Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 13, 2010

J. Scudder
Juniper Networks
C. Appanna
Cisco Systems

July 12, 2009

Multisession BGP
draft-ietf-idr-bgp-multisession-04

Status of this Memo

Copyright Notice

Abstract

This specification augments "Multiprotocol Extensions for BGP-4" (MP-BGP) by proposing a mechanism to facilitate the use of multiple sessions between a given pair of BGP speakers. Each session is used to transport routes related by some session-based attribute such as AFI/SAFI. This provides an alternative to the MP-BGP approach of multiplexing all routes onto a single connection.

Use of this approach is expected to provide finer-grained fault management and isolation as the BGP protocol is used to support more and more diverse services.

Table of Contents

1. Introduction

Most BGP [RFC4271] implementations only permit a single ESTABLISHED connection to exist with each peer. More precisely, they only permit a single ESTABLISHED connection for any given pair of IP endpoints.

BGP Capabilities [RFC5492] extend BGP to allow diverse information (encoded as "capabilities") to be associated with a session. In some cases, a capability may relate to the operation of the protocol machinery; an example is Route Refresh [RFC2918]. However, in other cases a capability may relate specifically to some common distinguishing characteristic of the routes carried over the session; an example is Multiprotocol BGP [RFC4760].

Multiprotocol BGP [RFC4760] extends BGP to allow information for multiple NLRI families and sub-families to be transported in BGP. Routes for different families are distinguished by AFI and SAFI. Routes for different families are commonly multiplexed onto a single BGP session.

A common criticism of BGP is the fact that most malformed messages cause the session to be terminated. While this behavior is necessary for protocol correctness, one may observe that the protocol machinery of a given implementation may only be defective with respect to a given AFI/SAFI. Thus, it would be desirable to allow the session related to that family to be terminated while leaving other AFI/SAFI unaffected. As BGP is commonly deployed, this is not possible.

A second criticism of BGP is that it is difficult or in some cases impossible to manage control plane resource contention when BGP is used to support diverse services over a single session. In contrast, if a single BGP session carries only information for a single service (or related set of services) it may be easier to manage such contention.

In this specification, we propose a mechanism by which multiple transport sessions may be established between a pair of peers. Each transport session is identified by a distinct set of BGP capabilities, notably the MP-BGP capability.

Each session is distinct from a BGP protocol point of view; an error or other event on one session has no implications for any other session. All protocol modifications proposed by this specification take place during the OPEN exchange phase of the session, there are no modifications to the operation of the protocol once a session reaches ESTABLISHED state.

Although AFI/SAFI is perhaps the most obvious way to group sets of

routes being exchanged between BGP peers, sessions can also be distinguished by other BGP capabilities. In general, any capability used in this fashion would be expected to have semantics of identifying some common distinguishing characteristic of a set of routes, just as AFI/SAFI does; however, specifics are beyond the scope of this document. For the sake of clarity, we generally use the MP-BGP capability (or interchangeably, AFI/SAFI) in this document. Such use is illustrative and is not intended to be limiting.

Routers implementing this specification MUST also implement the base criteria that is used to define sessions. For example if AFI/SAFI based sessions are desired then routers implementing this specification MUST also implement MP-BGP [RFC4760].

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Definitions

"MP-BGP capability" refers to the capability [RFC5492] with code 1, specified in MP-BGP [RFC4760] section 8.

A BGP speaker is said to "support" some feature or functionality (for example, to support this specification, or to support a particular AFI/SAFI) when the BGP implementation supports the feature AND the feature has not been disabled by configuration.

The Session Identifier is a capability or group of capabilities that will be used to differentiate individual BGP sessions between two IP endpoints. When the AFI/SAFI is used to distinguish sessions, the MP-BGP capability is the session identifier.

A pair of session identifiers are said to conflict when considering them as two sets, there is an intersection between them either in the capabilities or the values contained within the capabilities, but neither is a subset of the other. For example, a pair of MP-BGP capabilities is said to "conflict" when considering them as two sets (of AFI/SAFI values), there is an intersection between the sets but neither set is a subset of the other.

A BGP speaker is said to be the "active" speaker for a given connection if it was the party that initiated the transport open. The active speaker's transport endpoint will typically use an

ephemeral port number.

A BGP speaker is said to be the "passive" speaker for a given connection if it was the party that received the transport open. The passive speaker's transport endpoint typically uses the well-known BGP port number, 179, but this document introduces an exception detailed in Section 5.1.1.1.

