IDR
Internet-Draft
Intended status: Standards Track
Expires: September 26, 2019

J. Heitz
P. Narasimha
S. Gulrajani
Cisco

March 25, 2019

BGP Diagnostic Path Attribute
draft-heitz-idr-diagnostic-attr-01

Abstract

A BGP path attribute for the purpose of network diagnostics is described. It is non-transitive, such that BGP speakers will not forward it by default. It is structured as a list of elements. An element begins with the BGP identifier and AS number of the speaker that appends the element and includes a list of TLVs. Any speaker can add or remove an element to/from the list. TLVs for a timestamp and for a checksum are described.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on September 26, 2019.

Copyright Notice

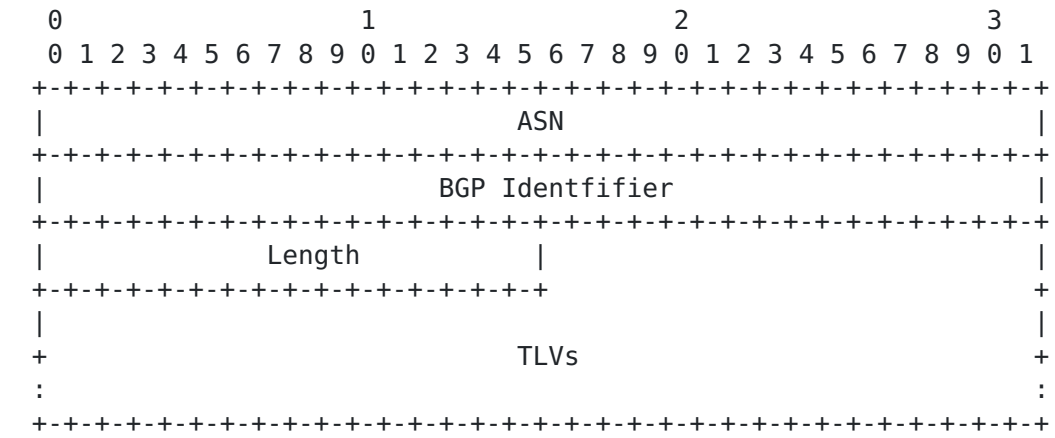
Table of Contents

1. Introduction

A BGP path attribute for the purpose of network diagnostics is described. It is non-transitive, such that BGP speakers that do not recognize the attribute will not propagate it by default. Even speakers that do recognize the attribute MUST NOT propagate it by default. A speaker MAY propagate the attribute if it is configured to do so and MAY add it's own information as it does so. The attribute is structured as a list of elements. An element begins with the BGP identifier and AS number of the speaker that appends the element and includes a list of TLVs. Any speaker can append or remove an element to/from the list. TLVs for a timestamp and for a checksum are described. The diagnostic attribute may be sent in a withdraw message.

2. Data Formats

The BGP diagnostic consists of a series of elements, each of which is formatted as follows.



The fields are as follows:

ASN - 4 octet Autonomous System Number of the speaker that appended this element.

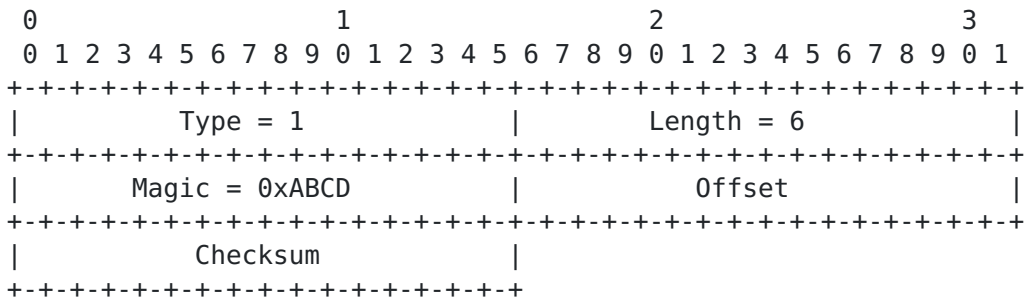
BGP Identfifier - BGP Identifier of the speaker that appended this element.

Length - The number of octets comprising the TLVs of this element. If there were no TLVs included, this length would be 0.

TLVs - Any number of TLVs as further described. Each TLV is optional. Each TLV comprises 2 octets of Type, then 2 octets specifying the number of octets in the value, then the octets of the value.

2.1. Checksum TLV

A checksum of the BGP message, including the marker field. The checksum is only valid between the sending and receiving speaker. Since a receiving speaker may propagate an update, it will likely change the set of attributes or their order in its own update message, thus making the checksum useless in the propagated update. A BGP speaker MAY remove the checksum TLV from a propagated Diagnostic Path Attribute.



The fields are as follows:

Type - 1.

Length - 6.

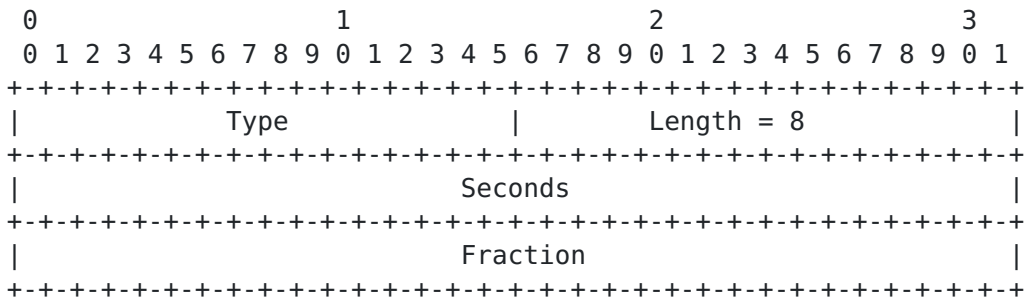
Magic - The value 0xABCD. This helps to diagnose corruption when looking at a hexdump.

