QUIC
Internet-Draft
Intended status: Informational
Expires: June 3, 2018

B. Trammell, Ed.
P. De Vaere
ETH Zurich
R. Even
Huawei
G. Fioccola
Telecom Italia
T. Fossati
Nokia
M. Ihlar
Ericsson
A. Morton
AT&T Labs
E. Stephan
Orange

November 30, 2017

The Addition of a Spin Bit to the QUIC Transport Protocol draft-trammell-quic-spin-00

Abstract

This document summarizes work to date on the addition of a "spin bit", intended for explicit measurability of end-to-end RTT on QUIC flows. It proposes a detailed mechanism for the spin bit, describes how to use it to measure end-to-end latency, discusses corner cases and workarounds therefor in the measurement, describes experimental evaluation of the mechanism done to date, and examines the utility and privacy implications of the spin bit. As the overhead and risk associated with the spin bit are negligible, and the utility of a passive RTT measurement signal at higher resolution than once per flow is clear, this document advocates for the addition of the spin bit to the protocol.

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 June 3, 2018.

Table of Contents

1. Introduction

The QUIC transport protocol [QUIC-TRANS] is a UDP-encapsulated protocol integrated with Transport Layer Security (TLS) [TLS] to encrypt most of its protocol internals, beyond those handshake packets needed to establish or resume a TLS session, and information required to reassemble QUIC streams (the packet number) and to route QUIC packets to the correct machine in a load-balancing situation (the connection ID). In other words, in contrast to TCP, QUIC's wire image (see [WIRE-IMAGE]) exposes much less information about

transport protocol state than TCP's wire image. Specifically, the fact that sequence and acknowledgement numbers and timestamps cannot be seen by on-path observes in QUIC as they can be in the TCP means that passive TCP loss and latency measurement techniques that rely on this information (e.g. [CACM-TCP], [TMA-QOF]) cannot be easily ported to work with QUIC.

This document proposes a solution to this problem by adding a "latency spin bit" to the QUIC short header. This bit is designed solely for explicit passive measurability of the protocol. It provides one RTT sample per RTT to passive observers of QUIC traffic. It describes the mechanism, how it can be added to QUIC, and how it can be used by passive measurement facilities to generate RTT samples. It explores potential corner cases and shortcomings of the mechanism and how they can be worked around. It summarizes experimental results to date with an implementation of the spin bit built atop a recent QUIC implementation. It additionally describes use cases for passive RTT measurement at the resolution provided by the spin bit. It further reviews findings on privacy risk researched by the QUIC RTT Design Team, which was tasked by the IETF QUIC Working Group to determine the risk/utility tradeoff for the spin bit.

The spin bit has low overhead, presents negligible privacy risk, and has clear utility in providing passive RTT measurability of QUIC that is far superior to QUIC's measurability without the spin bit, and equivalent to or better than TCP passive measurability.

1.1. About This Document

This document is maintained in the GitHub repository https://github.com/britram/draft-trammell-quic-spin, and the editor's copy is available online at https://britram.github.io/draft-trammell-quic-spin. Current open issues on the document can be seen at https://github.com/britram/draft-trammell-quic-spin/issues. Comments and suggestions on this document can be made by filing an issue there, or by contacting the editor.

2. The Spin Bit Mechanism

The latency spin bit enables latency monitoring from observation points on the network path. The bit is set by the endpoints in the following way:

- The server sets the spin bit value to the value of the spin bit in the packet received from the client with the largest packet number.

- The client sets the spin bit value to the opposite of the value set in the packet received from the server with the largest packet number, or to 0 if no packet as been received yet.

If packets are delivered in order, this procedure will cause the spin bit to change value in each direction once per round trip. Observation points can estimate the network latency by observing these changes in the latency spin bit, as described in Section 3.

