section 9.4 of [RFC4271]. Examples of these include redistribution of routes from other protocols into BGP or those otherwise locally originated (i.e., aggregate routes).

As described in Section 6.1.2, a subset of Loc-RIB routes MAY be sent to a BMP collector by setting the F flag.

5.1. Per-Peer Header

All peer messages that include a per-peer header as defined in Section 4.2 of [RFC7854] MUST use the following values:

Peer Type: Set to 3 to indicate Loc-RIB Instance Peer.

Peer Distinguisher: Zero-filled if the Loc-RIB represents the global instance. Otherwise, set to the route distinguisher or unique locally defined value of the particular instance to which the Loc-RIB belongs.

Peer Address: Zero-filled. The remote peer address is not applicable. The V flag is not applicable with the Loc-RIB Instance Peer Type considering addresses are zero-filled.

Peer Autonomous System (AS): Set to the primary router BGP autonomous system number (ASN).

Peer BGP ID: Set the ID to the router-id of the VRF instance if VRF is used; otherwise, set to the global instance router-id.

Timestamp: The time when the encapsulated routes were installed in the Loc-RIB, expressed in seconds and microseconds since midnight (zero hour), January 1, 1970 (UTC). If zero, the time is unavailable. Precision of the timestamp is implementation dependent.

5.2. Peer Up Notification

Peer Up notifications follow Section 4.10 of [RFC7854] with the following clarifications:

Local Address: Zero-filled; the local address is not applicable.

Local Port: Set to 0; the local port is not applicable.

Remote Port: Set to 0; the remote port is not applicable.

Sent OPEN Message: This is a fabricated BGP OPEN message. Capabilities MUST include the 4-octet ASN and all necessary capabilities to represent the Loc-RIB Route Monitoring messages. Only include capabilities if they will be used for Loc-RIB monitoring messages. For example, if ADD-PATH is enabled for IPv6 and Loc-RIB contains additional paths, the ADD-PATH capability should be included for IPv6. In the case of ADD-PATH, the capability intent of advertise, receive, or both can be ignored

since the presence of the capability indicates enough that additional paths will be used for IPv6.

Received OPEN Message: Repeat of the same sent OPEN message. The duplication allows the BMP receiver to parse the expected received OPEN message as defined in Section 4.10 of [RFC7854].

5.2.1. Peer Up Information

The following Peer Up Information TLV type is added:

* Type = 3: VRF/Table Name. The Information field contains a UTF-8 string whose value MUST be equal to the value of the VRF or table name (e.g., RD instance name) being conveyed. The string size MUST be within the range of 1 to 255 bytes.

  The VRF/Table Name TLV is optionally included to support implementations that may not have defined a name. If a name is configured, it MUST be included. The default value of "global" MUST be used for the default Loc-RIB instance with a zero-filled distinguisher. If the TLV is included, then it MUST also be included in the Peer Down notification.

The Information field contains a UTF-8 string whose value MUST be equal to the value of the VRF or table name (e.g., RD instance name) being conveyed. The string size MUST be within the range of 1 to 255 bytes.

The VRF/Table Name TLV is optionally included to support implementations that may not have defined a name. If a name is configured, it MUST be included. The default value of "global" MUST be used for the default Loc-RIB instance with a zero-filled distinguisher. If the TLV is included, then it MUST also be included in the Peer Down notification.

Multiple TLVs of the same type can be repeated as part of the same message, for example, to convey a filtered view of a VRF. A BMP receiver should append multiple TLVs of the same type to a set in order to support alternate or additional names for the same peer. If multiple strings are included, their ordering MUST be preserved when they are reported.

5.3. Peer Down Notification

The Peer Down notification MUST use reason code 6. Following the reason is data in TLV format. The following Peer Down Information TLV type is defined:

* Type = 3: VRF/Table Name. The Information field contains a UTF-8 string whose value MUST be equal to the value of the VRF or table name (e.g., RD instance name) being conveyed. The string size MUST be within the range of 1 to 255 bytes. The VRF/Table Name informational TLV MUST be included if it was in the Peer Up.

5.4. Route Monitoring

Route Monitoring messages are used for initial synchronization of the Loc-RIB. They are also used to convey incremental Loc-RIB changes.

As described in Section 4.6 of [RFC7854], "Following the common BMP header and per-peer header is a BGP Update PDU."

5.4.1. ASN Encoding

Loc-RIB Route Monitoring messages MUST use a 4-byte ASN encoding as indicated in the Peer Up sent OPEN message (Section 5.2) capability.

5.4.2. Granularity

State compression and throttling SHOULD be used by a BMP sender to reduce the amount of Route Monitoring messages that are transmitted to BMP receivers. With state compression, only the final resultant updates are sent.

For example, prefix 192.0.2.0/24 is updated in the Loc-RIB 5 times within 1 second. State compression of BMP Route Monitoring messages results in only the final change being transmitted. The other 4 changes are suppressed because they fall within the compression interval. If no compression was being used, all 5 updates would have been transmitted.

A BMP receiver should expect that the granularity of Loc-RIB Route Monitoring can vary depending on the BMP sender implementation.

5.5. Route Mirroring

Section 4.7 of [RFC7854] defines Route Mirroring for verbatim duplication of messages received. This is not applicable to Loc-RIB as PDUs are originated by the router. Any received Route Mirroring messages SHOULD be ignored.

5.6. Statistics Report

Not all Stat Types are relevant to Loc-RIB. The Stat Types that are relevant are listed below:

* Stat Type = 8: (64-bit Gauge) Number of routes in Loc-RIB.

