Operational Requirements for Enhanced Error Handling Behaviour in BGP-4

draft-ietf-grow-ops-reqs-for-bgp-error-handling-01

Abstract

BGP-4 is utilised as a key intra- and inter-Autonomous System routing protocol in modern IP networks. The failure modes as defined by the original protocol standards are based on a number of assumptions around the impact of session failure. Numerous incidents both in the global Internet routing table and within Service Provider networks have been caused by strict handling of a single invalid UPDATE message causing large-scale failures in one or more Autonomous Systems.
This memo describes the current use of BGP-4 within Service Provider networks, and outlines a set of requirements for further work to enhance the mechanisms available to a BGP-4 implementation when erroneous data is detected. Whilst this document does not provide specification of any standard, it is intended as an overview of a set of enhancements to BGP-4 to improve the protocol's robustness to suit its current deployment.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

Where BGP-4 [RFC4271] is deployed in the Internet and Service Provider
networks, numerous incidents have been recorded due to the manner in
which [RFC4271] specifies errors in routing information should be
handled. Whilst the behaviour defined in the existing standards retains
utility, the deployments of the protocol have changed within modern
networks, resulting in significantly different demands for protocol
robustness. Whilst a number of Internet Drafts have been written to
begin to enhance the behaviour of BGP-4 in terms of the handling of
erroneous messages, this memo intends to define a set of requirements
for ongoing work. These requirements are considered from the

perspective of a Network Operator, and hence this draft does not intend
to define the protocol mechanisms by which such error handling
behaviour is to be implemented.

1.1. Role of BGP-4 in Service Provider Networks

BGP was designed as an inter-Autonomous System (AS) routing protocol
and hence many of the error handling mechanisms within the protocol
specification are designed to be conducive to this role. In general,
this consideration as an inter-AS routing propagation mechanism results
in the view that a BGP session propagates a relatively small amount of
network-layer reachability information (NLRI) between two ASes. In this
case, it is the expectation of session resilience for those adjacencies
that are key to routing continuity (for example, it is expected that
two networks peering via BGP would connect multiple times in order to
safeguard equipment or protocol failure). In addition, there is some
expectation of multiple paths to a particular NLRI being available - it
would be expected that a network can fall back to utilising alternate,
less direct, paths where a failure of a more direct path occurs.
Traditional network architectures would deploy an Interior Gateway
Protocol (IGP) to carry infrastructure and customer prefixes, with an
Exterior Gateway Protocol (EGP) such as BGP being utilised to propagate
these prefixes to other Autonomous Systems. However, with the growth of
IP-based services, this is no longer considered best practice. In order
to ensure that convergence is within acceptable time bounds, the amount
of routing information carried within the IGP is significantly reduced
- and tends to be only infrastructure prefixes. iBGP is then utilised
to propagate both customer, and external prefixes within an AS. As
such, BGP has become an IGP, with traditional IGPs acting as a means by
which to propagate the routing information which is required to
establish a BGP session, and reach the egress node within the local
routing domain. This change in role presents different requirements for
the robustness of BGP as a routing protocol - with the expectation of
similar level of robustness to that of an IGP being set.
Along with this change in role, the nature of the IP routing
information that is carried has changed. BGP has become a ubiquitous
means by which service information can be propagated between devices.
For instance, BGP is utilised to carry routing information for IP/MPLS
VPN services as described in [RFC4364]. Since there is an existing
deployment of the protocol between PE devices in numerous networks, it
has been adapted to propagate this routing information, as its use
limits number of routing protocols required on each device. This
additional information being propagated represents a large change in
requirement for the error handling of the protocol - where session
failure occurs, it is likely a complete service outage for at least a
subset of a network's customers is experienced where an erroneous
packet may have occurred within a different sub-topology or even
service (a different address family for example). For this reason,
there is a significant demand to avoid service affecting failures that

may be triggered by routing information within a single sub-topology or
service.

Both within Internet and multi-service routing architectures, a number
of BGP sessions propagate a large proportion of the required routing
information for network operation. For Internet routing, these are
typically BGP sessions which propagate the global routing table to an
AS - failure of these sessions may have a large impact on network
service, based on a single erroneous update. In an multi-service
environment, typical deployments utilise a small number of core-facing
BGP sessions, typically towards route reflector devices. Failure of
these sessions may also result in a large impact to network operation.
Clearly, the avoidance of conditions requiring these sessions to fail
is of great utility to any network operator, and provides further
motivation for the revision of the existing behaviour.

