DCCP Working Group
Internet-Draft
Updates: 4342, 5622
(if approved)
Intended status: Standards Track
Expires: February 16, 2011

G. Renker
G. Fairhurst
University of Aberdeen
August 18, 2010

Sender RTT Estimate Option for DCCP
draft-renker-dccp-tfrc-rtt-option-01

Abstract

This document describes an update to CCID-3/4 that addresses parameter-estimation problems occurring with TFRC-based DCCP congestion control.

The fix uses a recommendation made in the original TFRC specification. It avoids the inherent problems of receiver-based RTT sampling, by utilising higher-accuracy RTT samples already available at the sender. It is integrated into the feature set of DCCP as an end-to-end negotiable extension, upward and downward compatible.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

This document lists and analyses problems observed with receiver-based RTT sampling in the actual implementation of TFRC congestion control [RFC4342], [RFC5622].

To fix these problems, this document presents a solution based on a concept first recommended in [RFC5348], 3.2.1; i.e. to measure the RTT at the sender. This results in a higher reliability and frequency of samples, and avoids the inherent problems of receiver-based RTT sampling discussed below.

We begin by listing the encountered problems in the next section. The proposed solution is presented in in Section 3. We then discuss security considerations in Section 4 and list the resulting IANA considerations in Section 5.

2. Problems caused by sampling the RTT at the receiver

There are at least six areas that make a TFRC receiver vulnerable to inaccuracies or absence of (receiver-based) RTT samples:

- the measured sending rate, X_recv ([RFC5348], 6.2);
- synthesis of the first loss interval ([RFC5348], 6.3.1);
- disambiguation of loss events ([RFC4342], 10.2);
- validation of loss intervals ([RFC4342], 6.1);
- ensuring that at least one feedback packet is sent per RTT ([RFC4342], 10.3);
- determining quiescence periods ([RFC4342], 6.4).

2.1. List of problems encountered with a real implementation

This section summarizes several years of experience using the Linux implementation of CCID-3 and CCID-4. It lists the problems encountered with receiver-based RTT sampling over real networks, in a variety of wired and wireless environments and under different link-layer conditions.

The Linux DCCP/TFRC implementation is based on the RTT-sampling algorithm specified in [RFC4342], 8.1. This algorithm relies on a coarse-grained window-counter (units of RTT/4), and uses packet inter-arrival times to estimate the current RTT of the network.

The algorithm is effective only for packets with modulo-16 CCVal differences between 2 and 4 (corresponding to RTT/2, 3/4RTT, and RTT). This limitation is noted in sections 8.1 and 10.3 of [RFC4342].

A second problem arises when there are holes in the sequence space. Because there may be wrap-around of the 4-bit CCVal window counter, it is not possible to determine window-counter wrap-around whenever sequence numbers of subsequent packets are not immediately adjacent. This problem occurs when packets are delayed, reordered, or lost in the network.

As a consequence, RTT sampling has to be paused during times of loss. This however aggravates the problem, since the sender now requires new feedback from the receiver, but the receiver is unable to provide accurate and up-to-date information: the receiver is unable to sample the RTT, accordingly also not able to estimate X_recv correctly,

which then in turn affects X_Bps at the sender.

The third limitation arises from using inter-arrival times as representatives of network inter-packet gaps. It is well known that the inter-packet gap is not constant along a network path. Furthermore, modern network interface cards do not necessarily deliver each packet at the time it is received, but rather in a bunch, to avoid overly frequent interrupts [MR97]. As a result, inter-packet arrival times may converge to zero, when subsequent packets are delivered at virtually the same time, served by the same interrupt routine.

The fourth problem is that of under-sampling and thus related to the first limitation. If loss occurs while the receiver has not yet had a chance to sample the RTT, it needs to fall back to some fixed RTT constant to plug into the equation of [RFC5348], 6.3.1. (The sender, for example, uses a fixed value of 1 second when it can not obtain an initial RTT sample, compare [RFC5348], 4.2).

In particular, if the loss is caused by a transient condition, this fourth problem causes a subsequent deterioration of the connection (rate reduction), further aggravated by the fact that TFRC takes longer than common window-based protocols to recover from a reduction of its allowed sending rate.

The fifth and last problem is starvation under burst loss, caused for instance by a sudden interference in a wireless transmission. The resulting burst loss sets off a vicious circle, where link-layer retransmissions and transmitter-backoff procedures and/or reverse-path loss eventually cause the nofeedback timer to be triggered at the sender. This in turn halves the sending rate, thereby doubling the inter-packet gap. Which in turn decreases X_recv sampled via RTT at the receiver. These factors contribute to an accelerated reduction of the sending rate towards zero, or rather 1 packet per 64 seconds (t_mbi). Under these conditions the connection is no longer in a usable state, unless buffering of more than 64 seconds (more is required because the sending rate is low) can be applied, which is impossible for interactive applications, and unacceptable for many audio/video applications.

Trying to smooth over these effects by imposing heavy filtering on the RTT samples did not substantially improve the situation, nor does it solve the problem of under-sampling.