2.1. Proposed Short Header Format Including Spin Bit

Since it is possible to measure handshake RTT without a spin bit (see Section 3.2), it is sufficient to include the spin bit in the short packet header. This proposal suggests to ues the second most significant bit (0x40) of the first octet in the short header for the spin bit.

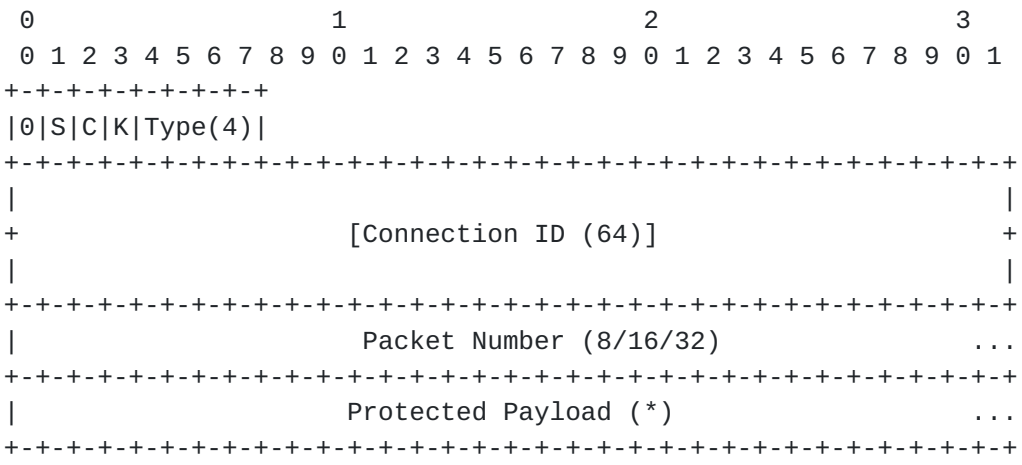

Figure 1: Short Header Format including proposed Spin Bit

This will shift the Connection ID flag and the Key Phase Bit to 0x20 and 0x10 respectively, and will limit the number of available short packet types to 16.

3. Using the Spin Bit for Passive RTT Measurement

When a QUIC flow is sending at full rate (i.e., neither application nor flow control limited), the latency spin bit in each direction changes value once per round-trip time (RTT). An on-path observer can observe the time difference between edges in the spin bit signal to measure one sample of end-to-end RTT. Note that this measurement, as with passive RTT measurement for TCP, includes any transport protocol delay (e.g., delayed sending of acknowledgements) and/or application layer delay (e.g., waiting for a request to complete). It therefore provides devices on path a good instantaneous estimate

of the RTT as experienced by the application. A simple linear smoothing or moving minimum filter can be applied to the stream of RTT information to get a more stable estimate.

We note that the Latency Spin Bit, and the measurements that can be done with it, can be seen as an end-to-end extension of a special case of the alternate marking method described in [ALT-MARK].

3.1. Limitations and Workarounds

Application-limited and flow-control-limited senders can have application and transport layer delay, respectively, that are much greater than network RTT. Therefore, the spin bit provides network latency information only when the sender is neither application nor flow control limited. When the sender is application-limited by periodic application traffic, where that period is longer than the RTT, measuring the spin bit provides information about the application period, not the RTT. Simple heuristics based on the observed data rate per flow or changes in the RTT series can be used to reject bad RTT samples due to application or flow control limitation.

Since the spin bit logic at each endpoint considers only samples on packets that advance the largest packet number seen, signal generation itself is resistent to reordering. However, reordering can cause problems at an observer by causing spurious edge detection and therefore low RTT estimates. This can be probabilistically mitigated by the observer tracking the low-order bits of the packet number, and rejecting edges that appear out-of-order.

3.2. Alternate RTT Measurement Approaches for Diagnosing QUIC flows

There are two broad alternatives to explicit signaling for passive RTT measurement for measuring the RTT experienced by QUIC flows.