Whilst the behaviour in [RFC4271] is suited to ensuring that BGP
messages with erroneous routing information in are limited in scope (by
means of session reset), with the above considerations, it is clear
that this mechanism is not suited to all deployments. It should,
however, be noted that the change in scope affects the handling only of
errors occurring after BGP session establishment. There is no current
operational requirement to amend the means by which error handling in
session establishment, or liveliness detection, are performed.

1.2. Overview of Operator Requirements for BGP-4 Error Handling

It is the intention of this document to define a set of criteria for
the manner in which a revised error handling mechanism in BGP-4 is
required to conform. The motivation for the definition of these
requirements can be summarised based on certain behaviour currently
present in the protocol that is not deemed acceptable within current
operational deployments, or where there is a short-fall in the tool set
available to an operator. These key requirements can be summarised as
follows:

    *It is unacceptable within modern deployments of the BGP-4
     protocol that a single erroneous UPDATE packet affects prefixes
     that it does not carry. This requirement therefore requires some
     modification to the means by which erroneous UPDATE packets are
     handled, and reacted to - with a particular focus on avoiding the
     use of the NOTIFICATION message.

    *It is recognised that some error conditions may occur within the
     BGP-4 protocol may not always be handled gracefully, and may
     result in conditions whereby an implementation cannot recover. In
     these (and similar) cases, it is unacceptable for an operator
     that this reset of the BGP-4 session results in interruption to
     forwarding packets (by means of withdrawing prefixes installed by
     BGP-4 into a device's RIB, and subsequently FIB). To this end,

there is a requirement to define a session reset mechanism which
provides session re-initialisation in a non-destructive manner.

*Further to the requirements to provide a more robust protocol,
 the current visibility into error conditions within the BGP-4
 protocol is extremely limited - where further modifications to
 this behaviour are to be made, complexity is likely to be added.
 Thus, to ensure that BGP-4 is manageable, there are requirements
 for mechanisms by which the protocol can be examined and
 monitored.

This document describes each of these requirements in further depth,
along with an overview of means by which they are expected to be
achieved. In addition, the mechanism by which the enhancements meeting
these requirements are to interact is discussed.

2. Errors within BGP-4 UPDATE Messages

Both through analysis of incidents occurring with the Internet DFZ, and
multi-service environments utilising BGP-4 to signal service or routing
information, a number of different classes of errors within BGP-4
UPDATE messages have been observed. In order to consider the
applicability of enhanced error handling mechanisms, it is possible to
divide these errors into a number of sub-classes, particularly focusing
around the location of the error within the UPDATE message.
Where an UPDATE message is considered invalid by a BGP speaker due to
an error within a path attribute that is not the NLRI (where the
definition of NLRI includes reachability information encoded in the
MP_REACH_NLRI and MP_UNREACH_NLRI attributes as specified in [RFC4760])
it is a requirement of any enhanced error handling mechanism to handle
the error in a manner focused on the NLRI contained within the message.
Since in this case, the message received from the remote peer is
syntactically valid, it is considered that such an UPDATE is indicative
of erroneous data within a path attribute - as such, it cannot be
assumed that the BGP speaker from whom the message was received is
directly responsible for the erroneous information - and hence
affecting all NLRI received via a specific session is disproportionate.
Two further error cases exist within UPDATE messages, both of which are
related to the mechanisms that are applicable to messages received
where some difficulty exists in parsing the entire BGP message. The two
cases concern those cases where a valid NLRI attribute can be
extracted, and those where such an attribute is not able to be parsed.
In these cases, errors in the packing of attributes within a BGP
message may have occurred. Such errors are likely indicative of an
error specifically caused by the remote BGP speaker. It is, however,
desirable to an operator that such errors are handled without affecting
all NLRI across a BGP session. As such, there is a key requirement to
maximise the number of cases in which it is possible to extract NLRI
from a BGP UPDATE message. To this end, it is required that where

possible the MP_REACH and MP_UNREACH attributes are utilised for encoding all NLRI (including IPv4 Unicast), and that this attribute is included as the first attribute of a BGP UPDATE message (as originally recommended in [I-D.chen-ebgp-error-handling]). Such a change to the order of inclusion of this attribute maximises the number of cases in which NLRI can be extracted from an UPDATE. Where this is possible, it is again required that the error handling mechanisms utilised should be directly applied to the NLRI included in the UPDATE.

