Workgroup: IDR Working Group
Published: 7 March 2023
Intended Status: Informational
Expires: 8 September 2023
Authors: E. Chen
Palo Alto Networks
R. Raszuk
Arrcus

Applying TCP User Timeout Parameter to BGP Sessions

Abstract

In this document we discuss the TCP "User Timeout" parameter and recommend using it to handle stuck BGP sessions.

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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 https://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 8 September 2023.

Copyright Notice

document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction

A BGP session [RFC4271] is said, informally, to be "stuck" when BGP messages are not transmitted over the session for an extended period of time. Certainly the stuck BGP session should have been terminated by the BGP holdtimer. Such a case could occur, though, due to software defects or under certain unusual circumstances. Currently it's difficult to know for sure due to lacking of automated, real-time detection mechanisms in BGP implementations.

It has been speculated that some BGP sessions may have stuck from time to time, and that may have contributed to stale routes (e.g., missing route withdraws) in the routing system.

Here is a specific scenario of a stuck BGP session between two BGP speakers A and B:

- Due to certain software defect, B stops reading data from TCP [RFC0793] for an extended period of time, resulting in B advertises a zero value for its TCP window size after its TCP buffer fills up.

- B fails to generate BGP HOLDTIME expiration, although it has not read from TCP and thus has not received any BGP KEEPALIVE from A during that time.

- B, however, continues to send BGP KEEPALIVE to A on time.

In this scenario A would not be able to send routing updates to B during that period of time. The routing system may become stale, not only on B, but on its BGP neighbors and beyond.

It's desirable for a BGP speaker (e.g., A in the example) to be able to detect and then terminate such a stuck session so that the stale routes are purged from the routing system.

The availability of such a mechanism may also help accelerate the resolution of the software defect involved.

In this document we discuss the TCP "User Timeout" parameter [RFC0793] and recommend using it to handle stuck BGP sessions.

2. Discussion on TCP User Timeout

The TCP "User Timeout" parameter is designed to terminate a connection in a variety of cases where a TCP session does not progress within certain time period. It is specified in [RFC0793] as follows:

> USER TIMEOUT
>
> For any state if the user timeout expires, flush all queues, signal the user "error: connection aborted due to user timeout" in general and for any outstanding calls, delete the TCB, enter the CLOSED state and return.

Clearly the TCP "User Timeout" applies when the application data is not delivered on time, including the cases that transmitted data may remain unacknowledged, or buffered data may remain untransmitted (due to zero window size).

The TCP "User Timeout" parameter is well summarized in [RFC5482], although the zero-window case is not explicitly called out:

> The Transmission Control Protocol (TCP) specification RFC0793 defines a local, per-connection "user timeout" parameter that specifies the maximum amount of time that transmitted data may remain unacknowledged before TCP will forcefully close the corresponding connection. Applications can set and change this parameter with OPEN and SEND calls.

Regarding the implementation of the TCP "User Timeout" parameter, one example is Linux's "TCP_USER_TIMEOUT" socket option documented in [LINUX-TCP].

3. Recommendations

As discussed in the introduction, a BGP session is considered "stuck" when BGP messages are not delivered for an extended period of time.

Given that BGP messages are TCP data, and TCP is responsible for delivering the data, thus it would be more natural and more complete to address the issue at the TCP layer rather than in BGP itself (particularly in the case of persistent TCP zero-window).

As the TCP "User Timeout" parameter is specifically defined to terminate the TCP connection when something in TCP is "stuck", we thus recommend using it to detect and terminate these stuck BGP sessions.

We RECOMMEND that the TCP "User Timeout" parameter be set for all BGP sessions, and the default timeout value be five times the configured BGP holdtime value but no less than ten minutes in order to tolerate certain short-lived, transient conditions. The TCP "User Timeout" value for a BGP session SHOULD be configurable.

We also RECOMMEND that the TCP "User Timeout" parameter be set only after the End-of-RIB marker [RFC4724], if expected, is received from each of the (AFI, SAFI) being exchanged over the BGP session, or otherwise thirty minutes after the BGP session is established. The delay for setting the parameter SHOULD be configurable.

When the TCP "User Timeout" for a BGP session expires, the BGP speaker SHOULD log the event locally. In addition, the administrator of the remote BGP speaker SHOULD be informed (by means outside the scope of this document) so that the issue can be investigated.

The procedures for BGP Graceful Restart [RFC4724] SHOULD be followed when the TCP session is terminated due to TCP "User Timeout" expiration.

4. IANA Considerations

This document has no request for IANA.

5. Security Considerations

The solution recommended in this document does not change the underlying security or confidentiality issues inherent in the existing BGP [RFC4271].

6. Acknowledgments

TBD

7. References

7.1. Normative References

[RFC0793]

Postel, J., "Transmission Control Protocol", RFC 793, DOI 10.17487/RFC0793, September 1981, <https://www.rfc-editor.org/info/rfc793>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <https://www.rfc-editor.org/info/rfc4724>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

7.2. Informative References

[LINUX-TCP] TCP(7), "Linux Man Pages", March 2021, <https://man7.org/linux/man-pages/man7/tcp.7.html>.

[RFC5482] Eggert, L. and F. Gont, "TCP User Timeout Option", RFC 5482, DOI 10.17487/RFC5482, March 2009, <https://www.rfc-editor.org/info/rfc5482>.

Authors' Addresses

Enke Chen
Palo Alto Networks

Email: enchen@paloaltonetworks.com

Robert Raszuk
Arrcus

Email: robert@raszuk.net