* Stat Type = 10: Number of routes in per-AFI/SAFI Loc-RIB. The value is structured as: 2-byte AFI, 1-byte SAFI, followed by a 64-bit Gauge.

6. Other Considerations

6.1. Loc-RIB Implementation

There are several methods for a BGP speaker to implement Loc-RIB efficiently. In all methods, the implementation emulates a peer with Peer Up and Down messages to convey capabilities as well as Route Monitor messages to convey Loc-RIB. In this sense, the peer that conveys the Loc-RIB is a locally emulated peer.

6.1.1. Multiple Loc-RIB Peers

There MUST be at least one emulated peer for each Loc-RIB instance, such as with VRFs. The BMP receiver identifies the Loc-RIB by the peer header distinguisher and BGP ID. The BMP receiver uses the VRF/Table Name from the Peer Up information to associate a name with the Loc-RIB.

In some implementations, it might be required to have more than one emulated peer for Loc-RIB to convey different address families for the same Loc-RIB. In this case, the peer distinguisher and BGP ID should be the same since they represent the same Loc-RIB instance. Each emulated peer instance MUST send a Peer Up with the OPEN message indicating the address family capabilities. A BMP receiver MUST process these capabilities to know which peer belongs to which address family.

6.1.2. Filtering Loc-RIB to BMP Receivers

There may be use cases where BMP receivers should only receive specific routes from Loc-RIB. For example, IPv4 unicast routes may include internal BGP (IBGP), external BGP (EBGP), and IGP, but only routes from EBGP should be sent to the BMP receiver. Alternatively, it may be that only IBGP and EBGP should be sent and IGP redistributed routes excluded. In these cases where the Loc-RIB is filtered, the F flag is set to 1 to indicate to the BMP receiver that the Loc-RIB is filtered. If multiple filters are associated with the same Loc-RIB, a table name MUST be used in order to allow a BMP receiver to make the right associations.

6.1.3. Changes to Existing BMP Sessions

In case of any change that results in the alteration of behavior of an existing BMP session, i.e., changes to filtering and table names, the session MUST be bounced with a Peer Down / Peer Up sequence.

7. Security Considerations

The same considerations as in Section 11 of [RFC7854] apply to this document. Implementations of this protocol SHOULD require that sessions only be established with authorized and trusted monitoring devices. It is also believed that this document does not introduce any additional security considerations.

8. IANA Considerations

IANA has assigned new parameters to the "BGP Monitoring Protocol (BMP) Parameters" registry (https://www.iana.org/assignments/bmp-parameters/).

8.1. BMP Peer Type

IANA has registered the following new peer type (Section 4.1):

Peer Type	Description

3	Loc-RIB Instance Peer

Table 1: BMP Peer Type

8.2. BMP Loc-RIB Instance Peer Flags

IANA has renamed "BMP Peer Flags" to "BMP Peer Flags for Peer Types 0 through 2" and created a new registry named "BMP Peer Flags for Loc-RIB Instance Peer Type 3".

This document defines peer flags that are specific to the Loc-RIB Instance Peer Type. IANA has registered the following in the "BMP Peer Flags for Loc-RIB Instance Peer Type 3" registry:

Flag	Description
0	F flag

Table 2: Loc-RIB Instance Peer Type

As noted in Section 4.2, the F flag indicates that the Loc-RIB is filtered. This indicates that the Loc-RIB does not represent the complete routing table.

Flags 1 through 7 are unassigned. The registration procedure for the registry is Standards Action.

8.3. Peer Up Information TLV

IANA has renamed the "BMP Initiation Message TLVs" registry to "BMP Initiation and Peer Up Information TLVs". Section 4.4 of [RFC7854] indicates that both Initiation and Peer Up share the same information TLVs. This document defines the following new BMP Peer Up Information TLV type (Section 5.2.1):

Type	Description
3	VRF/Table Name

Table 3: BMP Peer Up Information TLV Type

The Information field contains a UTF-8 string whose value MUST be equal to the value of the VRF or table name (e.g., RD instance name) being conveyed. The string size MUST be within the range of 1 to 255 bytes.

8.4. Peer Down Reason Code

IANA has registered the following new BMP Peer Down reason code (Section 5.3):

Type	Description
6	Local system closed, TLV data follows

Table 4: BMP Peer Down Reason Code

8.5. Deprecated Entries

Per this document, IANA has marked the F Flag entry in the "BMP Peer Flags for Peer Types 0 through 2" registry as "deprecated".

9. References

9.1. 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>.

[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>.

[RFC7854] Scudder, J., Ed., Fernando, R., and S. Stuart, "BGP Monitoring Protocol (BMP)", RFC 7854, DOI 10.17487/RFC7854, June 2016, <https://www.rfc-editor.org/info/rfc7854>.

[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>.

9.2. Informative References

[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>.

Acknowledgements

The authors would like to thank John Scudder, Jeff Haas, and Mukul Srivastava for their valuable input.

Authors' Addresses

Tim Evens
Cisco Systems
2901 Third Avenue, Suite 600

Seattle, WA 98121
United States of America

Email: tievens@cisco.com

Serpil Bayraktar
Menlo Security
800 W El Camino Real, Suite 250
Mountain View, CA 94040
United States of America

Email: serpil.bayraktar@menlosecurity.com

Manish Bhardwaj
Cisco Systems
3700 Cisco Way
San Jose, CA 95134
United States of America

Email: manbhard@cisco.com

Paolo Lucente
NTT Communications
Siriusdreef 70-72
2132 Hoofddorp
Netherlands

Email: paolo@ntt.net