

QUIC
Internet-Draft
Intended status: Informational
Expires: 30 January 2021

M. Duke
F5 Networks, Inc.
29 July 2020

Network Address Translation Support for QUIC
draft-duke-quic-natsupp-03

Abstract

Network Address Translators (NATs) are widely deployed to share scarce public IPv4 addresses among multiple end hosts. They overwrite IP addresses and ports in IP packets to do so. QUIC is a protocol on top of UDP that provides transport-like services. QUIC is better-behaved in the presence of NATs than older protocols, and existing UDP NATs should operate without incident if unmodified. QUIC offers additional features that may tempt NAT implementers as potential optimizations. However, in practice, leveraging these features will lead to new connection failure modes and security vulnerabilities.

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 30 January 2021.

Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights

and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Conventions	3
3. QUIC and NAT Rebinding	3
4. The Lure of the Connection ID	4
4.1. Resource Conservation	4
4.2. "Helping" with routing infrastructure issues	5
5. Filtering behavior	5
6. QUIC Detection	6
7. Security Considerations	6
8. IANA Considerations	6
9. Informative References	6
Appendix A. Acknowledgments	7
Appendix B. Change Log	7
B.1. since draft-duke-quic-natsupp-02	7
B.2. since draft-duke-quic-natsupp-01	7
B.3. since draft-duke-quic-natsupp-00	7
Author's Address	7

1. Introduction

Network Address Translators (NATs) are a widely deployed means of multiplexing multiple private IP addresses over scarce IPv4 public address space by replacing those addresses and using ports to distinguish those connections. The new address can also guarantee that packets move through a proxy throughout the life of a connection, so that the connection can continue with the required state at that proxy.

This document uses the colloquial term NAT to mean NAPT (section 2.2 of [RFC3022]), which overloads several IP addresses to one IP address or to an IP address pool, as commonly deployed in carrier-grade NATs or residential NATs.

QUIC [QUIC-TRANSPORT] is a protocol, operating over UDP, that provides many transport-like services to the application layer. Among these services is the mapping of multiple endpoint IP addresses to a single connection through use of a Connection ID (CID). Connection IDs are opaque byte fields that are expressed consistently across all QUIC versions [QUIC-INVARIANTS]. This feature may appear to present opportunities to optimize NAT port usage and simplify the work of the QUIC server. In fact, NAT behavior that relies on CID may instead cause connection failure when endpoints change Connection ID, and disable important protocol security features.

The remainder of this document explains how QUIC supports NATs better than other connection-oriented protocols, why NAT use of Connection ID might appear attractive, and how NAT use of CID can create serious problems for the endpoints. The conclusion of this document is that NATs should retain their existing 4-tuple-based operation and refrain from parsing or otherwise using QUIC connection IDs.

[RFC4787] contains some guidance on building NATs to interact constructively with a wide range of applications. This document extends the discussion to QUIC.

2. 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 RFC 2119 [RFC2119].

3. QUIC and NAT Rebinding

An explicit goal of QUIC is to be robust to NAT rebinding. When a connection is idle for a long time, the NAT may guess it has terminated and assign the client port to a new connection. As TCP defines a connection by its address and port 4-tuple, a TCP packet will not appear to belong to any existing connection at the receiver.

QUIC endpoints identify their connections using a CID that is encoded in every packet. If the client attempts to resume communication, the packet will be assigned a new source IP and/or port. Incoming packets from the server will be misrouted and dropped until the client sends a packet from its new address.

Therefore, QUIC connections can survive NAT rebindings as long as no routing function in the path is dependent on client IP address and port to deliver packets between server and NAT. Reducing the timeout on UDP NATs might be tempting in light of this property, but not all QUIC server deployments will be robust to rebinding.

4. The Lure of the Connection ID

There are a few reasons that CID-aware NATs could seemingly appear attractive.

4.1. Resource Conservation

NATs sometimes hit an operational limit where they exhaust available public IP addresses and ports, and must evict flows from their address/port mapping. CIDs offer a way to multiplex many connections over a single address and port.

However, QUIC endpoints may negotiate new connection IDs inside cryptographically protected packets, and begin using them at will. Imagine two clients behind a NAT that are sharing the same public IP address and port. The NAT is differentiating them using the incoming Connection ID. If one client secretly changes its connection ID, there will be no mapping for the NAT, and the connection will suddenly break.

