Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: November 15, 2018

D. Jovev
M. Proshin
Ericsson
May 14, 2018

Determining SCTP's Retransmission Timer
draft-jovev-tsvwg-sctp-rto-02

Abstract

This document defines a modification in the RFC 4960 [RFC4960] defined Stream Control Transmission Protocol's (SCTP's) Retransmission Timer (RTO) calculation method.

The modification is aimed to reduce the frequency of spurious T3 timeouts, which are caused by underestimated RTO values, derived by the [RFC4960] defend RTO calculation method. The proposed modification aligns the RTO calculation method with the characteristics of the statistical estimator algorithms, which are used for SRTT and RTTVAR calculation, the SCTP protocol data transfer rules and the characteristics of the data packets' arrival pattern in the telecom signalling networks.

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 November 15, 2018.

Table of Contents

1. Introduction

Like TCP, the SCTP's reliable transfer of data is ensured by limiting the time in which the acknowledgement for the reception of the transmitted data is received, after which expiration all unacknowledged data is retransmitted. The duration of this timer is referred to as Retransmission Timeout (RTO) and the actual timer is called T3-rtx or just T3.

The expiration of the T3 timer not only invokes retransmission of the unacknowledged data it also drastically reduces the congestion window (cwnd) to 1 MTU, which are both undesirable actions: data retransmission increases the amount of sent data in the network, and 1 MTU cwnd drastically reduces the SCTP association transmission capacity. Because of that, determining an RTO value which reflects the highest RTT, or the highest feedback time, as more appropriately called in [ALLMAN99], is critical for reducing the probability of spurious T3 timeouts, which is critically important for stable SCTP operation.

Namely, while in the conventional file transfer applications the transport layer transmission capacity reduction, due to T3 timeouts, only prolongs the time for completion of the file transfer, in the telecom signalling networks it often results in false congestion i.e., congestion caused by SCTP transmission capacity reduction not

by traffic increase, which can lead to unrepairable loss of data that adversely affects the services provided by the telecom networks.

This document defines a modification in the [RFC4960] defined SCTP's Retransmission Timer (RTO) calculation method. The modification is aimed to reduce the frequency of spurious T3 timeouts, which are caused by underestimated RTO values, by adjusting the RTO calculation method to the characteristics of the statistical estimator algorithms, which are used for SRTT and RTTVAR calculation, and to the SCTP protocol data transfer rules and the characteristics of the data packets' arrival pattern in the telecom signalling networks.

The modified RTO calculation affects only the sender side and it does not require introduction of new protocol variables or parameters nor change of the [RFC4960] recommended values for the existing RTO related protocol parameters.

The motivations for the modification in the [RFC4960] algorithm for RTO calculation are outlined in Section 2. The actual modification in the [RFC4960] algorithm for RTO calculation is specified in Section 3 whereas the technical background for the modification is elaborated in the Appendix A.

1.1. Conventions and Terminology

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. Problem description

The [RFC4960] defined process for RTO determination consists of two steps.

In the first step, using RTT measurements as input data, a calculated RTO value is derived from the mean/smooth RTT (SRTT) and RTT variation (RTTVAR) values, which are determined using a statistical estimator algorithm, originally published in [JAC88], and then, in the second step, the used RTO is determined as:

```
RTO <- min(RTO.Max, max(calculated RTO, RTO.Min)),
```

where RTO.Min and RTO.Max are configurable protocol parameters with [RFC4960] recommended values of 1 sec and 60 seconds.

By applying the [RFC4960] RTO calculation rules, the RTO value that will be used for the T3 timer will be:

* The value of the RTO.Min - if the calculated RTO is below RTO.Min.

* The calculated RTO - if the calculated RTO is above RTO.Min but below RTO.Max.

* The value of the RTO.Max - if the calculated RTO is above RTO.Max.

Diagram in Figure 1 illustrates the outcome of the above RTO determination rules.

