IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 1, 2012

D. Freedman
Claranet
R. Raszuk
NTT MCL Inc.
R. Shakir
BT

February 29, 2012

BGP OPERATIONAL Message
draft-ietf-idr-operational-message-00

Abstract

The BGP Version 4 routing protocol (RFC4271) is now used in many ways, crossing boundaries of administrative and technical responsibility.

The protocol lacks an operational messaging plane which could be utilised to diagnose, troubleshoot and inform upon various conditions across these boundaries, securely, during protocol operation, without disruption.

This document proposes a new BGP message type, the OPERATIONAL message, which can be used to effect such a messaging plane for use both between and within Autonomous Systems.

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 1, 2012.

Table of Contents

1. Introduction

In this document, a new BGP message type, the OPERATIONAL message is defined, creating a communication channel over which messages can be passed, using a series of contained TLV elements.

The messages can be human readable, for the attention of device operators or machine readable, in order to provide simple self test routines, which can be exchanged between BGP speakers.

A number of TLV elements will be assigned to provide for these message types, along with TLV elements to assist with description of the message data, such as describing precisely BGP prefixes and encapsulating BGP UPDATE messages to be sent back for inspection in order to troubleshoot session malfunctions.

The use of OPERATIONAL messages will be negotiated by BGP Capability [RFC5492], since the messages are in-band with the BGP session, they can be assumed to either be authenticated as originating directly from the BGP neighbor.

The goal of this document is to provide a simple, extensible framework within which new messaging and diagnostic requirements can live.

2. Applications

The authors would like to propose three main applications which BGP OPERATIONAL 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 organised in the following four functional groups comprising the three applications and some control messaging:

- ADVISE TLVs, designed to convey human readable information to be passed, cross boundary to operators, to inform them of past or upcoming error conditions, or provide other relevant, in-band operational information. The "Advisory Demand Message" ADM (Section 3.4.1.1) is an example of this.

- STATE TLVs, designed to carry information about BGP state across BGP neighbors, including both per-neighbor and global counters.

- DUMP TLVs, designed to describe or encapsulate data to assist in realtime or post-mortem diagnostics, such as structured representations of affected prefixes / NLRI and encapsulated raw UPDATE messages for inspection.

- CONTROL TLVs, designed to facilitate control messaging such as replies to requests which can not be satisfied.

Means concerning the reporting of information carried by these TLVs, either in reply or request processing are implementation specific but could include methods such as SYSLOG.

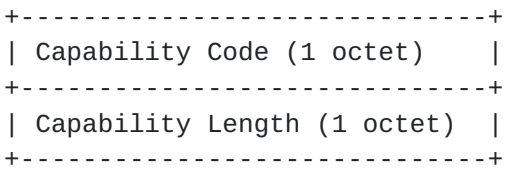
3. BGP OPERATIONAL message

3.1. BGP OPERATIONAL message capability

A BGP speaker that is willing to exchange BGP OPERATIONAL Messages with a neighbor should advertise the new OPERATIONAL Message Capability to the neighbor using BGP Capabilities advertisement [RFC5492] . A BGP speaker may send an OPERATIONAL message to its neighbor only if it has received the OPERATIONAL message capability from them.

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.

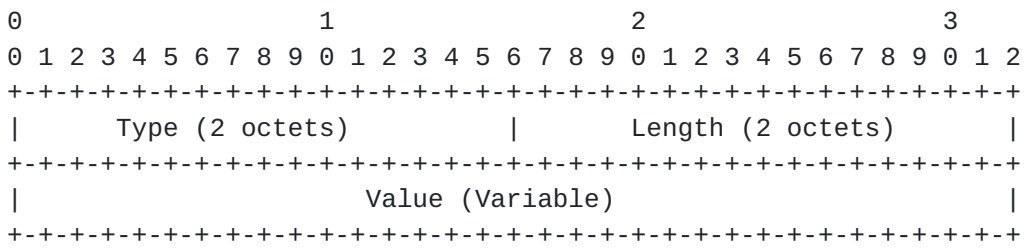


OPERATIONAL message BGP Capability Format

3.2. BGP OPERATIONAL message encoding

The BGP message as defined [RFC4271] consists of a fixed-size header followed by two octet length field and one octet of type value. The RFC limits the maximum message size to 4096 octets. As one of the applications of BGP OPERATIONAL message (through the MUD (Section 3.4.3.3) message) is to be able to carry an entire, potentially malformed BGP UPDATE, this specification mandates that when the neighbor has negotiated the BGP OPERATIONAL message capability, any further BGP message which may be subject to enclosure within a BGP OPERATIONAL message must be sent with the maximum size reduced to accommodate for the potential need of additional wrapping header size requirements. This is applicable to both the current BGP maximum message size limit or for any future modifications.