The first of these is handshake RTT measurement. As described in [QUIC-MGT], the packets of the QUIC handshake are distinguishable on the wire in such a way that they can be used for one RTT measurement sample per flow: the delay between the client initial and the server cleartext packet can be used to measure "upstream" RTT (between the observer and the server), and the delay between the server cleartext packet and the next client cleartext packet can be used to measure "downstream" RTT (between the client and the observer). When RTT measurements are used in large aggregates (all flows traversing a large link, for example), a methodology based on handshake RTT could be used to generate sufficient samples for some purposes without the spin bit.

However, this methodology would rely on the assumption that the difference between handshake RTT and nominal in-flow RTT is negligible. Specifically, (1) any additional delay required to compute any cryptographic parameters must be negligible with respect to network RTT; (2) any additional delay required to establish state along the path must be negligible with respect to network RTT; and (3) network treatment of initial packets in a flow must identical to that of later packets in the flow. When these assumptions cannot be shown to hold, spin-bit based RTT measurement is preferable to handshake RTT measurement, even for applications for which handshake RTT measurement would otherwise be suitable.

The second alternative is parallel active measurement: using ICMP Echo Request and Reply [RFC0792] [RFC4433], a dedicated measurement protocol like TWAMP [RFC5357], or a separate diagnostic QUIC flow to measure RTT. Regardless of protocol, the active measurement must be initiated by a client on the same network as the client of the QUIC flow(s) of interest, or a network close by in the Internet topology, toward the server. Note that there is no guarantee that ICMP flows will receive the same network treatment as the flows under study, both due to differential treatment of ICMP traffic and due to ECMP routing (see e.g. [TOKYO-PING]). TWAMP and QUIC diagnostic flows, though both use UDP, have similar issues regarding ECMP. However, in situations where the entity doing the measurement can guarantee that the active measurement traffic will traverse the subpaths of interest (e.g. residential access network measurement under a network architecture and business model where the network operator owns the CPE), active measurement can be used to generate RTT samples at the cost of at least two non-productive packets sent though the network per sample.

3.3. Experimental Evaluation

We have evaluated the effectiveness of the spin bit in an emulated network environment. The spin bit was added to a fork of [MINQ], using the mechanism described in Section 2, but with the spin bit appearing in a measurement byte added to the header for passive measurability experiments. Spin bit measurement support was added to [MOKUMOKUREN]. Full results of these ongoing experiments are available online in [SPINBIT-REPORT], but we summarize our findings here.

First, we confirm that the spin bit works as advertised: it provides one useful RTT sample per RTT to any passive observer of the flow. This sample tracks each sender's local instantaneous estimate of RTT as well as the expected RTT (i.e., defined by the emulation) fairly well. One surprising implication of this is that the spin bit provides _more_ information than is available by local estimation to

an endpoint which is mostly receiving data frames and sending mainly ACKs, and as such can also be useful in purely endpoint-local observations of the RTT evolution during the flow. The spin bit also works correctly under moderate to heavy packet loss and jitter.

Second, we confirm that the spin bit can be easily implemented without requiring deep integration into a QUIC implementation. Indeed, it could be implemented completely independently, as a shim, aside from the requirement that the spin bit value be integrity-protected along with the rest of the QUIC header.

Third, we performed experiments focused on the intermittent-sender problem described in Section 3.1. We confirm that the spinbit does not provide useful RTT samples after the handshake when packets are only sent intermittently. Simple heuristics can be used to recognize this situation, however, and to reject these RTT samples. We also find that a simple sender-side heuristic can be used to determine whether a sample will be useful. If a sender sends a packet more than a specified delay (e.g. 1ms) after the last packet received by the client, it knows that any latency spin observation of that packet will be invalid. If a second "spin valid" bit were available, the sender could then mark that packet "spin invalid". Our experiments show that this simple heuristic and spin validity bit are succesful in marking all packets whose RTT samples should be rejected.