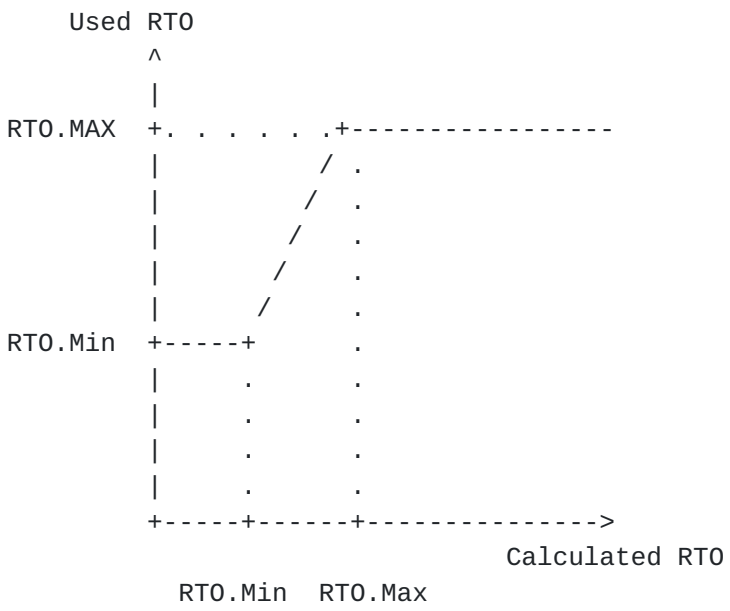

Figure 1: Relation between the calculated and used RTO values

The SCTP protocol has been operating in the telecom networks for more than fifteen years and spurious T3 timeouts have been one of the most frequently reported problems.

The results of the analysis of the spurious T3 timeouts problems, reported from the operating networks, indicated that the spurious T3 timeouts frequency increases when the SRTT value is closer to the RTO.Min value to the point where the association becomes unstable if the SRTT is longer than the RTO.Min value. The analysis of these problems also showed that the reported spurious T3 timeouts problems were resolved only by increasing the RTO.Min value well above the SRTT value.

The fact that the spurious T3 timeouts were successfully prevented only by setting the RTO.Min value considerably above the SRTT value, leads to conclusion that the RTO values, which are derived by the

[RFC4960] defend rules, are inadequate for the RTT variation pattern in the telecom signalling networks.

In other words, the fact that the SCTP association operation is stable only when the RTO.Min value is well above the SRTT value, makes the RTO calculation, which is specified by the [RFC4960] section 6.3.1. rules C1 C2 and C3, seemingly redundant.

To help visualise the problem, let assume, hypothetically, that the packets transmission pattern consists of high packet rate sequences longer than 500 msec with, for example, 200 packets/sec, which separated by 50 to 80 ms "idle" gaps. For such packet rate pattern, the statistical estimator algorithm for RTTVAR will produce a very low RTTVAR values, very likely well below 5 msec, because, during the long high packet rate sequences, the SACK delay will vary around 5 msec due to packet rate of 200 packets/sec.

Consequently, with the [RFC4960] RTO calculation rule:

RTO <- max(SRTT + 4 * RTTVAR, RTO.Min),

the RTO margin to absorb unexpected SACK delays, in this hypothetical case 50 to 80 msec due to the packet transmission gaps, is determined by the difference between the calculated RTO value and the measured (calculated) SRTT.

Since in case of low RTTVAR values the RTO is determined by the RTO.Min parameter, the RTO margin will be equal to the difference between the RTO.Min and SRTT (RTO margin = RTO.Min - SRTT). Thus, as illustrated in Figure 2, the [RFC4960] RTO calculation rules produce robust RTO values only when the SRTT is well below RTO.Min parameter value, which is the root cause of the problem.

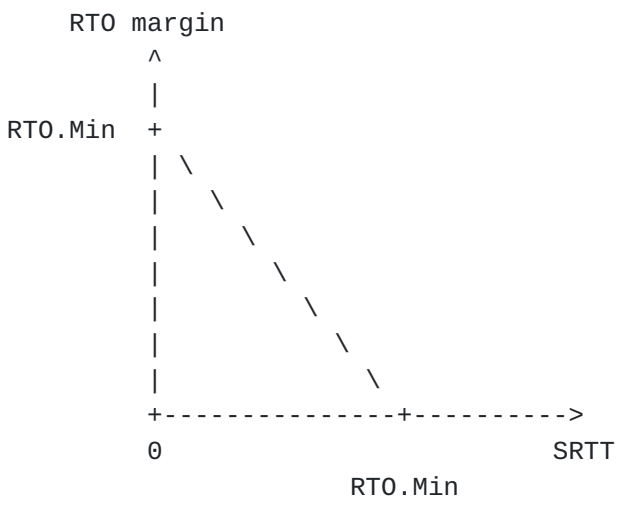

Figure 2: Relation between the RTO margin and SRTT

To rectify this anomaly, this document introduces modification in the [RFC4960] algorithm for RTO calculation. The actual modification is specified in Section 3 and it includes only change in the use of the RTO.Min protocol parameter; the technical background for the modification is elaborated in the Appendix A.