While mid-connection failure in some cases may seem superior to rejecting QUIC outright, HTTP/3 over QUIC falls back to TCP. This is preferable to a connection suddenly black holing and timing out. Furthermore, wide deployment of NATs with this behavior would make it risky to change Connection IDs in the internet, which would thwart various important protocol properties.

It is possible, in principle, to encode the client's identity in a connection ID using [QUIC-LB] and explicit coordination with the NAT. However, QUIC-LB makes assumptions about endpoint mobility and common configuration in server infrastructure that are almost never valid in client/NAT architectures. Deploying such a system would include the administrative overhead while not solving the problem described in this section if the client changes networks.

Note that using connection IDs in this manner would anyway violate the best common practice to avoid "port overloading" as described in [RFC4787].

4.2. "Helping" with routing infrastructure issues

One problem in QUIC deployment is router and switch server infrastructures that direct traffic based on address-port 4-tuple rather than connection ID. The use of source IP address means that a NAT rebinding or address migration will deliver packets to the wrong server. For the reasons described above, routers and switches will not have access to negotiated CIDs. This is a particular problem for low-state load balancers, and a QUIC extension exists [QUIC-LB] to allow some server-load balancer coordination for routable CIDs.

A NAT at the front of this infrastructure might save the effort of converting all these devices by decoding routable connection IDs and rewriting the packet IP addresses to allow consistent routing by legacy devices.

Unfortunately, the change of IP address or port is an important signal to QUIC endpoints. It requires a review of path-dependent variables like congestion control parameters. It can also signify various attacks that mislead one endpoint about the best peer address for the connection (see section 9 of [QUIC-TRANSPORT]). The QUIC PATH_CHALLENGE and PATH_RESPONSE frames are intended to detect and mitigate these attacks and verify connectivity to the new address. This mechanism cannot work if the NAT is bleaching peer address changes.

For example, an attacker might copy a legitimate QUIC packet and change the source address to match its own. In the absence of a bleaching NAT, the receiving endpoint would interpret this as a potential NAT rebinding and use a PATH_CHALLENGE frame to prove that the peer endpoint is not truly at the new address, thus thwarting the attack. A bleaching NAT has no means of sending an encrypted PATH_CHALLENGE frame, so it might start redirecting all QUIC traffic to the attacker address and thus allow an observer to break the connection.

5. Filtering behavior

[RFC4787] describes possible packet filtering behaviors that relate to NATs. Though the guidance there holds, a particularly unwise behavior is to admit a handful of UDP packets and then make a decision as to whether or not to filter it. QUIC applications are encouraged to fail over to TCP if early packets do not arrive at their destination. Admitting a few packets allows the QUIC endpoint to determine that the path accepts QUIC. Sudden drops afterwards will result in slow and costly timeouts before abandoning the connection.

6. QUIC Detection

Beyond the above difficulties, merely identifying that a UDP packet is part of a QUIC connection is not straightforward. Due to address migration, NATs cannot assume that QUIC version 1 application traffic is preceded by a handshake on the path. The short header prepended to version 1 application traffic has few consistent codepoints that reliably identify it as QUIC. Moreover, the protocol is designed to be extensible. [QUIC-INVARIANTS] describes the small set of QUIC protocol properties that will remain stable across versions.

For these reasons, applying generalized UDP policies will prevent accidental breakage of QUIC features and mishandled non-QUIC UDP packets.

7. Security Considerations

This document proposes no change in behavior in the internet, so there are no new security implications. However, ignoring the recommendations here could prevent existing security mechanisms in QUIC from working properly.

8. IANA Considerations

There are no IANA requirements.

9. Informative References

[QUIC-INVARIANTS]

Thomson, M., "Version-Independent Properties of QUIC", Work in Progress, Internet-Draft, draft-ietf-quic-invariants-latest, <<https://tools.ietf.org/html/draft-ietf-quic-invariants-latest>>.

[QUIC-LB]

Duke, M. and N. Banks, "QUIC-LB: Generating Routable QUIC Connection IDs", Work in Progress, Internet-Draft, draft-duke-quic-load-balancers-latest, <<https://tools.ietf.org/html/draft-duke-quic-load-balancers-latest>>.

[QUIC-TRANSPORT]

Iyengar, J., Ed. and M. Thomson, Ed., "QUIC: A UDP-Based Multiplexed and Secure Transport", Work in Progress, Internet-Draft, draft-ietf-quic-transport-latest, <<https://tools.ietf.org/html/draft-ietf-quic-transport-latest>>.