For the purpose of the OPERATIONAL message information encoding we will use one or more Type-Length-Value containers where each TLV will have the following format:



OPERATIONAL 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.

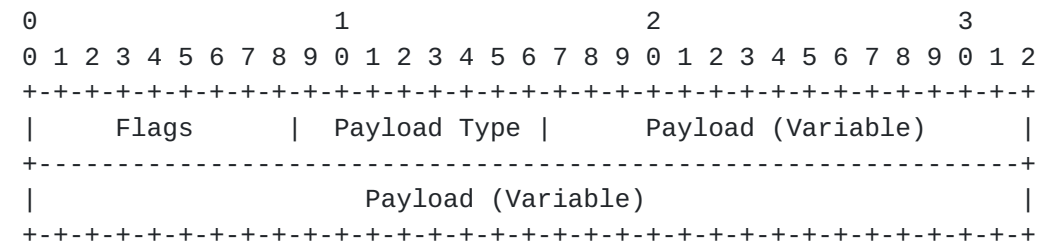
To work around continued BGP churn issues some types of TLVs will need to contain a sequence number to correlate a request with associated replies. The sequence number will consist of 8 octets and will be of the form: (4 octet bgp_router_id) + (local 4 octet number). When the local 4 octet number reaches 0xFFFF it should restart from 0x0000. The sequence number is only used if the TLV requires sequencing else it is not included.

The typical application scenario for use of the sequence number is for it to be included in a request TLV to be copied into associated reply messages in order to correlate requests with their associated replies.

3.3. PRI Format

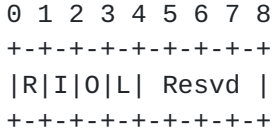
Prefix Reachability Indicators (PRI) are used to represent prefix NLRI and BGP attributes in a request and only prefix NLRI in a response, in this draft.

Each PRI is encoded as a 3-tuple of the form <Flags, Payload Type, Payload> whose fields are described below:



The use and the meaning of these fields are as follows:

a) Flags:
Four bits indicating NLRI Reachability:



aa) R Bit:
The R (Reachable) bit, if set represents that the prefixes were deemed reachable in the NLRI, else represents that the prefixes were deemed unreachable. This bit is meaningless in the context of all currently defined requests and can thus only be found in a response. If found in a request an implementation MUST ignore its state.

ab) I Bit:
The I (Adj-RIB-In) bit, if set in a query, indicates that the requestor wishes for the response to be found in the Adj-RIB-In of the neighbor representing this session, if cleared indicates that the Adj-RIB-In of the neighbor representing this session is not searched. If set in a response, indicates that the Adj-RIB-In of the neighbor representing this session contained this information, if cleared it did not.

ac) O Bit:
The O (Adj-RIB-Out) bit, if set in a query, indicates that the requestor wishes for the response to be found in the Adj-RIB-Out of the neighbor representing this session, if cleared indicates that the Adj-RIB-Out of the neighbor representing this session is not searched. If set in a response, indicates that the Adj-RIB-Out of the neighbor representing this session contained this information, if cleared it did not.

ad) L Bit:
The L (Loc-RIB) bit, if set in a query, indicates that the requestor wishes for the response to be found in the BGP Loc-RIB of the neighbor, if cleared indicates that the Loc-RIB of the neighbor is not searched. If set in a response, indicates that the Loc-RIB of the neighbor contained this information, if cleared it did not.