For all cases whereby NLRI can be obtained from an UPDATE message, it is expected that the requirements outlined in Section 3 should be considered by any enhancement to the BGP-4 protocol.

In the case that it is not possible to completely parse the NLRI attribute from the UPDATE message received from a peer, it is extremely likely that this is indicative of a serious error with either the process of attribute packing, or buffer usage on the remote BGP speaker. In this case, clearly, it is not possible to apply any error handling mechanism that is limited to a specific set of NLRI, since an implementation has no knowledge of the NLRI included within the UPDATE message. In addition, such errors are considered to be relatively fundamental to the operation of a BGP implementation, and hence may indicate a case whereby significant system errors have occurred. The current BGP-4 standard results in a BGP speaker restarting a session with the remote BGP speaker. However where such an error does occur, it is required that a graceful mechanism is utilised to provide a lower impact to network operation. The requirements for enhancements of this nature to BGP-4 are outlined in Section 5, with the requirements outlined therein focused on providing a means by which system integrity can be restored whilst allowing for continued network operation.

3. Avoiding use of NOTIFICATION

The error handling behaviour defined in RFC4271 is problematic due to the limited options that are available to an implementation. When an erroneous BGP message is received, at the current time, the implementation must either ignore the error, or send a NOTIFICATION message, after which it is mandatory to terminate the BGP session. It is apparent that this requirement is at odds with that of protocol robustness.

There is significant complexity to this requirement. The mechanism defined in [I-D.chen-ebgp-error-handling] describes a means by which no NOTIFICATION message is generated for all cases whereby NLRI can be extracted from an UPDATE. The NLRI contained within the erroneous UPDATE message is considered as though the remote BGP speaker has provided an UPDATE marking it as withdrawn. This results in a limit in the propagation of the invalid routing information, whilst also ensuring that no traffic is forwarded via a previously-known path that may no longer be valid. This mechanism is referred to as "treat-as-withdraw".

Whilst this behaviour results in avoiding a NOTIFICATION message, keeping other routing information advertised by the remote BGP speaker within the RIB, it may result in unreachability for a sub-set of the NLRI advertised by the remote speaker. Two cases should be considered - that where the entry for a prefix in the Adj-RIB-In of the neighbour propagating an erroneous packet is utilised, and that where the prefix installed in the device's RIB is learnt from another BGP speaker. In the former case, should the identified NLRI not be treated as withdrawn, the original NLRI is utilised within the global RIB. However, this information is potentially now invalid (i.e. it no longer provides a valid forwarding path), whilst an alternate (valid) path may exist in another Adj-RIB-In. By continuing to utilise the NLRI for which the UPDATE was considered invalid, traffic may be forwarded via an invalid path, resulting in routing loops, or black-holing. In the second case, no impact to the forwarding of traffic, or global RIB, is incurred, yet where treat-as-withdraw is implemented, possibly stale routing information is purged from the Adj-RIB-In of the neighbour propagating errors.

Whilst mechanisms such as "treat-as-withdraw" are currently documented, the proposals are limited in their scope - particularly in terms of restrictions to implementation only on eBGP sessions. This limitation is made based on the view that the BGP RIB must be consistent across an autonomous system. By implementing treat-as-withdraw for a iBGP session, one or more routers within the Autonomous System may not have reachability to a prefix, and hence blackholing of traffic, or routing loops, may occur. It should, however, be considered if this view is valid, in light of the manner in which BGP is utilised within operator networks. Inconsistency in a RIB based on a single UPDATE being treated as withdrawn may cause a inconsistency in a single sub-topology (e.g. Layer 3 VPN service), or a service not operating completely (in the case of an UPDATE carrying service membership information). Where a NOTIFICATION and teardown is utilised this is destructive to all sub-topologies in all address family identifiers (AFIs) carried by the session in question. Even where mechanisms such as multi-session BGP are utilised, a whole AFI is affected by such a NOTIFICATION message. In terms of routing operation, it is therefore far less costly to endure a situation where a limited sub-set of routing information within an AS is invalid, than to consider all routing information as invalid based on a single trigger.