- [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>>.
- [RFC3022] Srisuresh, P. and K. Egevang, "Traditional IP Network Address Translator (Traditional NAT)", RFC 3022, DOI 10.17487/RFC3022, January 2001, <<https://www.rfc-editor.org/info/rfc3022>>.
- [RFC4787] Audet, F., Ed. and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, DOI 10.17487/RFC4787, January 2007, <<https://www.rfc-editor.org/info/rfc4787>>.

Appendix A. Acknowledgments

Thanks to Dmitri Tikhonov, who first recognized that certain NAT behaviors could create problems for QUIC.

Appendix B. Change Log

RFC Editor's Note: Please remove this section prior to\$ publication of a final version of this document.\$

B.1. since draft-duke-quic-natsupp-02

- * Added discussion of QUIC identification

B.2. since draft-duke-quic-natsupp-01

- * Added brief discussion of impact of filtering.
- * Added references to RFC 4787.
- * Corrected normative reference to be informative.

B.3. since draft-duke-quic-natsupp-00

- * Tightened NAT terminology
- * Added additional clarifying examples
- * Added warning against using QUIC-LB for NATs that front clients.

Author's Address

Martin Duke
F5 Networks, Inc.

Email: martin.h.duke@gmail.com

QUIC
Internet-Draft
Intended status: Experimental
Expires: 30 October 2022

M. Duke
Google
28 April 2022

QUIC Version Aliasing
draft-duke-quic-version-aliasing-08

Abstract

The QUIC transport protocol preserves its future extensibility partly by specifying its version number. There will be a relatively small number of published version numbers for the foreseeable future. This document provides a method for clients and servers to negotiate the use of other version numbers in subsequent connections and encrypts Initial Packets using secret keys instead of standard ones. If a sizeable subset of QUIC connections use this mechanism, this should prevent middlebox ossification around the current set of published version numbers and the contents of QUIC Initial packets, as well as improving the protocol's privacy properties.

Discussion Venues

This note is to be removed before publishing as an RFC.

Discussion of this document takes place on the mailing list (quic@ietf.org), which is archived at <https://mailarchive.ietf.org/arch/browse/quic/>.

Source for this draft and an issue tracker can be found at <https://github.com/martinduke/quic-version-aliasing>.

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 30 October 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this 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	3
1.1. Terminology	4
2. Protocol Overview	5
2.1. Relationship to ECH and QUIC Protected Initials	6
3. The Version Alias Transport Parameter	7
3.1. Aliased Version Number Generation	7
3.2. Initial Token Extension (ITE) Generation	7
3.3. Salt and Packet Length Offset Generation	8
3.4. Packet Type Generation	8
3.5. Standard Version Number	9
3.6. Expiration Time	9
3.7. Format	10
3.8. Multiple Servers for One Domain	11
3.9. Multiple Entities With One Load Balancer	11
4. Client Behavior	12
4.1. The <code>aliasing_parameters</code> Transport Parameter	13
5. Server Actions on Aliased Version Numbers	14
6. Fallback	15
6.1. Bad Salt Packets	15
6.2. Client Response to Bad Salt	17
6.3. <code>version_aliasing_fallback</code> Transport Parameter	17
6.4. Server Response to <code>version_aliasing_fallback</code> Transport Parameter	18
7. Considerations for Retry Packets	19
8. Security and Privacy Considerations	19
8.1. Endpoint Impersonation	19
8.2. First-Connection Privacy	20
8.3. Forcing Downgrade	20
8.4. Initial Packet Injection	21
8.5. Retry Injection	21

8.6.	Increased Linkability	22
8.7.	Salt Polling	22
8.8.	Server Fingerprinting	22
8.9.	Increased Processing of Garbage UDP Packets	23
8.10.	Increased Retry Overhead	23
8.11.	Request Forgery	23
9.	IANA Considerations	23
9.1.	QUIC Version Registry	23
9.2.	QUIC Transport Parameter Registry	24
9.3.	QUIC Transport Error Codes Registry	24
10.	References	24
10.1.	Normative References	24
10.2.	Informative References	25
Appendix A.	Acknowledgments	25
Appendix B.	Change Log	25
B.1.	since draft-duke-quic-version-aliasing-07	25
B.2.	since draft-duke-quic-version-aliasing-05	26
B.3.	since draft-duke-quic-version-aliasing-04	26
B.4.	since draft-duke-quic-version-aliasing-03	26
B.5.	since draft-duke-quic-version-aliasing-02	26
B.6.	since draft-duke-quic-version-aliasing-01	26
B.7.	since draft-duke-quic-version-aliasing-00	26
Author's Address	27