The rest of the field is reserved for future use.

b) Payload Type:
This one octet type specifies the type and geometry of the payload.

ba) Type 0 - NLRI:
The payload contains (perhaps multiple) NLRI, the format of each NLRI is as defined in the base specification of such NLRI appropriate for the AFI/SAFI.

bb) Type 1 - Next Hop:
The payload contains a Next Hop address, appropriate for the AFI/SAFI. When used in an SSQ (Section 3.4.2.7) message the response is expected to contain prefixes from the selected RIBs which contain this next-hop in their next-hop attribute.

bc) Type 2 - AS Number:
The payload contains a 16 or 32 bit AS number (as defined in [RFC4893]), when used in an SSQ message the response is expected to contain prefixes from the selected RIBs which contain this AS number in their AS_PATH or AS4_PATH (as appropriate) attributes.

bc) Type 3 - Standard Community:
The payload contains a standard community (as defined in [RFC1997]), when used in an SSQ message the response is expected to contain prefixes from the selected RIBs which contain this standard community in their communities attribute.

bd) Type 4 - Extended Community:
The payload contains an extended community (as defined in [RFC4360]), when used in an SSQ message the response is expected to contain prefixes from the selected RIBs which contain this standard community in their extended communities attribute.

be) Types 5-65535 - Reserved:
Types 5-65535 are reserved for future use.

c) Payload:
Contains the actual payload, as defined by the payload type, the payload is of variable length, to be calculated from the remaining TLV length.

PRI are used for both request and response modes, a response MUST only contain an NLRI (type 0) payload but a request MAY contain payloads specifying a type to search for, an implementation MUST validate all PRI it receives in a request against the type of request which was made.

An implementation MUST NOT send a PRI in response with no NLRI (type 0) payload, this is considered to be invalid. If the implementation wishes to signal that a request did not yield a any valid results an implementation MAY respond with an NS TLV (Section 3.4.4.2), using the "Not Found" subcode, for example.

3.4. BGP OPERATIONAL message TLVs

3.4.1. ADVISE TLVs

ADVISE TLVs convey human readable information to be passed, cross boundary to operators, to inform them of past or upcoming error conditions, or provide other relevant, in-band operational information.

3.4.1.1. Advisory Demand Message (ADM)

TYPE: 1 - ADM

LENGTH: 3 Octets (AFI+SAFI) + Variable value (up to 2K octets)

USE: To carry a message, on demand, comprised of a string of UTF-8 characters (up to 2K octets in size), with no null termination. Upon reception, the string SHOULD be reported to the host's administrator.

Implementations SHOULD provide their users the ability to transmit a free form text message generated by user input.

3.4.1.2. Advisory Static Message (ASM)

TYPE: 2 - ASM

LENGTH: 3 Octets (AFI+SAFI) + Variable value (up to 2K octets)

USE: To carry a message, on demand, comprised of a string of UTF-8 characters, with no null termination. Upon reception, the string SHOULD be stored in the BGP neighbor statistics field within the router. The string SHOULD be accessible to the operator by executing CLI commands or any other method (local or remote) to obtain BGP neighbor statistics (e.g. NETCONF, SNMP).

The expectation is that the last ASM received from a BGP neighbor will be the message visible to the operator (the most current ASM).

Implementations SHOULD provide their users the ability to transmit a free form text message generated by user input.

3.4.2. STATE TLVs

STATE TLVs reflect, on demand, the internal state of a BGP neighbor as seen from the other neighbor's perspective.

3.4.2.1. Reachable Prefix Count Request (RPCQ)

TYPE: 3 - RPCQ

LENGTH: 3 Octets (AFI+SAFI) + Sequence Number

USE: Sent to the neighbor to request that an RPCP (Section 3.4.2.2) message is generated in response.

3.4.2.2. Reachable Prefix Count Reply (RPCP)

TYPE: 4 - RPCP

LENGTH: 3 Octets (AFI+SAFI) + Sequence Number + 4 Octet RX Prefix Counter (RXC) + 4 Octet TX Prefix Counter (TXC)

USE: Sent in reply to an RPCQ (Section 3.4.2.1) message from a neighbor, RXC is populated with the number of reachable prefixes accepted from the peer and TXC with the number of prefixes to be transmitted to the peer for the AFI/SAFI.

3.4.2.3. Adj-Rib-Out Prefix Count Request (APCQ)

TYPE: 5 - APCQ

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number

USE: Sent to the neighbor to request that an APCP (Section 3.4.2.4) message is generated in response.

APCQ can be used as a simple mechanism when an implementation does not permit or support the use of RPCQ.

3.4.2.4. Adj-Rib-Out Prefix Count Reply (APCP)

TYPE: 6 - APCP

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number + 4 Octet TX Prefix Counter (TXC)

USE: Sent in reply to an APCQ (Section 3.4.2.3) message from a neighbor, TXC is populated with the number of prefixes held in the Adj-Rib-Out for the neighbor for the AFI/SAFI.