It is considered that, if extended to cover iBGP, the mechanisms described in [I-D.chen-ebgp-error-handling] and [I-D.ietf-idr-optional-transitive] provide a means to avoid the transmission of a NOTIFICATION to a remote BGP speaker based on a single erroneous message, where at all possible, and hence meet this requirement. The failure cases whereby NLRI cannot be extracted from the UPDATE message represent a case whereby the receiving system cannot handle the error gracefully based on this mechanism.

4. Recovering RIB Consistency

The recommendations described in Section 3 may result in the RIB for a topology within an AS being inconsistent across the AS' internal routers. Alternatively, where such mechanisms are deployed at an AS boundary, interconnects between two ASes may be inconsistent with each other. There are therefore risks of traffic blackholing, due to missing routing information, or forwarding loops. Whilst this is deemed an acceptable compromise in the short term, clearly, it is suboptimal. Therefore, a requirement exists to provide mechanisms by which a BGP speaker is able to recover the consistency of the Adj-RIB-In for a particular neighbour.

It is envisaged that during such routing inconsistencies, the local BGP speaker is aware that some routing information was not able to be processed - due to the fact that an UPDATE message was not parsed correctly. If the 'treat-as-withdraw' mechanism described within Section 3 is utilised, it is also possible for the local BGP speaker to have determined the set of NLRI for which an erroneous UPDATE message was received. In this scenario, by utilising targeted mechanisms to re-request the specific NLRI that was unreachable, this routing information can be re-transmitted from the remote BGP speaker. Such a request requires extension to the existing BGP-4 protocol, in terms of specific UPDATE generation filters with a transient lifetime. It is envisaged that the work within [I-D.zeng-one-time-prefix-orf] provides a mechanism allowing targeted elements of the Adj-RIB-In for a BGP neighbour to be recovered.

In addition to such cases where specific routing information is known to be erroneous, the more general case where either a large amount of the Adj-RIB-In is contained in UPDATE messages subject to treat-as-withdraw, or the specific prefixes are unknown to the local BGP speaker must be considered. In this case, there is a requirement for a BGP speaker to re-request the entire RIB advertised by a remote neighbour. In this case, where such re-advertisement is required, it is envisaged that a ROUTE-REFRESH as per the description in [RFC2918] is utilised. [I-D.keyur-bgp-enhanced-route-refresh] provides a means by which the ROUTE-REFRESH mechanism can be extended in order to meet this requirement.

It is of particular note for both means of recovering RIB consistency described that these are effective only when considering transitive errors within an implementation - for instance, should an RFC interpretation error within an implementation be present, regardless of the number of times a specific UPDATE is generated, it is likely that this error condition will persist. For this reason, there is an requirement to consider the means by which such consistency recovery mechanisms are utilised. It is not advisable that a transitive filter and advertisement mechanism is triggered by all error handling events due to the load this is likely to place on the neighbour receiving such a request. Where this BGP speaker is a relatively centralised device - a route reflector (as described by [RFC4456]) for example - the act of

generation of UPDATE messages with such frequency is likely to cause
disproportionate load. It is therefore an operational requirement of
such mechanisms that means of request dampening be required by any such
extension.