1. Introduction

The QUIC version number is critical to future extensibility of the protocol ([RFC9000]). Past experience with other protocols, such as TLS1.3 [RFC8446], shows that middleboxes might attempt to enforce that QUIC packets use versions known at the time the middlebox was implemented. This deters deployment of experimental and standard versions on the internet.

Each version of QUIC has a "salt" [RFC9001] that is used to derive the keys used to encrypt Initial packets. As each salt is published in a standards document, any observer can decrypt these packets and inspect the contents, including a TLS Client Hello. A subsidiary mechanism like Encrypted Client Hello [ECHO] might protect some of the TLS fields inside a TLS Client Hello.

This document proposes "QUIC Version Aliasing," a standard way for servers to advertise the availability of other versions inside the cryptographic protection of a QUIC handshake. These versions are syntactically identical to the QUIC version in which the communication takes place, but use a different salt. In subsequent communications, the client uses the new version number and encrypts its Initial packets with a key derived from the provided salt. These version numbers and salts are unique to the client.

If a large subset of QUIC traffic adopts this technique, middleboxes will be unable to enforce particular version numbers or policy based on Client Hello contents without incurring unacceptable penalties on users. This would simultaneously protect the protocol against ossification and improve its privacy properties.

1.1. 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].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying significance described in RFC 2119.

A "standard version" is a QUIC version that would be advertised in a QUIC version negotiation and conforms to a specification. Any aliased version corresponds to a standard version in all its formats and behaviors, except for the version number field in long headers. To be compatible with version aliasing, there MUST be no more than four long header packet types, and the first client packet in a standard version MUST encode the token as if it were a QUIC version 1 initial packet. That is:

- * The most significant bit MUST be 1.
- * The first field after the Source Connection ID MUST be a variable-length integer including the length of a token.
- * The second field after the Destination Connection ID MUST be a field, with length indicated by the previous field, that contains opaque data generated by the server.
- * There must be a variable-length integer that encodes the packet length, unprotected in the header.

An "aliased version" is a version with a number generated in accordance with this document. Except for the version field in long headers, it conforms entirely to the specification of the standard version.

2. Protocol Overview

When they instantiate a connection, servers select an alternate 32-bit version number, and optionally an initial token extension, for the next connection at random and securely derive a salt, packet Length Offset, and long header packet type codepoints from those values using a repeatable process. They communicate this using a transport parameter extension including the version, initial token extension, Initial salt, Packet Length Offset, packet type codepoints, and an expiration time for that value.

If a client next connects to that server within the indicated expiration time, it MAY use the provided version number and encrypt its Initial Packets using a key derived from the provided salt. It uses the provided Initial packet codepoint. It adds the Packet Length Offset to the true packet length when encoding it in the long header. If the server provided an Initial Token Extension, the client puts it in the Initial Packet token field. If there is another token the client wishes to include, it appends the Initial Token Extension to that token. The server can reconstruct the salt and Packet Length Offset from the requested version and token, and proceed with the connection normally.

The Packet Length Offset provides a low-cost way for the server to verify it can derive a valid salt from the inputs without trial decryption. This has important security implications, as described in Section 8.5.

When generating a salt and Packet Length Offset, servers can choose between doing so randomly and storing the mapping, or using a cryptographic process to transform the aliased version number and token extension into the salt. The two options provide a simple tradeoff between computational complexity and storage requirements.

Long header packets are composed identically to their standard version, except that they use the provided packet type codepoint, version number, and packet length offset. Initial packets additionally use any provided token extension and are encrypted as described below.

Short header packets are unchanged when using this extension.

2.1. Relationship to ECH and QUIC Protected Initials

The TLS Encrypted Client Hello [ECH] shares some goals with this document. It encodes an "inner" encrypted Client Hello in a TLS extension in an "outer" Client Hello. The encryption uses asymmetric keys with the server's public key distributed via an out-of-band mechanism like DNS. The inner Client Hello contains any privacy-sensitive information and is only readable with the server's private key.