3.4.2.5. BGP Loc-Rib Prefix Count Request (LPCQ)

TYPE: 7 - LPCQ

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number

USE: Sent to the peer to request that an LPCP (Section 3.4.2.6) message is generated in response.

3.4.2.6. BGP Loc-Rib Prefix Count Reply (LPCP)

TYPE: 8 - LPCP

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number + 4 Octet Loc-Rib Counter (LC)

USE: Sent in reply to an LPCQ (Section 3.4.2.5) message from a neighbor, LC is populated with the number of prefixes held in the entire Loc-Rib for the AFI/SAFI.

3.4.2.7. Simple State Request (SSQ)

TYPE: 9 - SSQ

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number + Single request PRI (Variable)

USE: Using a PRI as a request form (See Section 3.3), an implementation can be asked to return information about prefixes found in various RIBs.

A single, simple PRI is used in the request, containing a single NLRI or attribute as the PRI payload. RIB response filtering may take place through the setting of the I, O and L bits in the PRI Flags

field.

An implementation MAY respond to an SSQ TLV in with an SSP (See Section 3.4.3.4) TLV (containing the appropriate data). An implementation MAY also respond to an SSQ with an NS TLV (with the appropriate subcode set) indicating why there will not be an SSP TLV in response. An implementation MAY also not respond at all (See Section 8).

3.4.3. DUMP TLVs

DUMP TLVs provide data in both structured and unstructured formats in response to events, for use in debugging scenarios.

3.4.3.1. Dropped Update Prefixes (DUP)

TYPE: 10 - DUP

LENGTH: 3 Octets(AFI+SAFI) + Variable number of dropped UPDATE Prefix Reachability Indicators (PRI) (See Section 3.3)

USE: To report to a neighbor a structured set of prefix reachability indicators retrievable from the last dropped UPDATE message, sent in response to an UPDATE message which was well formed but not accepted by the neighbor by policy.

For example, an UPDATE which was dropped and the rescued NLRI concerned a number of both reachable and unreachable prefixes, the DUP would encapsulate two PRI, one with the R-Bit (reachable) set, housing the rescued reachable NLRI and the other with the R-Bit cleared (unreachable), housing the rescued unreachable NLRI as payload.

3.4.3.2. Malformed Update Prefixes (MUP)

TYPE: 11 - MUP

LENGTH: 3 Octets(AFI+SAFI) + Variable number of dropped update Prefix Reachability Indicators (PRI) (See Section 3.3) due to UPDATE Malformation.

USE: To report to a neighbor a structured set of prefix reachability indicators retrievable from the last UPDATE message dropped through malformation, sent in response to an UPDATE message which was not well formed and not accepted by the neighbor, where a NOTIFICATION message was not sent. A MUP TLV may accompany a MUD (Section 3.4.3.3) TLV.

See the example from Section 3.4.3.1.

3.4.3.3. Malformed Update Dump (MUD)

TYPE: 12 - MUD

LENGTH: 3 Octets(AFI+SAFI) + Variable length representing retrievable malformed update octet stream.

USE: To report to a peer a copy of the last UPDATE message dropped through malformation, sent in response to an UPDATE message which was not well formed and not accepted by the neighbor, where a NOTIFICATION message was not sent. A MUD TLV may accompany a MUP (Section 3.4.3.2) TLV.

3.4.3.4. Simple State Response (SSP)

TYPE: 13 - SSP

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number + Single Response PRI (Variable)

USE: Using a PRI as a response form (See Section 3.3), an implementation uses the SSP TLV to return a response to an SSQ (See Section 3.4.2.7) TLV which should contain information about prefixes found in various RIBs. These RIBs should be walked to extract the information according to local policy.

A single, simple PRI is used in the response, containing multiple NLRI. The I, O and L bits in the PRI Flags field should be set indicating which RIBs the prefixes were found in.

An implementation MAY respond to an SSQ TLV in with an SSP TLV (containing the appropriate data). An implementation MAY also respond to an SSQ with an NS TLV (with the appropriate subcode set) indicating why there will not be an SSP TLV in response. An implementation MAY also not respond at all (See Section 8).