5. Reducing the Impact of Session Reset

Even where protocol enhancements allow errors in the BGP-4 protocol to
cease to trigger NOTIFICATION messages, and hence reset a BGP session,
it is clear that some error conditions may not be exited. In
particular, errors due to existing state, or memory structures,
associated with a specific BGP session will not be handled. It is
therefore important to consider how these error conditions are
currently handled by the protocol. It should be noted that the
following discussion and analysis considers only those NOTIFICATION
messages generated in response to errors in UPDATE messages (as defined
by Section 6.3 in [RFC4271]).
The existing NOTIFICATION behaviour triggers a reset of all elements of
the BGP-4 session, as described in Section 6 of [RFC4271]. It is
expected that session teardown requires an implementation to re-
initialise all structures and state required for session maintenance.
Clearly, there is some utility to this requirement, as error conditions
in BGP are, in general, exited from. However, this definition is
responsible for the forwarding outages within networks utilising BGP
for route propagation when each error is experienced. The requirement
described in Section 3 is intended to reduce the cases whereby a
NOTIFICATION is required, however, any mechanism implemented as a
response to this requirement by definition cannot provide a session
reset to the extent of that achieved by the current behaviour.
In order to address this, there is a requirement for a means by which a
BGP speaker can signal that an unhandled error condition in an UPDATE
message occurred - requiring a session reset - yet also continue to
utilise the paths advertised by the neighbour that are currently in use
within the RIB. In this case, the Adj-RIB-In received from the
neighbour is not considered invalid, despite a NOTIFICATION, and
session reset, being required. This set of requirements is akin to
those answered by the BGP Graceful Restart mechanism described in
[RFC4724]. Since the operational requirement in this case is to provide
a means to achieve a complete session restart without disrupting the
forwarding path of those prefixes in use within a BGP speaker's RIB, it
is expected that utilising a procedure similar to the Graceful Restart
mechanism meets the error handling requirement. By responding to an
error condition (repeated or otherwise) with a message indicating that
an error that cannot be handled has occurred, forcing session reset,
whilst retaining forwarding information within the RIB allows
forwarding to all prefixes within a system's RIB to continue, whilst
the session restarts. By placing a time bound on the restart lifetime,
should an error condition not be transient - for example, should an
error have occurred with the BGP process, rather than a specific of the

BGP session - the remote BGP speaker is still detected as an invalid device for forwarding.

It should, however, be noted that a protocol enhancement meeting this requirement is not able to solve all error conditions - however, a complete restart of the BGP and TCP session between two BGP speakers implements an identical recovery mechanism to that which is achieved by the existing behaviour. Where an error condition such as memory or configuration corruption has occurred in a BGP implementation, it is expected that a mechanism meeting this requirement continues to detect this, by means of a bound on time for session restart to occur. Whilst there may be some consideration that packets continue to be forwarded through a device which can be in an failure mode of this nature for a longer period, due to this requirement, the architecture of modern IP routers should be considered. A divided forwarding and control plane is common in many devices, as well as process separation for software-based devices - corruption of a specific protocol daemon does not necessarily imply forwarding is affected. Indeed, where forwarding behaviour of a device is affected, it is envisaged that a failure detection mechanism (be it Bidirectional Forwarding Detection, or indeed BGP KEEPALIVE packets) will detect such a failure in almost all cases, with the symptomatic behaviour of such a failure being an invalid UPDATE message in very few other cases.

6. Operational Toolset for Monitoring BGP

A significant complexity that is introduced through the requirements defined in this document is that of monitoring BGP session status for an operator. Although the existing error handling behaviour causes a disproportionate failure, session failure is extremely visible to most operational personnel within a Network Operator due to both existing definitions of SNMP trap mechanisms for BGP, along with the forwarding impact typically caused by such a failure. By introducing mechanisms by which errors of this nature are not as visible, this is no longer the case. There is a requirement that where subsets of the RIB on a device are no longer reachable from a BGP speaker, or indeed an AS, that some mechanism to determine the cause is available to an operator. Whilst, to some extent, this can be solved by mandating a sub-requirement of each of the aforementioned requirements that a BGP speaker must log where such errors occur, and are hence handled, this does not solve all cases. In order to clarify this requirement, the example of the transmission of an erroneous Optional Transitive attribute can be considered. Since, by definition, there is no requirement for all BGP speakers to parse such an attribute, a receiving router may treat NLRI as withdrawn based on an erroneous attribute not examined by its neighbour. In this case, the upstream device or network, propagating the UPDATE, has no visibility of this error. Operationally, however, it is of interest to the upstream router operator that such invalid information was propagated.

The requirement for logging of error conditions in transmitted BGP
messages, which are visible to only the receiver, cannot be achieved by
any existing BGP message, or capability. It is envisaged that each
erroneous event should be transmitted to the remote peer - including
the information as to the set of NLRI that were considered invalid.
Whilst with some mechanisms this is achieved by default (for example,
One-Time Prefix ORF *[I-D.zeng-one-time-prefix-orf]* (Outbound Route
Filtering) will transmit the set of prefixes that are required), the
operator requirement is to know which prefixes may have been
unreachable in all cases. It is envisaged that an extension to meet
this requirement will allow for such information to be transmitted
between peers, and hence logged. Such a mechanism may provide further
utility as a either a diagnostic, or logging toolset.
As such, it is possible to divide the messages that are required in
order to provide further visibility into BGP for an operator. Such a
division can be made both due to the required means of message
transmission, alongside the criticality of each request.