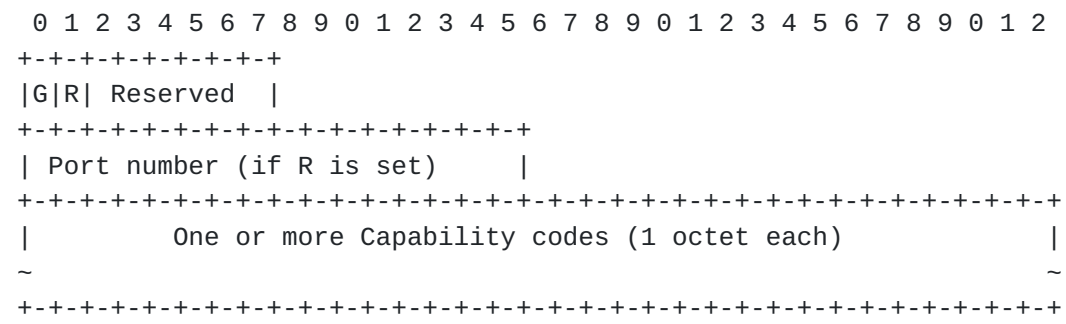
3. Use of BGP Capability Advertisement

This specification defines the Multisession capability [RFC5492]:

Capability code (1 octet): 68

Capability length (1 octet): variable

Capability value (1 octet): Flags followed by the list of capabilities that define a session.



The most significant bit is defined as the Grouping Support (G) bit. It can be used to indicate support for the ability to group multiple capability values into one session. When set (value 1) this bit indicates that the BGP speaker supports grouping. An example of grouping is if a BGP speaker wishes to use one session for AFI/SAFI values 1/1, 1/2 and 1/4, and another for AFI/SAFI values 2/1, 2/2 and 2/4.

The next bit is defined as the Redirect (R) bit. When set, it indicates that the sender wishes to continue the current BGP session using a different transport endpoint. This entails the active speaker dropping the current session and starting a fresh one using the proposed endpoint; this is detailed in Section 5.1.1.1 below. When set, the transport endpoint information is encoded in the port number field of the capability as detailed below.

The remaining bits are reserved, and should be set to zero by the

sender and ignored by the receiver.

If the R bit is set, following the reserved bits is the two-octet TCP port number to which the passive speaker wishes to redirect the session.

Following the reserved bits and the transport endpoint information if present is a list of one or more Capability codes defined in BGP. The size of the list is inferred from the length of the overall capability; it is the capability length minus one if the R bit is not set, or minus three if the R bit is set. The capabilities listed specify which capabilities in the OPEN message comprise the session identifier. The Multisession capability code itself MUST NOT be listed; if listed it MUST be ignored upon receipt.

For example, peers wishing to establish sessions based on AFI/SAFI would exchange the Multiprotocol Extensions capability code (1) only in the list. In this case the Multisession capability would have a length of two octets, or four octets if redirect is being requested.

4. New NOTIFICATION Subcodes

BGP [RFC4271] Section 4.5 provides a number of subcodes to the NOTIFICATION message, and Section 6.2 elaborates on the use of those subcodes.

This specification introduces five new subcodes:

OPEN Message Error subcodes:

7 - Capability Value Mismatch

8 - Grouping Conflict

9 - Grouping Required

10 - Redirecting Now

11 - Redirect Required

The Capability Value Mismatch code MAY be used when an OPEN message received contains one or more capabilities whose values are inconsistent with the corresponding capabilities of the local BGP speaker. The Data field MUST list the offending capability code(s).

The Grouping Conflict code MAY be used when an OPEN message contains one or more capabilities whose values conflict with the values of one

or more capability groups configured on the local BGP speaker. The Data field MUST indicate one of the conflicting locally-configured capability group, encoded as the appropriate capabilities.

The Grouping Required code MAY be used when a BGP speaker that is configured to require grouping attempts to establish a connection with a BGP speaker that does not support grouping. (While it is true that it might be possible to communicate much the same information using the Unsupported Capability NOTIFICATION message, this more explicit method is felt to be more transparent.)

If the MP-BGP capability is used as the session identifier, the notifications could be used as follows:

Capability Value Mismatch MAY be used when an OPEN message contains one or more MP-BGP capabilities, none of which lists an AFI/SAFI supported by the local BGP speaker. It is observed that this subcode may be useful for MP-BGP speakers in general, even if they do not (otherwise) implement this specification.