If no data is found to satisfy a query which is permitted to be answered, an implementation MAY respond with an NS TLV with the subcode "Not Found" to indicate that no data was found in response to the query. An implementation MUST NOT send a PRI in response with no NLRI payload, this is considered to be invalid.

3.4.4. CONTROL TLVs

CONTROL TLVs satisfy control mechanism messaging between neighbors, they are used for such functions as to refuse messages and

dynamically signal OPERATIONAL capabilities to neighbors during operation.

3.4.4.1. Max Permitted (MP)

TYPE: 65534 - MP

LENGTH: 3 Octets(AFI+SAFI) + 2 Octet Value

USE: The Max Permitted TLV is used to signal to the neighbor the maximum number of OPERATIONAL messages that will be accepted in a second of time (see Section 8, Security Considerations), an implementation MUST, on receipt of an MP TLV, ensure that it does not exceed the rate specified in the MP TLV for sending OPERATIONAL messages to the neighbor, for the duration of the session.

An implementation MAY send subsequent MP TLVs during the session's lifetime, updating the maximum acceptable rate

MP TLVs MAY be rate limited by the receiver as part of OPERATIONAL rate limiting (see Section 8, Security Considerations).

3.4.4.2. Not Satisfied (NS)

TYPE: 65535 - NS

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number + 2 Octet Error Subcode

USE: To respond to a query to indicate that the implementation can or will not answer this query. The following subcodes are defined:

0x01 - Request TLV Malformed: Used to signal to the neighbor that the request was malformed and will not be processed. A neighbor on receiving this message MAY re-transmit the request but MUST increment the sequence number. Implementations SHOULD ensure that the same request is not retransmitted excessively when repeatedly receiving this Error Subcode in response.

0x02 - TLV Unsupported for this neighbor: Used to signal to the neighbor that the request was unsupported and will not be processed. A neighbor on receiving this message MUST NOT retransmit the request for the duration of the session.

0x03 - Max query frequency exceeded: Used to signal to the neighbor that the request has exceeded the rate at which the neighbor finds acceptable for the implementation to transmit requests at, see Section 3.4.4.1 (MP TLV) and Section 8 and (Security Considerations) for more information.

0x04 - Administratively prohibited: Used to signal to the neighbor that the request was administratively prohibited and will not be processed. A neighbor on receiving this message MUST NOT retransmit the request for the duration of the session.

0x05 - Busy: Used to signal to the neighbor that the request will not be replied to, due to lack of resources estimated to satisfy the request. It is suggested that, on receipt of this error subcode a message is logged to inform the operator of this failure as opposed to automatically attempting to re-try the previous query.

0x06 - Not Found: Used to signal to the neighbor that the request would have been replied to but does not contain any data (i.e the data was not found). An implementation MUST NOT send a PRI response with no NLRI payload, this is considered to be invalid.

NS TLVs MAY be rate limited by the receiver as part of OPERATIONAL rate limiting (see Section 8, Security Considerations).

4. Use of the ADVISE TLVs

The BGP routing protocol is used with external as well as internal neighbors to propagate route advertisements. In the case of external BGP sessions, there is typically a demarcation of administrative responsibility between the two entities. While initial configuration and troubleshooting of these sessions is handled via offline means such as email or telephone calls, there is gap when it comes to advising a BGP neighbor of a behaviour that is occurring or will occur momentarily. There is a need for operators to transmit a message to a BGP neighbor to notify them of a variety of types of messages. These messages typically would include those related to a planned or unplanned maintenance action. These ADVISE messages could then be interpreted by the remote party and either parsed via logging mechanisms or viewed by a human on the remote end via the CLI. This capability will improve operator NOC-to-NOC communication by providing a communications medium on an established and trusted BGP session between two autonomous systems.

The reason that this method is preferred for NOC-to-NOC communications is that other offline methods do fail for a variety of reasons. Emails to NOC aliases ahead of a planned maintenance may have ignored the mail or may have not of recorded it properly within an internal tracking system. Even if the message was recorded properly, the staff that are on-duty at the time of the maintenance event typically are not the same staff who received the maintenance notice several days prior. In addition, the staff on duty at the time of the event may not even be able to find the recorded event in their internal tracking systems. The end result is that during a planned event, some subset of eBGP peers will respond to a session/ peer down event with additional communications to the operator who is initiating the maintenance action. This can be via telephone or via email, but either way, it may result in a sizeable amount of replies inquiring as to why the session is down.

