IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 12, 2011

R. Raszuk
E. Chen
Cisco Systems
B. Decraene
France Telecom

March 11, 2011

BGP Diagnostic Message
draft-raszuk-bgp-diagnostic-message-02

Abstract

BGP protocol lacks self diagnostic tools which would allow for monitoring and detection of any possible bgp state database differences between BGP_RIB_Out of the sender and BGP_RIB_In of the receiver over BGP peering session. It also lacks of build in mechanism to inform peer about subset of prefixes received over session which experienced some errors and which per protocol specification either resulted in attribute drop or "treat-as-withdraw" action.

The intention of this document is to start a new class of work which will make BGP protocol and therefor assuring services constructed with the help of BGP protocol to become much more reliable and robust.

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 http://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 12, 2011.

Copyright Notice

Table of Contents

1. Introduction

In this document we will first define a new diagnostic communication channel in the form of new BGP message then construct the set of basic message encoding to be used for simple diagnostic self test routines periodically exchanged between BGP speakers. We will also define set of other TLVs which can be very useful in precise description of prefixes affected by various cases of BGP session malfunctions.

The goal of this document is to provide the background which will in turn allow for very easy extensibility once new needs and new BGP diagnostic ideas surface.

2. Applications

Authors would like to propose four main applications which BGP Diagnostic TLVs are designed to address. New TLVs can be easily added to enhance further current applications or to propose new applications.

The set of TLVs is organized in the following application groups:

General TLVs used for operational purposes of the described mechanism.

Set of TLVs designed to carry information about BGP state across BGP peers that include per neighbor counters and global counters. There are two modes this functionality can be used - on demand by explicit query as well as periodic in an automated mode. The scope of messages is to be able to operate both on the iBGP as well as eBGP boundaries. It is in the control of the operator to decide which set of information would be send to a given set of peers.

Messages which operate in an automated push mode (as long as peer negotiated listen capability for them) and are designed to inform BGP peer on the list of impacted NLRIs which were received along with malformed attribute or within malformed update message.

Following recommendation from MP-BGP4 RFC4760 next group of messages are used to indicate which AFI/SAFIs were disabled for any further processing by BGP peer due to detection of an incorrect attribute present in the BGP Update message.

In number of troubleshooting efforts in real networks it is often very helpful to verify state of a given prefix in the neighboring

router's BGP database. This is particularly useful on the EBGP boundaries where there is no CLI/SNMP access to the router. Authors define a new way of query peer's BGP for the state of particular prefix.

Last set of messages is an attempt to allow for intra-domain better analysis of the BGP best path selection tie break decisions.

3. BGP diagnostic message

When defining any self test tool the critical element is to find a right separation balance between the test object and testing instruments.

For the vast majority of real BGP issues found in the life production networks authors believe that the right balance is the definition of new BGP message which could be exchanged along with any negotiated AFI/SAFI between those BGP speakers which will during initial OPEN message exchange new BGP diagnostic message capability.

The two extreme alternatives which were considered were the definition of new BGP attribute which may inherit and share potential issues of given BGP address family it is designed to diagnose and on the other extreme to build a separate and independent network diagnostic protocol. The use of BGP message seems to provide sufficient isolation from any service address family and is much easier to deploy then enabling an entire new intra and inter-domain protocol. Another very important issue with using any other protocol for detection of potential differences of BGP databases state is lack of synchronization with BGP UPDATE messages. This alone in the continuously churning BGP environment would not allow for any benefit.

3.1. BGP DIAGNOSTIC Message Encoding

BGP message as defined in RFC 4271 consists of a fixed-size header followed by two octet length field and one octet of type value. RFC 4271 limits maximum message size to 4096 octets. As one of the applications of BGP Diagnostic message is to be able to carry entire potentially malformed BGP message this specification extends the maximum size of BGP Diagnostic message to be always 128 octets bigger then any other BGP Message. Considering the current RFC 4271 maximum BGP message size to be 4096 octets maximum size of BGP diagnostic message would be 4224 octets.