The Grouping Conflict code MAY be used when an OPEN message contains several MP-BGP capabilities whose AFI/SAFI conflict with one or more AFI/SAFI groups configured on the local BGP speaker. The Data field MUST indicate one of the conflicting locally-configured AFI/SAFI groups, encoded as MP-BGP capabilities. (One might think of this as indicating "I'm not willing to combine AFI/SAFI foo and bar as you've tried to do.")

Use of the Redirecting Now and Redirect Required codes is detailed in Section 5.1.1.1.

The use of these subcodes is further elaborated below.

5. Overview of Operation

The operation section is divided into two main subsections.

The "Using Multisession" sections below discuss the BGP speaker's behavior when the peer does support this specification or is assumed to. The "Backward Compatibility" section discusses the BGP speaker's behavior when the peer does not support this specification, or is assumed not to. Both sections also discuss how to switch to the other mode.

A BGP speaker that supports this specification MUST always advertise the Multisession capability, regardless of its peer's known or presumed capability set.

In all cases until a BGP speaker has initiated or accepted one connection from a given peer, it is unknown whether the peer supports this specification or not. Two strategies can be considered for making this initial determination -- either the BGP speaker can initially assume that the peer does not support this specification, and switch modes if it is discovered that it does, or vice-versa. Either approach is acceptable.

As discussed previously, this section describes the operation from the point of view of the MP-BGP capability and the associated AFI/SAFI values as the session identifier. It can be replaced with any other capability or groups of capabilities without any changes to the behavior described below.

Note that if a BGP speaker only wishes to support a single AFI/SAFI in its communications with a given peer only one session is needed in any case, and so the "multisession" feature is moot. In such a case the behavior required would be indistinguishable from that given in the "backward compatibility" section below. In the illustrative examples in the following sections, it is generally assumed that a BGP speaker does wish to support multiple AFI/SAFI in its communications with a given peer.

5.1. Using Multisession

The following subsections discuss a BGP speaker's behavior towards a peer that is known or assumed to support this specification.

5.1.1. Initiating Connections

When a BGP speaker (the "active" speaker) attempts BGP communication with its peer (the "passive" speaker), it initiates one connection per group of AFI/SAFI it wishes to support. (This implies that a new local TCP port will be allocated for each new connection.) The OPEN sent on each connection MUST include the Multisession capability and one or more MP-BGP capabilities indicating the AFI/SAFI to be supported on that session. If a non-trivial group of AFI/SAFI (i.e., a group of two or more) is proposed, the BGP speaker MUST also set the G bit of the Multisession capability. Even if a trivial group of AFI/SAFI is proposed, the G bit SHOULD be set if grouping is supported. The active speaker MUST NOT set the R bit nor include an associated TCP port number.

Note that any "group of AFI/SAFI" may be a singleton group, i.e. the speaker may wish to use a separate BGP connection for each AFI/SAFI.

If the peer also supports this specification and also wishes to support the AFI/SAFI in question, it will respond with an OPEN that

includes the Multisession capability and the AFI/SAFI included in the active speaker's OPEN. If the active speaker's OPEN included a non-trivial group of AFI/SAFI that the peer supports, then the peer's Multisession capability will have the G bit set.

If the peer also supports this specification and wishes to support some but not all of the AFI/SAFI in question, it will respond with an OPEN that includes the Multisession capability and a subset of AFI/SAFI included in the active speaker's OPEN. The reason for listing only a subset may be because some of the AFI/SAFI are simply not supported, or because the peer does not wish to support the AFI/SAFI as a group (i.e. it may be configured to use a smaller group). In this case, the BGP speaker MAY consider the set of AFI/SAFI that were not included in the peer's OPEN to form a new group, and MAY try to initiate a new session using that group.

If the peer also supports this specification but does not support grouping, and a non-trivial group of AFI/SAFI has been proposed, then it will respond as given in the previous paragraph but with the additional proviso that the G bit will be clear. In this case, the BGP speaker MAY accept the connection as given in the previous paragraph, or it MAY reply with a NOTIFICATION message with ERROR Code OPEN Message Error and Error Subcode Grouping Required, and the connection will be closed.