The result of this is that the NOC responsible for initiating the maintenance can be inundated with calls/emails from a variety of parties inquiring as to the status of the BGP session. The NOC initiating the maintenance may have to further inquire with engineering staff (if they are not already aware) to find out the extent of the maintenance and communicate this back to all of the NOCs calling for additional information. The above scenario outlines what is typical in a planned maintenance event. In an unplanned maintenance event (the need for an immediate router upgrade/reload), the number of calls and emails will dramatically increase as more parties are unaware of the event.

With the ADVISE TLV set, an operator can transmit an OPERATIONAL

message just prior to initiating the maintenance specifying what event will happen, what ticket number this event is associated with and the expected duration of the event. This message would be received by BGP peers and stored in their logs as well as any monitoring system if they have this capability. Now, all of the BGP peers have immediate access to the information about this session, why it went down, what ticket number this is being tracked under and how long they should wait before assuming there is an actual problem. Even smaller networks without the network management capabilities to correlate BGP events and OPERATIONAL messages would typically have an operator login to a router and examine the logs via the CLI.

This draft specifies two types of ADVISE TLV, a DEMAND message (ADM) and a STATIC message (ASM), it is anticipated that the DEMAND message will be used to send a message, on demand to the BGP neighbor, to inform them of realtime events. The STATIC message can be used to provide continual, "Sticky" information to the neighbor, such as a contact telephone number or e-mail address should there be a requirement to have continual access to this information.

5. Use of the STATE TLVs

At the current time, the BGP-4 protocol, provides no mechanism by which the state of a remote system can be examined. Increasingly, as BGP-4 is utilised for additional applications, there is utility in providing in-band mechanisms for simple integrity checks, and diagnostic information to be exchanged between systems. As such, there are two sets of applications envisaged to be implemented utilising the STATE TLVs of the OPERATIONAL message.

5.1. Utilising STATE TLVs for Cross-Domain Debugging Functionality

In numerous cases, autonomous system boundaries represent a demarcation point between operational teams - in these cases, debugging the information received over a BGP session between the two systems is likely to result in human-to-human contact. In simple cases, this provides a particularly inefficient means by which specific queries regarding the routing information received via a BGP-4 session can be made. Whilst complex debugging is likely to continue to involve operational personnel, in a number of cases, it is advantageous for an operator to allow the remote administrative team to validate specific characteristics of the router's RIB. Such a means of debugging greatly enhances the speed of localising particular failures, and hence provides a potential reduction in the time to recovery of services dependent on the routing information transmitted via the BGP session. The STATE TLVs described in this document are intended to provide a mechanism by which requests for, and responses containing such debugging information can be implemented.

An example of the use of such a mechanism is on BGP-4 sessions making up a network-network interconnection carrying Layer 3 MPLS VPN [RFC4364] services - in these cases, such NNIs may be between particular administrative teams of the same network provider. The OPERATIONAL SSQ is intended to provide a simple query language that can be utilised to receive the subset of routing information that matches a particular query within the remote system's RIB. It is envisaged that such behaviour provides a simple means by which an operator can validate whether particular routing information is present, and as expected, on the remote system. Identification of inconsistencies quickly allows the device responsible for missing or incorrect information to be identified without direct interaction between humans.

5.2. Utilising STATE TLVs in the context of Error Handling

The enhancements to the BGP-4 protocol intended to provide more targeted error handling described in [I-D.ietf-idr-error-handling]

provide a number of cases whereby NLRI that are contained in particular UPDATEs may not be accepted by the remote BGP speaker. In this case, there is currently no mechanism by which an operator can identify whether the routing information received by the local speaker matches that which the remote speaker purports to have advertised. The Adj-Rib-Out Prefix Count Request (APCQ) and Reachable Prefix Count Request (RPCQ) are intended to provide means by which simple validation can be performed between two BGP speakers. It is envisaged that a BGP implementation can simply validate whether the remote system's RIB is consistent utilising such a mechanism, and hence trigger follow-up actions based on this. The extent of such follow-up actions is not intended to be defined by this document, however, it is envisaged that there is utility in such a state being flagged to an operational team to allow investigation of any inconsistency to be examined. Since many BGP-4 UPDATE message errors may be transient, validating the prefix counts in the local RIB against those received in response to the STATE TLV prefix count query messages described herein allows an operator to determine whether any inconsistency is persisting at the time of query, and hence whether any action is required.