Fourth, we performed experiments focused on the reordering problem described in Section 3.1. We find that while reordering can cause spurious samples at a naive observer, two simple approaches can be used to reject spurious RTT samples due to reordering. First, a two-bit spin signal that always advances in a single direction (e.g. 00 -> 01 -> 10 -> 11) successfully rejects all reordered samples, including under amounts of reordering that render the transport itself mostly useless. However, adding a bit is not necessary: having the observer keep the least significant bits of the packet number, and rejecting samples from packets that do not advance by one, as suggested in Section 3.1, is essentially as successful as a two-bit spin signal in mitigating the effects of reordering on RTT measurement.

Fifth, we performed parallel active measurements using ping, as described in Section 3.2. In our emulated network, the ICMP packets and the QUIC packets traverse the same links with the same treatment, and share queues at each link, which mitigates most of the issues with ping. We find that while ping works as expected in measuring end-to-end RTT, it does not track the sender's estimate of RTT, and as such does not measure the RTT experienced by the application layer as well as the spin bit does.

In summary, our experiments show that the spin bit is suitable for purpose, can be implemented with minimal disruption, and that most of the problems identified with it in specific corner cases can be easily mitigated. See [SPINBIT-REPORT] for more.

4. Use Cases for Passive RTT Measurement

This section describes use cases for passive RTT measurement. Most of these are currently achieved with TCP, i.e., the matching of packets based on sequence and acknowledgment numbers, or timestamps and timestamp echoes, in order to generate upstream and downstream RTT samples which can be added to get end-to-end RTT. These use cases could be achieved with QUIC by replacing sequence/ acknowledgement and timestamp analysis with spin bit analysis, as described in Section 3.

This section currently focuses one initial use case, interdomain troubleshooting. Additional use cases will be added in future revisions; see https://github.com/britram/draft-trammell-spin-bit/issues for use cases we are currently considering.

In any case, the measurement methodology follows one of a few basic variants:

- The RTT evolution of a flow or a set of flows can be compared to baseline or expected RTT measurements for flows with the same characterisitcs in order to detect or localize latency issues in a specific network.

- The RTT evolution of a single flow can also be examined in detail to diagnose performance issues with that flow.

- The spin bit can be used to generate a large number of samples of RTT for a flow aggregate (e.g., all flows between two given networks) without regard to temporal evolution of the RTT, in order to examine the distribution of RTTs for a group of flows that should have similar RTT (e.g., because they should share the same path(s)).

4.1. Interdomain Troubleshooting

Network access providers are often the first point of contact by their customers when network problems impact the performance of bandwidth-intensive and latency-sensitive applications such as video, regardless of whether the root cause lies within the access provider's network, the service provider's network, on the Internet paths between them, or within the customer's own network.

Many residential networks use WiFi (802.11) on the last segment, and WiFi signal strength degradation manifests in high first-hop delay, due to the fact that the MAC layer will retransmit packets lost at that layer. Measuring the RTT between endpoints on the customer network and parts of the service provider's own infrastructure (which have predictable delay characteristics) can be used to isolate this cause of performance problems.

Comparing the evolution of passively-measured RTTs between a customer network and selected other networks on the Internet to short- and medium-term baseline measurements can similarly be used to isolate high latency to specific networks or network segments. For example, if the RTTs of all flows to a given content provider increase at the same time, the problem likely exists between the access network and the content provider, or in the content provider's network itself. On the other hand, if the RTTs of all flows passing through the same access provider infrastructure change together, then the change is likely attributable to that infrastructure.

These measurements are particularly useful for traffic which is latency sensitive, such as interactive video applications. However, since high latency is often correlated with other network-layer issues such as chronic interconnect congestion [IMC-CONGESTION], it is useful for general troubleshooting of network layer issues in an interdomain setting.

In this case, multiple RTT samples per flow are useful less for observing intraflow behavior, and more for generating sufficient samples for a given aggregate to make a high-quality measurement.