3. The modified algorithm for RTO Calculation

The modified rules governing the computation of SRTT, RTTVAR and RTO are as follows:

C1) Until an RTT measurement has been made for a packet sent to the given destination transport address, set RTO to the protocol parameter 'RTO.Initial'.

C2) When the first RTT measurement R is made, set

SRTT <- R,

RTTVAR <- R/2, and

RTO <- SRTT + max(4 * RTTVAR, RTO.Min).

C3) When a new RTT measurement R' is made, set

RTTVAR <- (1 - RTO.Beta) * RTTVAR + RTO.Beta * |SRTT - R'|

and

SRTT <- (1 - RTO.Alpha) * SRTT + RTO.Alpha * R'

Note: The value of SRTT used in the update to RTTVAR is its value before updating SRTT itself using the second assignment.

After the SRTT and RTTVAR computation, update RTO:

RTO <- SRTT + max(4 * RTTVAR, RTO.Min).

C4) When data is in flight and when allowed by rule C5 below, a new RTT measurement MUST be made each round trip. Furthermore, new RTT measurements SHOULD be made no more than once per round trip for a given destination transport address. There are two reasons for this recommendation: First, it appears that measuring more frequently often does not in practice yield any significant benefit [ALLMAN99]; second, if measurements are made more often, then the values of RTO.Alpha and RTO.Beta in rule C3 above should be adjusted so that SRTT and RTTVAR still adjust to changes at roughly the same rate (in terms of how many round trips it takes them to reflect new values) as they would if making only one measurement per round-trip and using RTO.Alpha and RTO.Beta as given in rule C3. However, the exact nature of these adjustments remains a research issue.

C5) Karn's algorithm: RTT measurements MUST NOT be made using packets that were retransmitted (and thus for which it is ambiguous whether the reply was for the first instance of the chunk or for a later instance).

IMPLEMENTATION NOTE: RTT measurements should only be made using a chunk with TSN r if no chunk with TSN less than or equal to r is retransmitted since r is first sent.

C6) A maximum value may be placed on RTO provided it is at least RTO.max seconds.

There is no requirement for the clock granularity G used for computing RTT measurements and the different state variables, other than:

G1) Whenever RTTVAR is computed, if RTTVAR = 0, then adjust RTTVAR <- G.

Experience [ALLMAN99] has shown that finer clock granularities (<= 100 msec) perform somewhat better than more coarse granularities.

4. IANA Considerations

This document does not create any new registries or modify the rules for any existing registries managed by IANA.

5. Security Considerations

This document does not add any security considerations to those given in [RFC4960].

6. References

6.1. Normative References

[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>.

[RFC4960] Stewart, R., "Stream Control Transmission Protocol", September 2007, <https://tools.ietf.org/html/rfc4960>.

6.2. Informative References

[ALLMAN99] Mark Allman and Vern Paxson, "On Estimating End-to-End Network Path Properties", 1999, <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20000004338.pdf>.

[JAC88] Van Jacobson and Michael J. Karels , "Congestion Avoidance and Control", November 1988, <https://people.eecs.berkeley.edu/~sylvia/cs268/papers/congavoid.pdf>.

Appendix A. Technical background for the modifications in the RTO calculation algorithm

As indicated in Section 2, with the [RFC4960] RTO calculation rules, the frequency of spurious T3 timeouts increases when the SRTT value is close to the RTO.Min value to the point where, under heavy load, the association becomes unstable if the SRTT is longer than the RTO.Min value.