In addition to allowing a manually-triggered validation of the RIB prefix counts, such a mechanism provides a simple means by which automated consistency checking can be enhanced on a BGP session. A device initiating a periodic check based on the RPCQ or APCQ TLVs can validate basic information regarding the number of entries in a particular RIB of a remote neighbor. Such consistency checks may trigger further (more detailed) sets of consistency validation mechanisms, or be flagged to a local operator. In this case, the potential forwarding black-holes that can be caused by inconsistency in the RIB of two systems can be quickly identified, and examined by an operator, or recovered from via an automated means such as a ROUTE-REFRESH message. As such, the use of the OPERATIONAL TLV in this case allows the resources on the BGP speakers involved to be minimised by allowing the speakers to perform a lightweight check prior to triggering any further action.

6. Use of the DUMP TLVs

Where a notable condition is experienced by a BGP-4 speaker, currently a limited set of responses are available to the speaker to make human network administrators aware of the condition. Within a local administrative boundary, logging functionality such as SNMP and SYSLOG can be used to record the occurrence of the event, as such, this provides visibility in an effective manner to the local administrator of the device. Whilst this provides a mechanism to make the router operator aware of erroneous states, or messages, where the condition is a direct result of an input from a remote system, or the information is of note to the remote BGP speaker, there is no means to communicate the detection of an erroneous condition to the remote device. As described in [I-D.ietf-grow-ops-reqs-for-bgp-error-handling] such conditions are likely to occur within the context of the handling of erroneous UPDATE messages.

The OPERATIONAL message intends to provide a number of message types to a BGP speaker that can be used to communicate information to a remote system. Whilst clearly free-text mechanisms such as the ADM provide a means by which arbitrary information can be transmitted, the use of a structured message type indicating particular message data can be transmitted back to the remote speaker provides means by which this information can be processed and reported directly. As such, the knowledge that particular OPERATIONAL messages relate to particular erroneous conditions that may be affecting network operation allows a system to determine any specific response actions, or prioritise any reporting to network management systems.

Where an UPDATE message's NLRI attribute can be wholly parsed, the pertinent information as to the prefixes that have been identified to be in the message is available to the receiving BGP speaker. Clearly, this information is of relevance to the administrators of the remote device, and is likely to provide some information regarding the contents of the message which is considered erroneous. The Malformed UPDATE Prefixes (MUP) TLV defined herein is intended to allow the receiving speaker to transmit the minimum required information regarding an UPDATE identified as malformed to the remote speaker without the overhead of additional path attributes (which may not be available to the receiving speaker). It is envisaged that the Dropped Update Prefixes (DUP) TLV provides analogous behaviour in the case where the UPDATE message is dropped due to local administrative policy, or implementation characteristics.

In some cases in order to determine the exact condition resulting in an error, there is a requirement for a network operator (or equipment implementor) to have an exact copy of the protocol message

transmitted to a remote system. The operational requirements presented in [I-D.ietf-grow-ops-reqs-for-bgp-error-handling] describe the operational advantage of logging a copy of such a message locally, however, where the message is erroneous due to a bug in the formation or transmission of the message by the sender, and the error is identified on the receiving speaker, this information is not available to the operator responsible for the erroneous network element. The Malformed UPDATE Dump (MUD) TLV is intended to be utilised to transmit an encapsulated copy of such a message back to the remote BGP speaker, and hence allow the operator to determine the exact formation of the invalid message.

7. Error Handling

An implementation MUST NOT send an OPERATIONAL message to a neighbor in response to an erroneous or malformed OPERATIONAL message. Any erroneous or malformed OPERATIONAL message received SHOULD be logged for the attention of the operator and then MAY be discarded.

8. Security considerations

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

Where a request type is not supported or allowed by an implementation for some reason, the implementation MAY send an NS (Section 3.4.4.2) TLV in response, the Error subcode of this TLV SHOULD be set according to the reason that this request will not be responded to.

Implementations MUST rate-limit the rate at which they transmit and receive OPERATIONAL messages. Specifically, an implementation MUST NOT allow the handling of OPERATIONAL messages to negatively impact any other functions on a router such as regular BGP message handling or other routing protocols.

Although an NS error subcode is provided to indicate that a request was rate-limited, an implementation need not reply to a request at all, this is the suggested course of action when rate-limiting the sending of responses to a neighbor.