  *Messages required to replace NOTIFICATION - In cases where the
   error handling mechanisms defined by [RFC4271] currently result
   in a NOTIFICATION message being generated, a number of the
   requirements detailed within this document result this message
   being suppressed. Despite this change, the error condition's
   occurrence is still of interest to an operator, since some form
   of invalid data has been received on a session in order to
   provide both monitoring and troubleshooting capabilities. It
   therefore considered that an implementation must generate a
   message both locally, and transmitted to the remote peer, based
   on the such a condition. Where such a message is transmitted to
   the remote peer, it is considered that the BGP session via which
   the erroneous UPDATE message was received as transport to the
   remote peer. The information transmitted in such a message should
   be minimised to allow identification of the paths which were
   considered erroneous (i.e. restricting the information to that
   which is directly relevant to a network operator in the case of
   an error condition occurring). Any delay to convergence on the
   session in question is considered to be acceptable, given the
   suboptimal nature of the reception of invalid routing information
   via a BGP session. Further concerns regarding such a mechanism
   relate to the load generated on the BGP speaker in question,
   however, it must be considered that in the case of an erroneous
   UPDATE being received, and the 'treat-as-withdraw' mechanism
   being utilised, where the erroneous path is removed from the Loc-
   RIB, there is likely to be a requirement to generate UPDATE
   messages withdrawing the prefix from all further BGP speakers to
   which the prefix is advertised. The load generated by the
   generation of such UPDATEs is likely to be much greater than that
   of transmitting error information via a logging message type back
   to the speaker from which it was received. It is envisaged that

light-weight BGP message-based signalling mechanisms such as [I-D.ietf-idr-advisory] provide a suitable means to satisfy this requirement.

*Additional Diagnostic Capabilities for BGP - In a number of cases, there is an operational requirement to further debug erroneous BGP UPDATE messages, along with the particulars of the state of a BGP speaker. For instance, where an invalid BGP UPDATE message is transmitted between two BGP speakers, the exact format of the UPDATE message is of interest to an operator, as this information provides a clear indication of an message considered to be erroneous by the BGP speaker to which it was transmitted. In this case, it is considered of great utility that the entire UPDATE message is transmitted back to the advertising speaker, in order to allow for further debugging to occur. Whilst such information is particularly useful to an operator, it clearly provides information that is not key to protocol operation - for this reason, it is expected that some of the concerns regarding the additional complexity, and load that a BGP speaker is subjected to is not acceptable. For this reason, it is required that where mechanisms are developed to support this requirement, messages of this nature can be supported both within an existing BGP session, and via a dedicated separate session, be it BGP carrying messages such as DIAGNOSTIC *[I-D.raszuk-bgp-diagnostic-message]* or ADVISORY *[I-D.ietf-idr-advisory]* or a dedicated monitoring protocol akin to BMP described in [I-D.ietf-grow-bmp].

Whilst the operational requirement for such monitoring tools to allow for visibility into BGP is clearly agreed upon, the means by which such messages are transmitted between two BGP speakers is likely to be dependent upon both the positions of the speakers in question (for instances, the requirements for such a protocol may differ where a session is between two ASBRs under separate administration). The introduction of additional message types to the BGP protocol clearly introduces further complexity - and leaves room for further implementation and standardisation errors that may compromise the robustness of the BGP protocol. In addition, the queuing and scheduling of these BGP messages must be interleaved with the transmission of the key protocol messages - such as KEEPALIVE and UPDATE packets. It is therefore a concern that should a large number of messages specifically for operational visibility be transmitted, this will delay the transmission of UPDATE packets, and hence adversely affect the end-to-end convergence time for NLRI carried within BGP. The operational requirement for why messages are advantageous to be in-band to a protocol should also be considered. In particular, it should be noted that where such information is to be transmitted between administrative boundaries a BGP session represents an existing channel exists between the two ASes. This channel is considered to be secure insofar as the routing information, and requests sent via the session are considered