Offset - The number of octets from the start of the UPDATE message (start of the marker) to the start of this TLV. This can help to identify corruption due to misaligned segment reassembly.

Checksum - The 16 bit checksum computed according to [RFC1071].

2.2. Timestamp TLV

The time at which the indicated speaker processed the indicated set of NLRI in the UPDATE message. There are several stages of processing for each NLRI, each of which may be timestamped. These stages vary widely between implementations. Therefore, the type code used for each stage is implementation dependent. One stage is universal. That is the stage when BGP hands the UPDATE message to TCP for transmission. The Type code for the timestamp for that stage is 256. The accuracy of the timestamp depends upon the diagnostic application that requires it and is out of scope of this document. The timestamp has enough bits to describe a time point within a period of 136 years. As time passes, the timestamp will simply wrap from one period to the next. For example, there exist some time points in the year 1900 and the year 2036 with identical timestamps. The determination of the period in which a timestamp occurs is out of scope of this document.



The fields are as follows:

Type - 256: The time when BGP hands the UPDATE message off to TCP.

- 257 - 511: Timestamps for any other stages of processing within BGP. The actual values and stages are implementation dependent.

Length - 8.

Seconds - The number of Seconds since 0 h 1 January 1900 UTC as further described in Section 6 of [RFC5905].

Fraction - A fraction of the above seconds also described in Section 6 of [RFC5905].

3. Usage

The Checksum TLV is useful to narrow down a source of corruption of UPDATE messages in each of the software and hardware layers between the actual BGP processes.

Because the Diagnostic Attribute contains a list of speakers that propagated an UPDATE and the attribute can be attached to a withdraw message, it can assist in the diagnosis of route oscillations.

The timestamp TLV is used to narrow down delays in UPDATE processing between BGP speakers and between the various stages of processing within a BGP speaker.

4. Operational Considerations

As with any new BGP attribute, if it is propagated, NLRI packing into BGP UPDATE messages may be affected. This needs to be taken into consideration when using the Timestamp TLV to measure bulk update message timing.

The Diagnostic Path Attribute MAY be sent in an UPDATE message that does not contain an NLRI field [RFC4271] or an MP_REACH_NLRI Path Attribute [RFC4760]. When carried in such a message, it is unlikely to be propagated, although it is possible.

If the addition or extension of the Diagnostic Path Attribute would cause the UPDATE message length to be exceeded, then the attribute SHOULD NOT be added or extended.

If the timestamps among participating speakers are not well synchronized, then the timestamps added by each speaker may appear out of order. In any case, the order in which elements are added to the Diagnostic Attribute can always be determined, because each speaker appends its element to the attribute.

The experimental TLV types may clash. That means that multiple vendors may use the same expreimental TLV type code for different purposes unbeknown to each other. To reduce the chance of TLV type code clashes, the type code for an experimental TLV SHOULD be configurable on the speaker. Because the propagation of the Diagnostic Attribute must be configured on each speaker, it is unlikely that two uncoordinated experiments will interfere with each other.

5. Error Handling

A checksum error SHALL NOT be treated as a protocol error. The response is implementation dependent.

A TLV length error SHALL be treated as attribute-discard according to [RFC7606].

An unrecognized TLV MUST not be treated as a protocol error.

6. Security Considerations

This attribute is not forwarded by default. Therefore, it should cause no unexpected ill effects.

7. IANA Considerations

IANA is requested to assign a BGP path attribute value for the BGP Diagnostic Path Attribute.

IANA is requested to create and maintain a registry for the TLV types. The allocation policies as per [RFC8126] are stated for the range of values.

Range	Allocation Policy
0-32767	First Come First Served
32768-65535	Experimental

Value	Description	Reference
0	Reserved	This RFC
1	Checksum	This RFC
256 - 511	Timestamp	This RFC

8. Acknowledgements

The authors would like to thank Shyam Sethuram and Bruno Decraene for suggestions that have been added to the document.

9. Normative References

[RFC1071] Braden, R., Borman, D., and C. Partridge, "Computing the Internet checksum", RFC 1071, DOI 10.17487/RFC1071, September 1988, <https://www.rfc-editor.org/info/rfc1071>.

[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>.

[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>.

[RFC5905] Mills, D., Martin, J., Ed., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, DOI 10.17487/RFC5905, June 2010, <https://www.rfc-editor.org/info/rfc5905>.

[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>.

Authors' Addresses

Jakob Heitz
Cisco
170 West Tasman Drive
San Jose, CA, CA 95134
USA

Email: jheitz@cisco.com

Prasad S. Narasimha
Cisco
170 West Tasman Drive
San Jose, CA, CA 95134
USA

Email: snprasad@cisco.com

Sameer Gulrajani
Cisco
170 West Tasman Drive
San Jose, CA, CA 95134
USA

Email: sameerg@cisco.com