If the peer wishes to continue the BGP connection on a different transport endpoint, in addition to responding as detailed above, it will set the R bit and will include the TCP port number that should be used to continue the connection. See Section 5.1.1.1 for details regarding how this is handled.

If the peer does not wish to support the AFI/SAFI in question, it will reply with a NOTIFICATION message with Error Code OPEN Message Error, and Error Subcode Capability Value Mismatch, and the connection will be closed.

A BGP speaker MUST NOT attempt to initiate connections for any AFI/SAFI for which a connection already exists.

If the peer does not support this specification, it will respond with an OPEN that does not include the Multisession capability. In this case the connection SHOULD be terminated, and future connections to the peer should be attempted in the "backward compatibility" mode discussed in Section 6.

5.1.1.1. Continuing a Redirected Connection

When the active speaker receives an OPEN from the passive speaker that includes transport redirect information, it MUST reply with an Open Message Error NOTIFICATION with its subcode set to Redirecting Now and close the session. Subsequently, it MUST attempt to initiate a new session using the transport endpoint that the passive speaker has proposed in lieu of the original one (which typically would have been the well-known BGP port, 179). The new session should proceed exactly as the original one did; that is, the active speaker SHOULD send an OPEN with the same content, and can expect to receive from the passive speaker an OPEN with the same content as previously with the exception that the R bit should be clear and no associated port number should be present. If the R bit is not clear it (and the accompanying port number) SHOULD be disregarded.

Note that although the OPEN messages exchanged on the reinitiated session can be expected to be the same as or similar to those from the previous session as discussed above, an implementation MUST NOT rely on or enforce this assumption when handling the received OPEN. The new session MUST be handled as any other new session would be in this respect.

As discussed above, when the passive speaker requests a redirect, the active speaker is expected to drop the current session and initiate a new one. If it does not do so, the passive speaker MAY elect to continue the session, or it MAY elect to terminate the session by sending a Redirect Required NOTIFICATION.

5.1.2. Accepting Connections

When processing a connection attempt, the BGP speaker MUST wait until the peer's OPEN message has been received before proceeding. This is at variance with the behavior specified in the finite state machine (FSM) of [RFC4271], but is interoperable with that FSM. The FSM changes are specified in Section 7.

Once the peer's OPEN message has been received, if it includes the Multisession capability and one or more MP-BGP capabilities indicating a group of AFI/SAFI that the BGP speaker wishes to support, then the BGP speaker responds with an OPEN message that includes the Multisession capability and one or more MP-BGP capabilities indicating the same AFI/SAFI.

If the OPEN includes the Multisession capability and one or more MP-BGP capabilities indicating a group of AFI/SAFI that conflicts with an AFI/SAFI grouping that has been configured on the BGP speaker then the BGP speaker MAY reply with an OPEN listing a set of AFI/SAFI that

intersect with those proposed by the peer (in effect overriding the locally configured set) or it MAY close the connection with a NOTIFICATION message with Error Code OPEN Message Error and Error Subcode Grouping Conflict. The former behavior is suggested as the default if grouping is supported.

If the BGP speaker does not support AFI/SAFI grouping it MAY reply with an OPEN listing one of the AFI/SAFI out of those proposed by the peer. It MUST also set the G bit in the Multisession capability to zero.

If the passive speaker wishes to continue the session for this particular grouping on a different port number, it sets the R bit in its OPEN and includes the TCP port number on which it will continue the session. The passive speaker MUST be prepared to accept a connection on the given port immediately following transmission of its OPEN.

If the received OPEN message does not include any MP-BGP capability indicating an AFI/SAFI the BGP speaker wishes to support, it SHOULD close the connection with a NOTIFICATION message with Error Code OPEN Message Error and Error Subcode Capability Value Mismatch.

If the received OPEN message does not include the Multisession capability, then the peer does not support this specification. The connection MAY be continued in the "backward compatibility" mode discussed in Section 6, or it MAY be terminated and future connections to the peer attempted in the "backward compatibility" mode.

5.1.3. Collision Detection, Graceful Restart

[RFC4271] Section 6.8 (BGP connection collision detection) considers a pair of connections to have collided if the source and destination IP addresses of both connections match. With respect to peers that support this specification, the AFI/SAFI groups associated with the connections must also intersect for them to be considered to have collided.

This consideration also applies to Section 4.2 of BGP Graceful Restart [RFC4724], when determining whether a new connection should be considered equivalent to a reset of a previous TCP session.