We are not aware of an alternative (published) algorithm to better estimate the RTT at the receiver.

The TFRC sender, on the other hand, is much better equipped to

estimate the RTT and can do this more accurately. This is in particular due to the use of timestamps and elapsed time information ([RFC5348], 3.2.2), which are mandatory in CCID-3 (sections 6 and 8.2 of [RFC4342]).

2.2. Other areas affected by the RTT sampling problems

We here analyse the impact that unreliability of receiver-based RTT sampling has on the areas listed at the begin of this section.

In addition, benefits of sender-based RTT sampling have already been pointed out in [RFC5348], and in the specification of CCID-3 [RFC4342], at the end of section 10.2.

2.2.1. Measured Receive Rate X_recv

A key problem is that the reliability of X_recv [RFC4342] depends directly upon the reliability and accuracy of RTT samples. This means that failures propagate from one parameter to another.

Errata IDs 610 and 611 update [RFC4342] to use the definition of the receive rate as specified in [RFC5348].

Having an explicit (rather than a coarse-grained) RTT estimate allows measurement of X_recv with greater accuracy, and isolates failure.

An explicit RTT estimate also enables the receiver to more accurately perform the test in step (2) of [RFC4342], 6.2, i.e. to check whether less or more than one RTT has passed since the last feedback.

2.2.2. Disambiguation and Accuracy of Loss Intervals

Since a loss event is defined as one or more lost (ECN-marked) data packets in one RTT ([RFC5348], 5.2), the receiver needs accurate RTT estimates to validate and accurately separate loss events. Moreover, [RFC5348], 5.2 expressly points out the sender RTT estimate as RECOMMENDED for this purpose.

Having the sender RTT Estimate available further increases the accuracy of the information reported by the receiver. The definition of Loss Intervals in [RFC4342], 6.1 needs the RTT to separate the lossy parts; in particular, lossy parts spanning a period of more than one RTT are invalid.

A similar benefit arises in the computation of the loss event rate: as discussed in section 9.2 of [RFC4342], it may happen that sender and receiver compute different loss event rates, due to differences in the available timing information. An explicit RTT estimate

increases the accuracy of information available at the receiver, thus the sender may not need to recompute the (less reliable) loss event rate reported by the receiver.

2.2.3. Determining Quiescence

The quiescence period is defined as max(2 * RTT, 0.2 sec) in section 6.4 of [RFC4342]. An explicit RTT estimate avoids under- and over-estimating quiescence periods.

2.2.4. Practical Considerations

Using explicit RTT estimates contributes to greater robustness and can also result in simpler implementation:

First, it becomes easier to separate adjacent loss events. The 4-bit counter value wraps relatively frequently, which requires complex computations to avoid aliasing effects.

Second, the receiver is better able to determine when to send feedback packets. It can perform the test described in step (2) of [RFC5348], 6.2 more accurately. Moreover, unnecessary expiration of the nofeedback timer (as described in [RFC4342], 10.3) can be avoided.

Lastly, a sender-based RTT estimate option can be used by middleboxes for verification [RFC4342], 10.2.

3. Specification

3.1. Conventions

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 [RFC2119].

This document uses the conventions of [RFC5348], [RFC4340], [RFC4342], and [RFC5622].

3.2. Options and Features

This document defines a single TFRC-specific option, RTT Estimate, described in the next subsection.

Following the guidelines in [RFC4340], section 15, the use of the RTT Estimate option is governed by an associated feature, Send RTT Estimate. This feature is described in the second subsection.

3.2.1. RTT Estimate Option

The sender communicates its current RTT estimate to the receiver using a RTT Estimate option.

==> RFC Editor's Note:

Please replace 'XX' with IANA value when published and delete this note.

Type	Option Length	Meaning	DCCP Data?
XX	6	RTT Estimate	Y

Table 1: The RTT Estimate option defined by this document

Column meanings are as per [RFC4340], section 5.8 (table 3). This option is permitted in any DCCP packet, has option number XX and a length of 6 bytes.

```
+--------+--------+--------+--------+--------+--------+
|xxxxxxxx|00000110|        Sender RTT Estimate        |
+--------+--------+--------+--------+--------+--------+
 Type=XX  Length=6
```

The four bytes of option data carry the current RTT estimate of the sender, using a granularity of 1 microsecond (senders sampling with a lower resolution can multiply their RTT estimates to achieve this granularity).

A value of zero indicates that the sender does not have a valid RTT sample yet.

Senders SHOULD send long-term RTT estimates (sampled over a longer period of time) rather than instantaneous RTT samples.

3.2.2. Send RTT Estimate Feature

The Send RTT Estimate feature lets endpoints negotiate whether the sender MUST provide RTT Estimate options on its data packets.

==> RFC Editor's Note:

Please replace 'YY' with IANA value when published and delete this note.

Send RTT Estimate has feature number YY and is server-priority. It takes one-byte Boolean values. Values greater than 1 are invalid and MUST be ignored.