5. Privacy and Security Considerations

The privacy considerations for the latency spin bit are essentially the same as those for passive RTT measurement in general.

A concern was raised during the discussion of this feature within the QUIC working group and the QUIC RTT Design Team that high-resolution RTT information might be usable for geolocation. However, an evaluation based on RTT samples taken over 13,780 paths in the Internet from RIPE Atlas anchoring measurements [TRILAT] shows that the magnitude and uncertainty of RTT data render the resolution of geolocation information that can be derived from Internet RTT is limited to national- or continental-scale; i.e., less resolution than is generally available from free, open IP geolocation databases.

One reason for the inaccuracy of geolocation from network RTT is that Internet backbone transmission facilities do not follow the great-circle path between major nodes. Instead, major geographic features

and the efficiency of connecting adjacent major cities influence the facility routing. An evaluation of ~3500 measurements on a mesh of 25 backbone nodes in the continental United States shows that 85% had RTT to great-circle error of 3ms or more, making location within US State boundaries ambigous [CONUS].

Therefore, in the general case, when an endpoint's IP address is known, RTT information provides negligible additional information.

RTT information may be used to infer the occupancy of queues along a path; indeed, this is part of its utility for performance measurement and diagnostics. When a link on given path has excessive buffering (on the order of hundreds of milliseconds or more; a situation colloquially referred to as "bufferbloat"), such that the difference in delay between an empty queue and a full queue dwarfs normal variance and RTT along the path, RTT variance during the lifetime of a flow can be used to infer the presence of traffic on the bottleneck link. In practice, however, this is not a concern for passive measurement of congestion-controlled traffic, since any observer in a situation to observe RTT passively need not infer the presence of the traffic, as it can observe it directly.

In addition, since RTT information contains application as well as network delay, patterns in RTT variance from minimum, and therefore application delay, can be used to infer or fingerprint application-layer behavior. However, as with the case above, this is not a concern with passive measurement, since the packet size and interarrival time sequence, which is also directly observable, carries more information than RTT variance sequence.

We therefore conclude that the high-resolution, per-flow exposure of RTT for passive measurement as provided by the spin bit poses negligible marginal risk to privacy.

As shown in Section 2, the spin bit can be implemented separately from the rest of the mechanisms of the QUIC transport protocol, as it requires no access to any state other than that observable in the QUIC packet header itself. We recommend that implementations take advantage of this property, to reduce the risk that a errors in the implementation could leak private transport protocol state through the spin bit.

Since the spin bit is disconnected from transport mechanics, a QUIC endpoint implementing the spin bit that has a model of the actual network RTT and a target RTT to expose can "lie" about its spin bit transitions, even without coordination with and the collusion of the other endpoint. This is not the case with TCP, which requires coordination and collusion to expose false information via its

sequence and acknowledgment numbers and its timestamp option. When passive measurement is used for purposes where one endpoint might gain a material advantage by representing a false RTT, e.g. SLA verification or enforcement of telecommunications regulations, this situation raises a question about the trustworthiness of spin bit RTT measurements.

This issue must be appreciated by users of spin bit information, but mitigation is simple, as QUIC implementations designed to lie about RTT through spin bit modification are subject to dynamic analysis along paths with known RTTs. We consider the ease of verification of lying in situations where this would be prohibited by regulation or contract, combined with the consequences of violation of said regulation or contract, to be a sufficient incentive in the general case not to do it.

6. Acknowledgments

Many thanks to Christian Huitema, who originally proposed the spin bit as pull request 609 on [QUIC-TRANS]. Thanks to the QUIC RTT Design Team for discussions leading especially to the measurement limitations and privacy and security considerations sections.

This work is partially supported by the European Commission under Horizon 2020 grant agreement no. 688421 Measurement and Architecture for a Middleboxed Internet (MAMI), and by the Swiss State Secretariat for Education, Research, and Innovation under contract no. 15.0268. This support does not imply endorsement.