For the purpose of diagnostic message information encoding we will

use one or more Type-Length-Value containers where each TLV will have the following format:

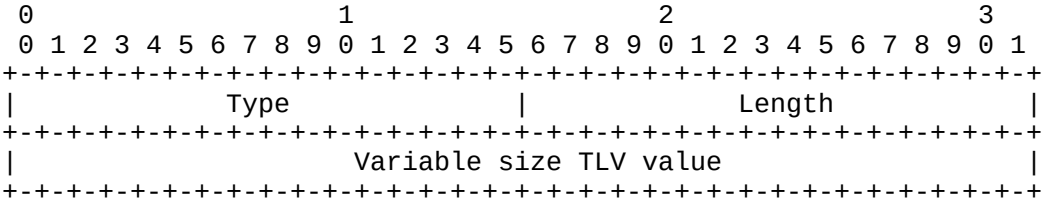

Figure 1: DIAGNOSTIC message TLV Format

Type - 2 octet value indicating the TLV type
Length - 2 octet value indicating the TLV length in octets
Value - Variable length value field depending on the type of the TLVs carried.

To work around continued BGP churn issue some types of TLVs will need to contain a sequence number to correlate request with associated to it replies. The sequence number will consist of 8 octets and will be of form: 4 octet bgp_router_id + local 4 octet number. When local 4 octet number reaches 0xFFFF it should restart from 0x0000.

Typical application scenario for use of sequence number is to include it in the diagnostic request message and during reply to copy it into reply messages triggered by such request message.

3.2. BGP DIAGNOSTIC Message TLVs

This document defines the following diagnostic TLV types:

* Operational TLVs

* BGP database counters exchange

* Diagnostics for encoding errors in BGP messages

* AFI/SAFI signaling when malformed update

* Prefix specific BGP debugging

* Intra-domain bgp decision monitoring

* Exchange of Route Target filters

* Errors and warnings detected when validating BGP paths and prefixes

3.2.1. Operational TLVs

Type 1 - Diagnostic Message Periodic Request
Length - 2 octets - variable value

Value (N x 2 octets):
TLV type - 2 octets

Use: To indicate the request to periodically receive listed TLV information. TLV type of 0xFFFF indicates request to receive all available diagnostic TLVs from the peer.

Type 2 - Max frequency permitted
Length - 2 octets - variable value

Value (N x 4 octets):
TLV type - 2 octets
Frequency value in seconds two octets 0..65535

Special values:
0 - never send given diagnostic TLV
65535 - no TLV inter-gap minimum set

Use: To indicate in seconds the maximum frequency given TLV may be periodically sent to the bgp speaker

Type 3 - Diagnostic Message Query
Length - 2 octets - variable value
Sequence number - 8 octets

Value (N x 2 octets):
TLV type - 2 octets

Use: To interactively (during debugging/troubleshooting) request to receive listed TLV information. TLV type of 0xFFFF indicates request to receive all available diagnostic TLVs from the peer. TLV of type 0x0000 indicates request to receive a list of all enabled and available diagnostic TLV types from the peer towards querying BGP speaker. The support of this TLV type is mandatory.

Type 4 - Counter’s reset request
Length - 2 octets - variable value

Value (N x 2 octets):
TLV type - 2 octets - List of TLVs subject to counter’s reset.

Use: To request rest of per neighbor counters of a given TLV type. TLV type of 0xFFFF indicates request to zero all per neighbor counters.

```
   Type 5 - Not supported TLV reply
   Length - 2 octets - variable value

     Value (N x 3 octets):
       TLV type - 2 octets - TLV that is not supported by the peer
                  but where part of TLV Request or TLV Query message
       Error Code - 1 octet - Error code

           Error codes:

           0x01 - Wrong TLV value
           0x02 - TLV not supported for this peer
           0x03 - Max query frequency exceeded
           0x04 - Administratively disabled

     Use: To indicate to the peer that the TLV he has requeste
          either in TLV Request or in TLV Query message is not
          supported. The support of this TLV type is mandatory.


   Type 6 - Enabled and supported TLV types
   Length - 2 octets - variable value

     Value (N x 2 octets):
       TLV type - 2 octets - TLV that is enabled and supported
                             by the peer

     Use: To indicate to the peer that the enclosed list of TLVs
          can be requested either in TLV Request or in TLV Query
          messages. The support of this TLV type is mandatory.
```

3.2.2. BGP database counters exchange

Type 7 - Number of Reachable Prefixes Transmitted/Received
Length - 2 octets - variable value
Sequence number - 8 octets

Value (N x 11 octets):
AFI/SAFI - 3 octets
Number of prefixes transmitted - 4 octets
Number of prefixes received - 4 octets

Use: To indicate number of reachable prefixes exchanged for a given AFI/SAFI between two bgp speakers. This message can be sent only based on the remote query Type 3 which contains the query sequence number to be placed in the reply.