Number	Meaning	Rec'n Rule	Initial Value	Req'd
YY	Send RTT Estimate	SP	0	N

Table 2: The Send RTT Estimate feature defined by this document

The column meanings are described in [RFC4340], section 6.4. In particular, the feature is by default off (initial value of 0), and the extension is not required to be understood by every DCCP implementation (cf. [RFC4340], section 15).

DCCP B sends a "Change R(Send RTT Estimate, 1)" to ask DCCP A to send RTT Estimate options as part of its data traffic.

3.3. Usage

When the Send RTT Estimate Feature is enabled, the sender MUST provide an RTT Estimate Option on all of its Data, DataAck, Sync, and SyncAck packets. It MAY in addition provide the RTT Estimate Option on other packet types, such as DCCP-Ack.

When the receiver has requested the use of the RTT Estimate Option, it MUST use the RTT value reported by that option in all places that require a RTT (listed at the begin of Section 2), and MUST NOT estimate the RTT based on CCVal window counter values. The receiver MAY keep a moving-average of these sender-based RTT estimates, in the manner of [RFC5348], section 4.3.

When the Send RTT Estimate is disabled, the sender MUST NOT send RTT Estimate options on any of its packets, the receiver MUST ignore the RTT Estimate option on all incoming packets, and MUST try to estimate the RTT in some other way (not specified by this document).

The sender MUST implement and continue to update CCVal window counter RTT values as specified in [RFC4342], section 8.1, even when the Send RTT Estimate Feature is on.

4. Security Considerations

Security considerations for CCID-3 have been discussed in section 11 of [RFC4342]; for CCID-4 these have been discussed in section 13 of [RFC5622], referring back to the same section of [RFC4342].

This document introduces an extension to communicate the current RTT estimate of the sender to the receiver of a TFRC communication.

By altering the value of the RTT Estimate option, it is possible to interfere with the behaviour of the flow. In particular, since accuracy of the RTT estimate directly influences the accuracy of the measured sending rate X_recv, it would be possible to obtain either higher or lower sending rates than are warranted by the current network conditions.

This is only possible if an attacker is on the same path as the DCCP sender and receiver, and is able to guess valid sequence numbers. Therefore the considerations in section 18 of [RFC4340] apply.

5. IANA Considerations

This document requests identical allocation in the dccp-ccid3-parameters and the dccp-ccid4-parameters registries.

5.1. Option Types

This document defines a single CCID-specific option for communicating RTT estimates from the HC-sender to the HC-receiver. Following [RFC4340], 10.3, this requires an option number for the RTT Estimate option in the range 128...191.

Note to IANA and the RFC editor

When the IANA has allocated an option number for the `RTT Estimate' option, please replace all occurrences of the placeholder `XX' in this text with that number and delete this note.

(Due to [RFC4340], 19.3 and [RFC4342], 12.2, the option number would be allocated in the range 128...183/191.)

5.2. Feature Numbers

This document defines a single CCID-specific feature number for the Send RTT Estimate feature which is located at the HC-sender. Following [RFC4340], 10.3, a feature number in the range 128...191 is required.

Note to IANA and the RFC editor

When the IANA has allocated an option number for the `Send RTT Estimate' feature, please replace all occurrences of the placeholder `YY' in this text with that number and delete this note.

(Due to [RFC4340], 19.4 and [RFC4342], 12.3, the feature number would be allocated in the range 128...183/191.)

6. References

6.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4340] Kohler, E., Handley, M., and S. Floyd, "Datagram Congestion Control Protocol (DCCP)", RFC 4340, March 2006.

[RFC4342] Floyd, S., Kohler, E., and J. Padhye, "Profile for Datagram Congestion Control Protocol (DCCP) Congestion Control ID 3: TCP-Friendly Rate Control (TFRC)", RFC 4342, March 2006.

[RFC5348] Floyd, S., Handley, M., Padhye, J., and J. Widmer, "TCP Friendly Rate Control (TFRC): Protocol Specification", RFC 5348, September 2008.

[RFC5622] Floyd, S. and E. Kohler, "Profile for Datagram Congestion Control Protocol (DCCP) Congestion ID 4: TCP-Friendly Rate Control for Small Packets (TFRC-SP)", RFC 5622, August 2009.

6.2. Informative References

[MR97] Mogul, J. and K. Ramakrishnan, "Eliminating Receive Livelock in an Interrupt-Driven Kernel", ACM Transactions on Computer Systems (TOCS), 15(3):217-252, August 1997.

Note to the RFC Editor:

Please remove the following Change Log when published, and delete this note.

Appendix A. Change Log

This document is a rewrite of Revision 00. The wording has changed, and as a result of more experience with CCID-3/4, the list of problems has been added to. The specification itself remains unchanged from Revision 00.

Authors' Addresses

Gerrit Renker
University of Aberdeen
Department of Engineering
Fraser Noble Building
Aberdeen AB24 3UE
Scotland

Email: gerrit@erg.abdn.ac.uk
URI: http://www.erg.abdn.ac.uk

Godred Fairhurst
University of Aberdeen
Department of Engineering
Fraser Noble Building
Aberdeen AB24 3UE
Scotland

Email: gorry@erg.abdn.ac.uk
URI: http://www.erg.abdn.ac.uk