7. Informative References

[ALT-MARK] Fioccola, G., Capello, A., Cociglio, M., Castaldelli, L., Chen, M., Zheng, L., Mirsky, G., and T. Mizrahi, "Alternate Marking method for passive and hybrid performance monitoring", draft-ietf-ippm-alt-mark-13 (work in progress), October 2017.

[CACM-TCP] Strowes, S., "Passively Measuring TCP Round-Trip Times (in Communications of the ACM)", October 2013.

[CONUS] Morton, A., "Comparison of Backbone Node RTT and Great Circle Distances (https://github.com/acmacm/FIXME-TBD)", September 2017.

[IMC-CONGESTION]
Luckie, M., Dhamdhere, A., Clark, D., Huffaker, B., and k. claffy, "Challenges in Inferring Internet Interdomain Congestion (in Proc. ACM IMC 2014)", November 2014.

[MINQ] Rescorla, E., "MINQ, a simple Go implementation of QUIC (https://github.com/ekr/minq)", November 2017.

[MOKUMOKUREN]
Trammell, B., "Mokumokuren, a lightweight flow meter using gopacket (https://github.com/britram/mokumokuren)", November 2017.

[QUIC-MGT]
Kuehlewind, M. and B. Trammell, "Manageability of the QUIC Transport Protocol", draft-ietf-quic-manageability-01 (work in progress), October 2017.

[QUIC-TRANS]
Iyengar, J. and M. Thomson, "QUIC: A UDP-Based Multiplexed and Secure Transport", draft-ietf-quic-transport-07 (work in progress), October 2017.

[RFC0792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, DOI 10.17487/RFC0792, September 1981, <https://www.rfc-editor.org/info/rfc792>.

[RFC4433] Kulkarni, M., Patel, A., and K. Leung, "Mobile IPv4 Dynamic Home Agent (HA) Assignment", RFC 4433, DOI 10.17487/RFC4433, March 2006, <https://www.rfc-editor.org/info/rfc4433>.

[RFC5357] Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and J. Babiarz, "A Two-Way Active Measurement Protocol (TWAMP)", RFC 5357, DOI 10.17487/RFC5357, October 2008, <https://www.rfc-editor.org/info/rfc5357>.

[SPINBIT-REPORT]
De Vaere, P., "Latency Spinbit Implementation Experience", November 2017.

[TLS] Rescorla, E., "The Transport Layer Security (TLS) Protocol Version 1.3", draft-ietf-tls-tls13-21 (work in progress), July 2017.

[TMA-QOF] Trammell, B., Gugelmann, D., and N. Brownlee, "Inline Data Integrity Signals for Passive Measurement (in Proc. TMA 2014)", April 2014.

[TOKYO-PING]
Pelsser, C., Cittadini, L., Vissicchio, S., and R. Bush, "From Paris to Tokyo - On the Suitability of ping to Measure Latency (ACM IMC 2014)", October 2014.

[TRILAT] Trammell, B., "On the Suitability of RTT Measurements for Geolocation (https://github.com/britram/trilateration/blob/paper-rev-1/paper.ipynb)", August 2017.

[WIRE-IMAGE]
Trammell, B. and M. Kuehlewind, "The Wire Image of a Network Protocol", draft-trammell-wire-image-00 (work in progress), November 2017.

Authors' Addresses

Brian Trammell (editor)
ETH Zurich

Email: ietf@trammell.ch

Piet De Vaere
ETH Zurich

Email: piet@devae.re

Roni Even
Huawei

Email: roni.even@huawei.com

Giuseppe Fioccola
Telecom Italia

Email: giuseppe.fioccola@telecomitalia.it

Thomas Fossati
Nokia

Email: thomas.fossati@nokia.com

Marcus Ihlar
Ericsson

Email: marcus.ihlar@ericsson.com

Al Morton
AT&T Labs

Email: acmorton@att.com

Emile Stephan
Orange

Email: emile.stephan@orange.com