Type 8 - Number of prefixes in BGP_RIB_Out
Length - 2 octets - variable value

Value (N x 7 octets):
AFI/SAFI - 3 octets
Number of prefixes 4 octets

Use: To indicate number of prefixes kept in BGP_RIB_Out between bgp speakers for a given AFI/SAFI between two bgp speakers.

Type 9 - Number of paths in BGP_RIB_Out
Length - 2 octets - variable value

Value (N x 6 octets):
AFI/SAFI - 3 octets
Number of paths 4 octets

Use: To indicate number of paths kept in BGP_RIB_Out between bgp speakers for a given AFI/SAFI between two bgp speakers.

```
   Type 10 - Number of prefixes present in BGP_RIB
   Length - 2 octets - variable value

     Value (N x 6 octets):
       AFI/SAFI - 3 octets
       Number of prefixes 4 octets

     Use: To indicate number of prefixes kept in BGP RIB for a given
          AFI/SAFI.


   Type 11 - Number of paths present in BGP_RIB
   Length - 2 octets - variable value

     Value (N x 7 octets):
       AFI/SAFI - 3 octets
       Number of prefixes 4 octets

     Use: To indicate number of paths kept in BGP RIB for a given
          AFI/SAFI.
```

3.2.3. Diagnostics for encoding errors in BGP messages

```
   Type 12 - Reachable prefixes present in dropped attribute UPDATE msg
   Length - 2 octets - variable value

     Value (N octets):
       AFI/SAFI - 3 octets
       1 .. M - List of prefixes

     Use: To list reachable prefixes present in the update message
          where optional transitive attribute with partial bit set
          was malformed and has been removed from the update message.
          Prefix encoding should follow given AFI/SAFI definition.
```

Type 13 - Unreachable prefixes present in dropped attribute UPDATE msg
Length - 2 octets - variable value

Value (N octets):
AFI/SAFI - 3 octets
1 .. M - List of prefixes

Use: To list unreachable prefixes present in the update message where optional transitive attribute with partial bit set was malformed and has been removed from the update message. Prefix encoding should follow given AFI/SAFI definition.

Type 14 - Reachable prefixes present in malformed UPDATE msg
Length - 2 octets - variable value

Value (N octets):
AFI/SAFI - 3 octets
1 .. M - List of prefixes

Use: To list reachable prefixes present in the malformed update message which were subject to "treat-as-withdraw" behaviour. Prefix encoding should follow given AFI/SAFI definition.

Type 15 - Entire malformed update message enclosure
Length - 2 octets - variable value
Sequence number - 8 octets

Value:
Malformed message

Use: Propagate the malformed message to the peer upon it’s request or at the event of error detection. That includes propagation of messages which had malformed attribute, unparsable content or any other abnormal encoding. If more then a single message has been determined as malformed the subsequent replies will contain the same sequence number and should not be treated as an override.

3.2.4. AFI/SAFI signaling when malformed update

Type 16 - List of ignored AFI/SAFIs by the peer over given session
Length - 2 octets - variable value

Value (N octets):
1..M AFI/SAFI - 3 octets each

Use: To list those AFI/SAFIs which were detected to be malformed by the peer and while session is up were transitioned to IGNORE state.

Such case is inline with Multiprotocol Extensions RFC 4760 as per it's section 7 Error Handling:

"For the duration of the BGP session over which the UPDATE message was received, the speaker then SHOULD ignore all the subsequent routes with that AFI/SAFI received over that session".

3.2.5. Prefix specific BGP debugging

Type 17 - Prefix specific BGP query
Length - 2 octets - variable value

Value (N octets):
AFI/SAFI - 3 octets
Prefix under query
Prefix mask (optional)