The reasons for such outcome can be contributed to the following factors:

a) The characteristic of the statistical estimator algorithms for SRTT and RTTVAR calculation;

b) The anomalies in the distribution of the RTT measurement values caused by the [RFC4960] SACK generation rules, specifically, the delay of SACK sending; and

c) Inappropriate solution for protection against underestimated RTO values.

The characteristics of the statistical estimator algorithms for SRTT and RTTVAR, which are the foundation for RTO calculation, are well known and widely investigated in terms of improving the outcome (reduction of spurious T3 timeouts) by adjustment of the statistical estimator algorithms' configurable parameters. For example, the investigation results published in [ALLMAN99] indicate that lower gain factors RTO.Alpha and RTO.Beta, in the SRTT and RTTVAR calculations formulas, reduces the probability of computing a low RTO value that will result in T3 timeout. The same source also states that lower spurious T3 timeouts probability is also achieved by increasing the RTTVAR component i.e., the value of the factor K in the RTO calculation formula:

RTO <- SRTT + K * RTTVAR.

This behaviour can be related to the well-known characteristic of the statistical estimator algorithms for SRTT and RTTVAR estimation, which can be described as follows: If the RTT measurements values converge to a single RTT value, the calculated RTTVAR converge to zero (0) and the calculated RTO converge to SRTT. As a result, a relatively short sequence of moderately low RTT values, which are within the RTT values range, simultaneously lowers the SRTT and RTTVAR values to the point where the calculated RTO value is below the highest value in the RTT variation range, which may result in spurious T3 timeout if the next RTT is at the top of the RTT variation range.

This 'problem' is further exacerbated by the SCTP protocol rules for sending SACK which allow SACK delay of up to 500 msec. Namely, the SACK delay rules, combined with burst nature of the data packets' arrival pattern in the telecom signalling networks, drastically increase the jitteriness of the RTT measurements. That, in turn, adversely affect the results obtained by statistical estimator algorithms for SRTT and RTTVAR calculations in terms of underestimated RTO values that are prone to spurious T3 timeouts.

Obviously, and as proven in the operating networks, an RTO determined by application of rule C6, with an RTO.Min value in seconds,

practically eliminates underestimated RTO values and with that the spurious T3 timeouts. That is because the 1 second RTO will be well above the delay inserted by the terrestrial transport networks, which operate with latency below 100 msec, and because the SACK delay is also well below 1 second.

However, an RTO value in seconds, coupled with the RTO back-off rule RTO <- RTO * 2, results in too long detection of remote endpoint failure or complete failure of the physical layer. For example, with the [RFC4960] recommended RTO.Min of 1 second, RTO.Max of 60 seconds and Association.Max.Retrans of 4 attempts, the association closure time will be 31 seconds, which is an unacceptably long time that, under high load, can potentially destabilise the operation of the network.

Namely, in the telecom networks where the client nodes are connected to redundant server nodes and where multiple load sharing SCTP associations are used between the nodes, a timely detection of the SCTP remote peer endpoint failure, or complete failure of the physical layer, is critical to enables failover to the redundant resources.

Thus, instead of using an arbitrary long RTO defend by RTO.Min parameter, which practically makes the calculated RTO value by rules C1, C2 and C3 redundant, the RTO value should reflect, as close as possible, the real conditions in the network in terms of the time to transport the packets between two endpoints, the time delays induced by the SCTP protocol rules and to also include adequate additional time as protection against underestimated RTO values. To achieve that, the subsequent paragraphs first analyse the characteristics of the RTT components and then specify a modified RTO calculation algorithm which is derived from the characteristics of the statistical estimator algorithms for SRTT and RTTVAR and the characteristics of the RTT components.

Specifically, an RTT measurement starts at transmission of data, or at transmission of HEARTBEAT, and it is completed at reception of the corresponding SACK or HEARTBEAT ACK from the remote peer endpoint.

The RTT measurements results, which are based on data transfer and SACK reception, will be influenced by the following main components:

a) Transport network's physical layer propagation times in forward and backward directions.

b) IP network layer IP packets' sending, receiving and processing times in forward and backward directions.

c) The time to send, receive and process SCTP packet at the transmitting and receiving SCTP endpoints.

d) SACK sending delay when SACK is not sent for every received packet.