An implementation MAY send an MP (Section 3.4.4.1) TLV to indicate the maximum rate at which it will accept OPERATIONAL messages from a neighbor, upon receipt of this TLV the sender MUST ensure it does not transmit above this rate for the duration of the session.

An implementation, considering a request to be too computationally expensive, MAY reply with the "Busy" NS error subcode to indicate such, though the implementation need not reply to the request.

Implementations MUST provide a mechanism for preventing access to information requested by SSR (Section 3.4.2.7) messages for the operator. Implementations SHOULD ensure that responses concerning the Loc-RIB (PRI with L-Bit set or responses which would set the L-Bit) are filtered in the default configuration.

9. IANA Considerations

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

This document requests IANA to define and maintain a new registry named: "OPERATIONAL Message Type Values". The allocation policy is on a first come first served basis.

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

ADVISE:

* Type 1 - Advisory Demand Message (ADM)

* Type 2 - Advisory Static Message (ASM)

STATE:

* Type 3 - Reachable Prefix Count Request (RPCQ)

* Type 4 - Reachable Prefix Count Response (RPCP)

* Type 5 - Adj-RIB-Out Prefix Count Request (APCQ)

* Type 6 - Adj-RIB-Out Prefix Count Response (APCP)

* Type 7 - Loc-Rib Prefix Count Request (LPCQ)

* Type 8 - Loc-Rib Prefix Count Response (LPCP)

* Type 9 - Simple State Request (SSQ)

DUMP:

* Type 10 - Dropped Update Prefixes (DUP)

* Type 11 - Malformed Update Prefixes (MUP)

* Type 12 - Malformed Update Dump (MUD)

* Type 13 - Simple State Response (SSP)

CONTROL:

* Type 65534 - Max Permitted (MP)

* Type 65535 - Not Satisfied (NS)

10. Acknowledgements

This memo is based on existing works [I-D.ietf-idr-advisory] and [I-D.raszuk-bgp-diagnostic-message] which describe a number of operational message types documented here. The authors would like to thank Enke Chen, Bruno Decraene, Alton Lo, Tom Scholl, John Scudder and Richard Steenbergen for their valuable input.

11. References

11.1. Normative References

[RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.

[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.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007.

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

11.2. Informative References

[I-D.ietf-grow-ops-reqs-for-bgp-error-handling]
Shakir, R., "Operational Requirements for Enhanced Error Handling Behaviour in BGP-4", draft-ietf-grow-ops-reqs-for-bgp-error-handling-02 (work in progress), October 2011.

[I-D.ietf-idr-advisory]
Scholl, T., Scudder, J., Steenbergen, R., and D. Freedman, "BGP Advisory Message", draft-ietf-idr-advisory-00 (work in progress), October 2009.

[I-D.ietf-idr-error-handling]
Scudder, J., Chen, E., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", draft-ietf-idr-error-handling-01 (work in progress), December 2011.

[I-D.jasinska-ix-bgp-route-server]
Jasinska, E., Hilliard, N., Raszuk, R., and N. Bakker, "Internet Exchange Route Server",

draft-jasinska-ix-bgp-route-server-03 (work in progress), October 2011.

[I-D.nalawade-bgp-inform]
Nalawade, G., Scudder, J., and D. Ward, "BGPv4 INFORM message", draft-nalawade-bgp-inform-02 (work in progress), August 2002.

[I-D.nalawade-bgp-soft-notify]
Nalawade, G., "BGPv4 Soft-Notification Message", draft-nalawade-bgp-soft-notify-01 (work in progress), July 2005.

[I-D.raszuk-bgp-diagnostic-message]
Raszuk, R., Chen, E., and B. Decraene, "BGP Diagnostic Message", draft-raszuk-bgp-diagnostic-message-02 (work in progress), March 2011.

[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.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

Authors' Addresses

David Freedman
Claranet
21 Southampton Row, Holborn
London WC1B 5HA
UK

Email: david.freedman@uk.clara.net

Robert Raszuk
NTT MCL Inc.
101 S Ellsworth Avenue Suite 350
San Mateo, CA 94401
US

Email: robert@raszuk.net

Rob Shakir
BT
pp C3L
BT Centre
81, Newgate Street
London EC1A 7AJ
UK

Email: rob.shakir@bt.com
URI: http://www.bt.com/