Network Working Group
Internet Draft
Expiration Date: May 2004
File name: draft-scudder-bgp-multisession-00.txt

John G. Scudder
Chandra Appanna
Cisco Systems
November 2003

Multisession BGP
draft-scudder-bgp-multisession-00.txt

Status of this Memo

Abstract

This specification augments "Multiprotocol Extensions for BGP-4" [MP-BGP] by proposing a mechanism to allow multiple sessions to be used between a given pair of BGP speakers. Each session is used to transport routes for one or more AFI/SAFI. This provides an alternative to the current [MP-BGP] approach of multiplexing routes for all AFI/SAFI onto a single connection.

Use of this approach is expected to increase the robustness of the BGP protocol as it is used to support more and more diverse AFI/SAFI.

1. Introduction

Most BGP [BGP, BGP-DRAFT] 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.

Multiprotocol BGP [MP-BGP] 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.

In this specification, we propose a mechanism by which multiple transport sessions may be established between a pair of peers. Each transport session can be used for one or more AFI/SAFI. 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.

Routers implementing this specification MUST also implement [MP-BGP].

2. Definitions

"MP-BGP capability" refers to the capability [BGP-CAP] with code 1, specified in [MP-BGP] section 10.

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.

A pair of AFI/SAFI groups is said to "conflict" when considering the two groups as two sets, there is an intersection between the groups but neither group is a subset of the other.

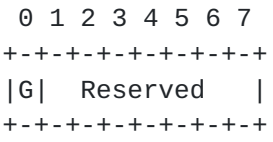
3. Use of BGP Capability Advertisement

This specification defines the Multisession capability [BGP-CAP]:

Capability code (1 octet): TBD

Capability length (1 octet): 1

Capability value (1 octet): Flags as below



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 AFI/SAFI into one session. When set (value 1) this bit indicates that the BGP speaker supports grouping.

The remaining bits are reserved, and should be set to zero by the sender and ignored by the receiver.

4. New NOTIFICATION Subcodes

[BGP, BGP-DRAFT] 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 two new subcodes:

OPEN Message Error subcodes:

7 - No Supported AFI/SAFI.

8 - Grouping Conflict

9 - Grouping Required

The No Supported AFI/SAFI code MAY be used when an OPEN message contains one or more MP-BGP capabilities, none of which list 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 SHOULD indicate one of the conflicting locally-configured AFI/SAFI groups, encoded as MP-BGP capabilities.

The Grouping Required code MAY be used when a BGP speaker which is configured to require grouping attempts to establish a connection with a BGP speaker which 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.)

The use of these subcodes is further elaborated below.

5. Overview of Operation

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.

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 discuss how to switch to the other mode.

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

5.1. Using Multisession:

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

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 following sections, it is generally assumed that a BGP speaker does wish to support multiple AFI/SAFI in its communications with a given peer.

5.1.1. Initiating Connections:

When a BGP speaker attempts BGP communication with its peer, 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.

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 which 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 which 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 which 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 which 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 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 No Supported AFI/SAFI, and the connection will be closed.

A BGP speaker SHOULD 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 which 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 below.

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 [BGP-DRAFT], but is interoperable with that FSM. The FSM changes are specified in a later section.

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 which the BGP speaker wishes to support, then the BGP speaker responds with an OPEN message which 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 which 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 which 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 SHOULD also set the G bit in the Multisession capability to zero.

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 No Supported AFI/SAFI.

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 below, or it MAY be terminated and future connections to the peer attempted in the "backward compatibility" mode.

5.1.3. Collision Detection, Graceful Restart:

[BGP, BGP-DRAFT] 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 which 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 6.2 of [BGP-GR], when determining whether a new connection should be considered equivalent to a reset of a previous TCP session.

5.2. Backward Compatibility:

This subsection discusses a BGP speaker's behavior towards a peer which 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 [BGP, BGP-DRAFT] 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 which 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 [BGP-DYN-CAP] for on-the-fly switching of session modes are beyond the scope of this document.

6. 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 which may exist to a given peer. This implies that, for example, any session flap dampening that may exist is performed per AFI/SAFI.

The specific state machine modifications to [BGP-DRAFT] Section 8.2.2 are as follows.

6.1. 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.

6.2. 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.

7. 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.

In addition to the simple mode of supporting one AFI/SAFI per connection, the procedures described here also permit 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.

8. Acknowledgements

To be supplied.

9. References

[BGP4]
Rekhter, Y. and T. Li, "A Border Gateway Protocol 4 (BGP-4)," RFC 1771, March 1995.

[BGP-DRAFT]
Rekhter, Y., T. Li and S. Hares, "A Border Gateway Protocol 4 (BGP-4)," Work in Progress (draft-ietf-idr-bgp4-20), April 2003.

[MP-BGP]
Bates, T., R. Chandra, D. Katz, Y. Rekhter, "Multiprotocol Extensions for BGP-4," Work in Progress (draft-ietf-idr-rfc2858bis-03), July 2003.

[BGP-GR]
Sangli, S., Y. Rekhter, R. Fernando, J. Scudder, E. Chen, "Graceful Restart Mechanism for BGP," Work in Progress (draft-ietf-idr-restart-06), January 2003.

[BGP-CAP]
Chandra, R., J. Scudder, "Capabilities Advertisement with BGP-4," RFC 2842, May 2000.

[BGP-DYN-CAP]
Chen, E. and S. Sangli, "Dynamic Capability for BGP-4," Work in Progress (draft-ietf-idr-dynamic-cap-03), December 2002.

10. Security Considerations

This document introduces no new security vulnerabilities to BGP or other specifications referenced in this document.

11. IANA Considerations

TBD

12. Authors' Addresses

John G. Scudder
Cisco Systems, Inc.
100 S. Main Suite 200
Ann Arbor, MI 48104
Email: jgs@cisco.com

Chandra Appanna
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
e-mail: achandra@cisco.com

13. Full Copyright Statement

TASK FORCE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.