A similar RTT structuring can be constructed for the RTT measurements based on HEARTBEAT and HEARTBEAT ACK however, since HEARTBEAT ACK is sent for every HEARTBEAT with no delay, the HEARTBEAT based RTT estimation is less 'challenging' and it will not be examined in detail in this document.

The component 'a)', the transport network's physical layer propagation time is a stable component determined primarily by the length of the connection between two endpoints and to a very small degree by the nature of the physical medium (coper, coax cable, radio link, etc.). This component determines the theoretical/absolute minimum RTT time and it changes only when the physical properties of the connection, primarily the length, are changed.

The components 'b)' and 'c)', the IP network layer and SCTP endpoints packets sending, receiving and processing times are proportional to the traffic level (A) by factor 1/(1-A), which is the mean value of the waiting queues length. However, the actual time durations are derived as a product of the waiting queue length (the number of packets waiting to be processed) and the time to process a packet (the time to transmit/receive packet or the time to process a packet by the protocol stack's layers). Since the waiting queues' lengths are variable the aggregated time to send, receive and process SCTP packet will be variable too. Because the networks' load variation's gradient is generally small and because the telecom networks' signalling traffic is normally carried over high speed IP backbone networks with engineered capacity i.e., with no congestion, the variation of this timing components values will be significantly smaller than the variation range due to SACK delay.

The time component due to bullet 'd)' is the delay time inserted by the SCTP protocol rules and it is applicable only when the SACK is not returned on every packet.

Namely, when SACK is returned on every received packet, the RTT measurement value R is determined only by the combined time from components 'a)', 'b)' and 'c)', which in this context will be called NRTT (Network RTT). However, when the SACK is not returned on every packet i.e., when the SACK is returned on every 'N-th' received packet, and N > 1, the RTT measurement value R is determined by NRTT and the allowed SACK delay time.

Specifically, if the packets' arrival rate/frequency F is low, relative to the value of the protocol parameter SACK delay timer (SACK.Delay.timer), i.e., if the relation

(N - 1) * 1/F >= SACK.Delay.timer

is true, the RTT measurement value will be determined by the NRTT and the SACK.Delay. In that case, the RTT measurement value R can be expressed as follows:

R = NRTT + SACK.Delay.timer.

Alternatively, if the packets' arrival rate F is high, relative to the SACK.Delay, i.e., if the inequation

(N - 1) * 1/F < SACK.Delay.timer

is true, the RTT measurement value will be determined by the NRTT and the time to receive the number of packets required to trigger sending of SACK. In that case, the RTT measurement value can be expressed as follows:

R = NRTT + (N - 1) * 1/F.

Since by the [RFC4960] specifications the number of received packets that is required to trigger sending of SACK is limited to 2 (N = 2), the expression for the RTT measurement value can be simplified as follows:

R = NRTT + 1/F.

Thus, in general, the RTT measurement value can be expressed as follows:

R = NRTT + min(SACK.Delay.timer, 1/F).

In other words, for any packet arrival rate F, the shortest RTT measurement value is greater than the NRTT and the longest RTT measurement value does not exceed NRTT plus SACK.Delay i.e., the following relation is true:

NRTT + 1/maxF < R <= NRTT + SACK.Delay.timer,

where maxF is the highest packets arrival rate. Consequently, the range of the RTT measurements R is given by the following relation:

NRTT + 1/maxF <= R <= NRTT + SACK.Delay.timer,

Or in other words, the values of the RTT measurements R will be between a minimum value (minR) that is determined as:

minR = NRTT + 1/maxF,

and a maximum value (maxR) that is determined as:

maxR = NRTT + SACK.Delay.timer.

The above presented RTT related relations are illustrated in Figure 3.

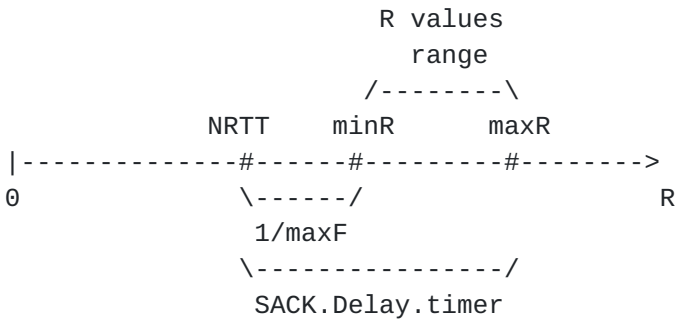

Figure 3: The expected values range of the RTT measurements R

The above analysis also shows that the SACK delay, in practical terms, significantly increases the RTT (R'), which leads to conclusion that the calculated SRTT (mean RTT) by formula:

SRTT <- (1 - RTO.Alpha) * SRTT + RTO.Alpha * R';

converges to a value greater than NRTT + 1/maxF i.e., to a value greater than the lowest RTT, regardless of the variation pattern of the measured RTTs.