to come from a trusted source. Since error information relates to both
a particular attachment, and is key to ensuring that such a session is
operating as expected, it is considered of great operational benefit
that this information is transmitted over this channel. In addition,
the overall system scalability is improved by such in-band
transmission. It is expected that erroneous information resulting in
the 'treat-as-withdraw' mechanism being utilised is relatively
infrequently transmitted between two peers (when compared to the
frequency of UPDATE messages transmission). The impact of including an
additional BGP message type for such operational visibility is
relatively small from a resource utilisation perspective - additional
processing overhead is only experienced when such a message is
received. Where a separate session is maintained, particular network
elements within a service provider topology may require hundreds, or
thousands, of additional sessions for the transmission of this
information. Such an resource consumption overhead is likely to be
unacceptable to some network operators.
For the reasons explained above, it is expected that mechanisms
specified to meet the requirements for event visibility consider the
relative impacts of additional monitoring sessions, or message
inclusion in band to BGP in order not to compromise the security,
scalability and robustness of the BGP-4 protocol.

7. Operational Complexities Introduced by Altering RFC4271

The existing NOTIFICATION and subsequent teardown of a BGP session upon
encountering an error has the advantage that a consistent approach to
error handling is required of all implementations of the BGP-4
protocol. This is of operational advantage, as it provides a clear
expectation of the behaviour of the protocol. The requirements defined
herein add further complexity to the error-handling within BGP, and
hence are liable to compromise the existing deterministic protocol
behaviour. It is therefore deemed that there is a further requirement
to provide a clear method by which an erroneous UPDATE should be
reacted to, in order that all protocol implementations provide a
consistent means by which recovery is achieved. A further complexity is
introduced due to the disparate nature of the work items altering the
BGP error handling behaviour - since all items are likely to be
implemented as a BGP capability [RFC5492], situations are likely to
occur between devices (especially those with different BGP
implementations), where some of the mechanisms referenced are
unsupported. This adds further barriers to a standard definition of the
BGP-4 error handling behaviour.
In general, the approach considered ideal upon encountering an
erroneous UPDATE message can be divided into two cases - those where
the NLRI can be determined from the message, and those where it cannot
be. The latter case is the simpler of the two. In this case, there is a
requirement for the implementation to reset the BGP session, utilising
the reduced-impact approach, described in Section 5. In the case where

the remote BGP speaker is in a transient error condition related to specific peer data structures, or state, a single instance of this behaviour is likely to exit the error condition. In the case of implementation errors, it is possible that the BGP session in question may enter a continuous loop of being reset, with a partial RIB being held by one or more of the BGP speakers due to an non-deterministic order of UPDATE propagation. It is therefore a requirement that within this reduced-impact procedure any subsequent UPDATE messages that would result in further session resets are ignored. Whilst this results in a condition where an undetermined amount of the RIB is inconsistent, partial reachability is maintained. In this case, the operational toolsets discussed in [Section 6](#) is likely to provide mechanisms by which this condition can be brought to the attention of the relevant operators. This requirement to accept a partial RIB, which results in potential invalid traffic forwarding is a direct result of the deployments of BGP-4, as described in [Section 1.1](#).

The case where NLRI can be determined from an erroneous UPDATE provides further complexities. In this case, a BGP speaker is aware of the sub-set of the RIB which have been identified as being contained within invalid UPDATE messages. This allows a local BGP speaker to re-request single prefixes, utilising a mechanism such as "one-time prefix ORF". However, a similar result is achieved by re-requesting the entire RIB - albeit with greater resource requirements. It is therefore expected that the process of recovery utilises a staged set of mechanisms to attempt to restore consistency of the RIB:

1. Where available, a mechanism capable of requesting only the NLRI determined to have been contained within a invalid UPDATE should be utilised. However, since it is possible that such an error condition can be transient in nature, it is likely that more than one request is to be transmitted (assuming the first does not return a valid UPDATE message). In order to allow a deterministic process, there is a requirement for a limit on the number of specific requests transmitted to be defined.

2. Where a specific refresh mechanism is not available, a peer should re-request the entire RIB. Again, there is a requirement to limit the number of complete RIB requests that should be sent via an implementation, in order to provide a bound both on the expected level of load a device may experience, and on the time for which the RIB may be inconsistent.

3. Finally, a session reset should be performed, as per the reduced-impact NOTIFICATION requirement defined in [Section 5](#). At this point, a similar challenge to that discussed above exists, should the error condition persist. In this case, as defined above, there is a requirement to ignore those UPDATE messages that continue to be erroneous.

It is envisaged that where limits are required, these will be defined on a per memo-basis, or within a further revision of the requirements described herein.