Use: To query peer for the status of prefix under examination. When prefix mask is present the request is for exact match. When prefix mask is not present the request is for the longest match. Prefix encoding should follow given AFI/SAFI definition.

```
   Type 18 - Prefix specific BGP response
   Length - 2 octets - variable value

     Value (N octets):
       AFI/SAFI - 3 octets
       Prefix under query
           Prefix mask (optional)
           Prefix status (1 octet)

         Status:
           0x01 - prefix not found in BGP table
           0x02 - prefix in BGP table and active (in FIB)
           0x03 - prefix in BGP table and not-active (not in FIB)
                   0x04 - administratively disabled

     Use: To inform peer querying about the status of particular
          prefix status. Prefix encoding should follow given AFI/SAFI
          definition.


   Type 19 - BGP attribute based prefix query
   Length - 2 octets - variable value

     Value (N octets):
       AFI/SAFI - 3 octets
           Query Parameters - 1 octet
       BGP Attribute TLV

       Defined Query Parameters:
         Bit 0 - value 0 - Exact match
         Bit 0 - value 1 - Partial match

     Use: To query peer for the list of prefixes which paths contain
          given BGP attribute. BGP attribute encoding should follow
              given attribute's specification.
```

```
   Type 20 - BGP attribute based prefix reply
   Length - 2 octets - variable value

     Value (N octets):
       AFI/SAFI - 3 octets
           Query Parameters - 1 octet
       1 .. M - List of prefixes

            Defined Query Parameters:
            Bit 0 - value 0  - Exact match
            Bit 0 - value 1  - Partial match

     Use: To inform bgp peer about presence of set of prefixes
          which contain with exact or partial match the BGP
          Attribute as specified in the query. Prefix encoding
          should follow given AFI/SAFI definition.
```

3.2.6. Intra-domain bgp decision monitoring

```
   Type 21 - Number of IGP metric best path tie breaks executed
   Length - 2 octets - variable value

     Value (N x 7 octets):
       AFI/SAFI - 3 octets
       Number of tie breaks 4 octets

     Use: To indicate number of prefixes with their best path selected
          by tie break of IGP metric to their BGP next hop distance
          step of BGP best path selection algorithm.


   Type 22 - Number of BGP best path tie breaks in each selection step
   Length - 2 octets - variable value

     Value (N x 7 octets):
       AFI/SAFI - 3 octets
           Best path selection step N - Number of tie breaks 4 octets

     Use: To indicate number of cases where in BGP best path selection
          algorithm given step has been used as a tie break during
          overall best path selection process for a given prefix.
```

3.2.7. Exchange of installed Route Target filters

```
Type 23 - Request for reception of route target filters
          installed towards given peer by RFC4684
Length - 2 octets - variable value
Sequence number - 8 octets

  Value (N x 7 octets):
    AFI/SAFI - 3 octets
    BGP Router ID of the peer - 4 octets

  Use: To request reception of full table of route target
       fiters installed towards listed BGP peer for a requested
       AFI/SAFI. Single request may contain multiple pairs of
       AFI/SAFIs and/or BGP Router IDs.


Type 24 - Reply containing all route target filters installed
          towards given peer
Length - 2 octets - variable value
Sequence number - 8 octets

  Value (7 + N * 12 or 24 octets):
    AFI/SAFI - 3 octets
    BGP Router ID of the peer - 4 octets
        List of route targets - each 12 or 24 octets

  Use: Allows for troubleshooting purposes to share list of
       route targets installed for a given AFI/SAFI towards
       indicated BGP peer. In the event that RT filtering
       table size will not fit in single BGP Diagnostic
       Message reply the subsequent reply should include
       the same sequence number.
```

4. Operation

BGP implementation which supports DIAGNOSTIC message can support all or subset of defined diagnostic types. The range of supported TLV types will be signaled in the new BGP capability message during BGP connection establishment phase.

The operation of this extension can be realized on a pool/query based or push based principles. An implementation may provide, a timer to periodically send selected Diagnostic types TLVs to the peer or to the management station.

Similarly BGP peer may periodically or by manual cli request the reception of selected or all of the defined diagnostic TLV types.

The received values are then compared against local counters. When discrepancy is found operator is alarmed and further analysis should follow. The repair actions is out of scope of this document.

Example:

Under some situations when determined that the discrepancy is detected an automated or manual Route Refresh message can be triggered with it’s extension for Start_of_Refresh and End_of_Refresh markers . That would allow for purge of any stalled data across two BGP databases.

An important point which needs to be discussed is the exchange of counter’s values in light of continued BGP churn presence. As BGP is never stable it is expected that any sort of described counters will also be subject to continues value change making any comparison of their values questionable.

There are three classes of counters defined in this document: sent counters, received counters and current table state counters.

Only "sent" counters can be used for not correlated comparison and problem detection between any two BGP speakers. They are not subject to BGP churn issue due to the fact that DIAGNOSTIC messages would be exchanged inline with BGP UPDATE messages on a given session. An implementation must be able to freeze the received counters when comparing or displaying the received "sent" counters from BGP peer.