6. Backward Compatibility

This subsection discusses a BGP speaker's behavior towards a peer that is known or assumed not to support this specification. In

short, the BGP speaker's behavior towards such a peer should be as otherwise defined for the BGP protocol, according to [RFC4271] and any other extension supported by the BGP speaker.

As previously mentioned, the BGP speaker SHOULD always advertise the Multisession capability in its OPEN message, even towards "backward compatibility" peers.

If, in opening a BGP connection with such a peer, an OPEN that includes the Multisession capability is received from the peer, then the peer SHOULD be changed to "multisession" mode. How this is done depends on whether the BGP speaker has already sent an OPEN or not --

If the BGP speaker has not yet sent an OPEN to the peer, then the connection MAY be continued in the "multisession" mode discussed above, or it MAY be terminated and future connections to the peer attempted in "multisession" mode.

If the BGP speaker has sent an OPEN to the peer, then the current session SHOULD be terminated and future connections to the peer attempted in "multisession" mode.

Use of techniques such as dynamic capabilities [I-D.ietf-idr-dynamic-cap] for on-the-fly switching of session modes is beyond the scope of this document.

7. State Machine

As mentioned under "accepting connections" above, this specification modifies the BGP finite state machine, albeit in a backward-compatible fashion.

In addition, note that one state machine is considered to exist for each of the connections that may exist to a given peer. This implies that, for example, any session flap dampening that may exist is performed per session identifier.

The specific state machine modifications to [RFC4271] Section 8.2.2 are as follows.

7.1. Modifications to Connect State and Active State

In the actions in response to the events Open Delay timer expires [Event 12] and TCP connection succeeds [Event 16 or Event 17], an OPEN is not sent and the state changes to WaitForOpen and not to OpenSent.

7.2. Addition of WaitForOpen State, Deletion of OpenSent State

The WaitForOpen state is the same in all respects to OpenSent, except for the action in response to reception of a valid OPEN message [Event 19]. In that event, the local system sends an OPEN message prior to sending a KEEPALIVE message.

The OpenSent state is deleted. All references to OpenSent are replaced by references to WaitForOpen.

8. Discussion

Note that many BGP implementations already permit multiple sessions to be used between a given pair of routers, typically by configuring multiple IP addresses on each router and configuring each session to be bound to a different IP address. The principal contribution of this specification is to allow multiple sessions to be created automatically, without additional configuration overhead or address consumption.

The specification supports the simple case of one capability being used as the session identifier and one connection per session identifier value. It also permits connections be established based on multiple capabilities as a session identifier with multiple values per capability grouped together per connection.

In the context of MP-BGP based connections, which we believe may be the most prevalent use of this specification, it permits supporting one AFI/SAFI per connection, and also permits arbitrary grouping of AFI/SAFI onto BGP connections. For such grouping to function pleasingly, both peers participating in a connection need to agree on what AFI/SAFI groupings will be used. If conflicting groupings are configured, the connections may not establish, or more connections may be established than were expected (in the degenerate case, one connection per AFI/SAFI could be established despite configured groupings). We observe that the potential for misbehavior in the presence of conflicting configuration is not unusual in BGP, and that support for, and configuration of grouping is purely optional.

9. Security Considerations

The ability to redirect to a port other than the well-known BGP port implies that a legitimate BGP session may exist for which neither port is equal to 179. This may have implications for firewall filters used to protect the control processor.

In other respects, this document does not change the BGP security model.

10. Acknowledgements

The authors would like to thank Pedro Marques, Keyur Patel, Robert Raszuk, Yakov Rekhter and David Ward for their valuable comments.

11. IANA Considerations

IANA has allocated BGP Capability Code 68 as the Multisession BGP Capability.

This document requests IANA to allocate five new OPEN Message Error subcodes:

7 - Capability Value Mismatch

8 - Grouping Conflict

9 - Grouping Required

10 - Redirecting Now

11 - Redirect Required

12. References

12.1. 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.

[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.

12.2. Informative References

[I-D.ietf-idr-dynamic-cap]
Chen, E. and S. Sangli, "Dynamic Capability for BGP-4", draft-ietf-idr-dynamic-cap-09 (work in progress), November 2006.

[RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, September 2000.

Authors' Addresses

John G. Scudder
Juniper Networks

Email: jgs@juniper.net

Chandra Appanna
Cisco Systems

Email: achandra@cisco.com