Whilst the approach described above provides a standard means by which error recovery may be handled on a per UPDATE basis, further complexities are raised where multiple errors occur. Clearly, following this procedure causes control-plane load on both the BGP speakers - for this reason, consideration of how repeated use of the mechanisms discussed in this document is required. It is notable that errors may not occur with UPDATE messages relating to only a single NLRI, independent errors in multiple NLRIs may be experienced. For this reason, it is required that an implementation rate limits the number of error handling events sourced towards a particular neighbour. It is expected that such rate limiting, or event suppression is achieved on a per-session basis, where state information is already held, rather than on a per-prefix basis as it is envisaged that such behaviour presents significant scaling problems, and introduces further state requirements for an implementation of the protocol. It is recommended that where a flag indicative of erroneous behaviour is implemented, the state of such a value is maintained independently of session establishment.

8. IANA Considerations

This memo includes no request to IANA.

9. Security Considerations

The requirements outlined in this document provide mechanisms by which erroneous BGP messages may be responded to with limited impact to forwarding operation. This is of benefit to the security of a BGP speaker in general. Where UPDATE messages may have been propagated by a single malicious Autonomous System or router within a network (or the Internet default free zone - DFZ), which are then propagated to all devices within the same routing domain, all other NLRI available over the same session become unreachable. This mechanism may provide means by which an Autonomous System can be isolated from required routing domains (such as the Internet), should the relevant UPDATE messages be propagated via specific paths. By reducing the impact of such failures, it is envisaged that this possibility may be constrained to a specific set of NLRI, or a specific topology.

Some mechanisms meeting the requirements specified in this document, particularly those within Section 6 may provide further security concerns, however, it is envisaged that these are addressed in per-enhancement memos.

10. Acknowledgements

The author would like to thank Shane Amante, Bruno Decraene, Rob Evans, David Freedman, Tom Hodgson, Sven Huster, Jonathan Newton, Neil McRae,

Thomas Mangin, Tom Scholl and Ilya Varlashkin for their review and valuable feedback.

11. References

11.1. Normative References

[RFC2918]	Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, September 2000.
[RFC4271]	Rekhter, Y., Li, T. and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
[RFC4364]	Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
[RFC4456]	Bates, T., Chen, E. and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.
[RFC4724]	Sangli, S., Chen, E., Fernando, R., Scudder, J. and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, January 2007.
[RFC4760]	Bates, T., Chandra, R., Katz, D. and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
[RFC5492]	Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.
[I-D.keyur-bgp-enhanced-route-refresh]	Patel, K, Chen, E and B Venkatachalapathy, "Enhanced Route Refresh Capability for BGP-4", Internet-Draft draft-keyur-bgp-enhanced-route-refresh-02, March 2011.
[I-D.ietf-idr-optional-transitive]	Scudder, J, Chen, E, Mohapatra, P and K Patel, "Revised Error Handling for BGP UPDATE Messages", Internet-Draft draft-ietf-idr-optional-transitive-04, October 2011.
[I-D.chen-ebgp-error-handling]	Chen, E, Mohapatra, P and K Patel, "Revised Error Handling for BGP Updates from External Neighbors", Internet-Draft draft-chen-ebgp-error-handling-01, September 2011.
[I-D.zeng-one-time-prefix-orf]	Zeng, Q and J Dong, "One-time Address-Prefix Based Outbound Route Filter for BGP-4", Internet-Draft draft-zeng-one-time-prefix-orf-01, October 2010.
[I-D.raszuk-bgp-diagnostic-message]	Raszuk, R, Chen, E and B Decraene, "BGP Diagnostic Message", Internet-Draft draft-raszuk-bgp-diagnostic-message-02, March 2011.
[I-D.ietf-idr-advisory]	

	Scholl, T, Scudder, J, Steenbergen, R and D Freedman, "BGP Advisory Message", Internet-Draft draft-ietf-idr-advisory-00, October 2009.
[I-D.ietf-grow-bmp]	Scudder, J, Fernando, R and S Stuart, "BGP Monitoring Protocol", Internet-Draft draft-ietf-grow-bmp-05, December 2010.

11.2. Informational References

	Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, June 2010.
[RFC5881]	

Author's Address

Rob Shakir Shakir Cable&Wireless Worldwide London, UK EMail: rjs@cw.net URI: http://www.cw.com/