Received counters send in the Diagnostic messages are only meaningful in the context of explicit request trigger situation generated by the BGP speaker. BGP speaker should stop transmitting any BGP message of a given AFI/SAFI or freeze corresponding counter after sending diagnostic message request to the peer and before reception of actual diagnostic message reply. In order to correlate diagnostic message requests with associated replies use of build in sequence numbers is provided.

Table state counters (for example number of BGP RIB entries) are exchanged only for informational reasons and they should not be subject to comparison with any local counter values.

5. Capability negotiation

A BGP speaker that is willing to send or receive the BGP DIAGNOSTIC

Messages from its peer should advertise the new DIAGNOSTIC Messages Capability to the peer using BGP Capabilities advertisement [BGP-CAP]. A BGP speaker may send a DIAGNOSTIC message to its peer only if it has received the DIAGNOSTIC message capability from its peer.

The Capability Code for this capability is specified in the IANA Considerations section of this document.

The Capability Length field of this capability is 2 octets. The Capability Value field consists of reserved flags field.

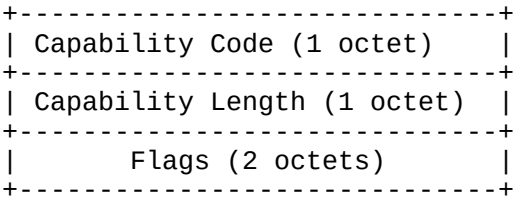

Figure 2: DIAGNOSTIC message BGP Capability Format

6. Security considerations

No new security issues are introduced to the BGP protocol by this specification.

7. IANA Considerations

IANA is requested to allocate a type code for the DIAGNOSTIC message from the BGP Message Types registry, as well as requesting a type code for the new Diagnostic Message Capability negotiation from BGP Capability Codes registry.

This document requests IANA to define and maintain a new registry named: "DIAGNOSTIC Message Type Values". The reserved types are: 0x0000 0xFFFF. The allocation policy is on a first come first served basis.

This document makes the following assignments for the DIAGNOSTIC Message Type Values:

Type 1 - Diagnostic Message TLV(s) Request
Type 2 - Max frequency permitted
Type 3 - Diagnostic Message TLV(s) Query
Type 4 - Counter's reset request
Type 5 - Not supported TLV
Type 6 - Enabled and supported TLV types

Type 7 - Number of Reachable Prefixes Transmitted/Received
Type 8 - Number of prefixes in BGP_RIB_Out
Type 9 - Number of paths in BGP_RIB_Out
Type 10 - Number of prefixes present in BGP_RIB
Type 11 - Number of paths present in BGP_RIB

Type 12 - Reachable prefixes present in dropped attribute message
Type 13 - Unreachable prefixes present in dropped attribute message
Type 14 - Reachable prefixes present in malformed UPDATE message
Type 15 - Entire malformed update message enclosure

Type 16 - List of ignored AFI/SAFIs by the peer over given session

Type 17 - Prefix specific BGP query
Type 18 - Prefix specific BGP response
Type 19 - BGP attribute based prefix query
Type 20 - BGP attribute based prefix reply

Type 21 - Number of IGP metric best path tie breaks executed
Type 22 - Number of BGP best path tie breaks in each selection step

Type 23 - Request for reception of route target filters
Type 24 - Reply containing all route target filters installed

Type 25 - 65534 Free for future allocation.
Type 65535 - Reserved

8. Acknowledgments

Authors would like to thank Alton Lo for his valuable input.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

9.2. Informative References

[I-D.retana-bgp-security-state-diagnostic] Retana, A. and R. Raszuk, "BGP Security State Diagnostic Message", draft-retana-bgp-security-state-diagnostic-00 (work in progress), March 2011.

[I-D.shakir-idr-ops-reqs-for-bgp-error-handling] Shakir, R., "Operational Requirements for Enhanced Error Handling Behaviour in BGP-4", draft-shakir-idr-ops-reqs-for-bgp-error-handling-01 (work in progress), February 2011.

Authors' Addresses

Robert Raszuk
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: raszuk@cisco.com

Enke Chen
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: enkechen@cisco.com

Bruno Decraene
France Telecom
38-40 rue du General Leclerc
Issi Moulineaux cedex 9 92794
France

Email: bruno.decraene@orange-ftgroup.com