At that same time, the above analysis shows that the SACK delay significantly increases the RTT measurement (R') variation range but it does not alter the RTTVAR convergence to 0, or rather low values when calculated by formula:

RTTVAR <- (1 - RTO.Beta) * RTTVAR + RTO.Beta * |SRTT - R'|.

Or in other words, the RTTVAR calculation can still yield low values even though the SACK delay increases the RTT measurement (R') variation range (refer to Figure 3).

That, combined with the fact that RTTVAR contribution to the RTO value is 4 times of SRTT (RTO <- SRTT + 4 * RTTVAR), leads to conclusion that the RTO underestimations are primarily due to low

RTTVAR values. Thus, instead of setting low threshold for the calculated RTO, which is the role of rule C6, the compensation for underestimated RTOs should be achieved by setting low threshold for RTTVAR as follows:

After calculating RTTVAR by formula:

RTTVAR <- (1 - RTO.Beta) * RTTVAR + RTO.Beta * |SRTT - R'|,

if RTTVAR is less than RTTVAR.Min set RTTVAR to RTTVAR.Min.

Or by altering the RTO calculation formula as follows:

RTO <- SRTT + max(4 * RTTVAR, RTTVAR.Min).

However, to avoid introduction of new protocol parameter, and because the existing RTO.Min protocol parameter is no longer used, RTO.Min can take the role of the RTTVAR.Min. In that case, the RTO calculation formula will be expressed as follows:

RTO <- SRTT + max(4 * RTTVAR, RTO.Min).

The above formula ensures that, in case of low RTTVAR values, the RTO margin to absorb unexpected SACK delays is determined by the RTO.Min (the RTTVAR.Min alias) only, thus, it is constant and independent of the SRTT (refer to the illustration in Figure 4).

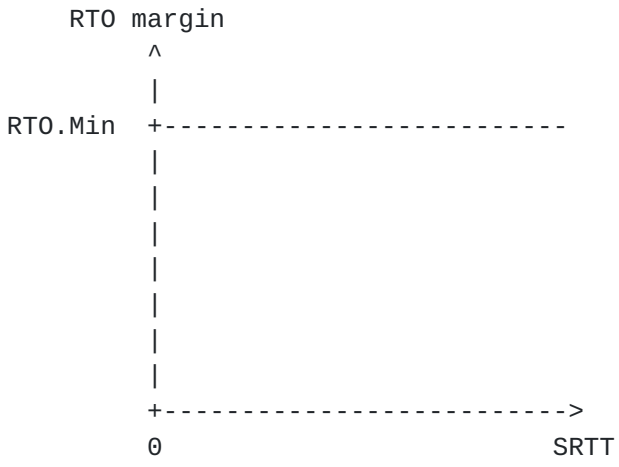

Figure 4: Relation between the RTO margin and SRTT with the new RTO calculation rules

Since the RTT variation range introduced by SACK delay is predictable i.e., the RTT variation range introduced by SACK delay is, in practical terms, determined by the SACK delay time (refer to Figure 2), the value of the RTTVAR low threshold should be determined based on the SACK delay time used at the remote peer.

The [RFC4960] recommended value for RTO.Min does not require change when the RTO.Min is used as RTTVAR low threshold in the above modified formula for RTO calculation. Namely, the recommended 1 sec correspond to 2 times the allowed SACK delay time, which is 500 msec.

Authors' Addresses

Dimitar Jovev
Ericsson
818 Bourke St.
Melbourne, Victoria 3008
Australia

Email: dimitar.jovev@gmail.com

Maksim Proshin
Ericsson
Kistavaegen 25
Stockholm 164 80
Sweden

Email: mproshin@tieto.mera.ru