Significantly, unlike QUIC Version Aliasing, ECH can operate on the first connection between a client and server. However, from the second connection QUIC version aliasing provides additional benefits. It:

- * greases QUIC header fields and packet formats;
- * protects all of the TLS Client Hello and Server Hello;
- * mitigates Retry injection attacks;
- * does not require a mechanism to distribute the public key;
- * uses smaller Client Hello messages, which might allow a larger ORTT packet in the same datagram; and
- * relies on computationally cheap symmetric encryption.

If ECH is operating in "Split Mode", where a client-facing server is using the SNI information to route to a backend server, the client-facing server MUST have the cryptographic context relevant to version aliasing at the backend server to successfully extract the SNI for routing purposes. Furthermore, either all backend servers must share this context, or the client-facing server must decrypt the incoming packet with all possible derived salts.

Note that in the event of the server losing state, the two approaches have a similar fallback: ECH uses information in the outer Client Hello, and Version Aliasing requires a connection using a standard version. In either case, maintaining privacy requires the outer or standard version Client Hello to exclude privacy-sensitive information. However, ECH will allow confidential transmission of data in 1 RTT, while Version Aliasing requires 2 RTTs to resume. This mechanism is also relevant to mitigation of downgrade attacks (see Section 8.3).

Similarly, the QUIC Protected Initial [QUIC-PI] uses the ECH distribution mechanism to generate secure initial keys and Retry integrity tags. While still dependent on a key distribution system, asymmetric encryption, and relatively large Initial packets, it offers similar protection properties to Version Aliasing while still not greasing the version field.

A maximally privacy-protecting client might use Protected Initials for any connection attempts for which it does not have an unexpired aliased version, and QUIC version aliasing otherwise.

See also section 1.1 of [QUIC-PI] for further discussion of tradeoffs.

3. The Version Alias Transport Parameter

3.1. Aliased Version Number Generation

Servers MUST use a random process to generate version numbers. This version number MUST NOT correspond to a QUIC version the server advertises in QUIC Version Negotiation packets or transport parameters. Servers SHOULD also exclude version numbers used in known specifications or experiments to avoid confusion at clients, whether or not they have plans to support those specifications.

Servers MAY use version numbers reserved for grease in Section 15.1 of [RFC9000], even though they might be advertised in Version Negotiation Packets.

Servers MUST NOT use client-controlled information (e.g. the client IP address) in the random process, see Section 8.7.

Servers MUST NOT advertise these versions in QUIC Version Negotiation packets.

3.2. Initial Token Extension (ITE) Generation

Servers SHOULD generate an Initial Token Extension (ITE) to provide additional entropy in salt generation. Two clients that receive the same version number but different extensions will not be able to decode each other's Initial Packets.

Servers MAY choose any length that will allow client Initial Packets to fit within the minimum QUIC packet size of 1200 octets. A four-octet extension is RECOMMENDED. The ITE MUST appear to be random to observers.

The server MUST be able to distinguish ITEs from Resumption and Retry tokens in incoming Initial Packets that contain an aliased version number. As the server controls the lengths and encoding of each, there are many ways to guarantee this.

3.3. Salt and Packet Length Offset Generation

The salt is an opaque 20-octet field. It is used to generate Initial connection keys using the process described in [RFC9001].

The Packet Length Offset is a 64-bit unsigned integer with a maximum value of $2^{62} - 1$.

To reduce header overhead, servers MAY consistently use a Packet Length Offset of zero if and only if it either (1) never sends Retry packets, or (2) can guarantee, through the use of persistent storage or other means, that it will never lose the cryptographic state required to generate the salt before the promised expiration time. Section 8.5 describes the implications if it uses zero without meeting these conditions.

Servers MUST either generate a random salt and Packet Length Offset and store a mapping of aliased version and ITE to salt and offset, or generate the salt and offset using a cryptographic method that uses the version number, ITE, and only server state that is persistent across connections.

If the latter, servers MUST implement a method that it can repeat deterministically at a later time to derive the salt and offset from the incoming version number and ITE. It MUST NOT use client controlled information other than the version number and ITE; for example, the client's IP address and port.

3.4. Packet Type Generation

The server generates the packet type codepoint for each of the four long header packet types (Initial, 0RTT, Handshake, and Retry). Each of these codepoints is two bits.

Future versions of QUIC with 4 or fewer long header packet types can specify a mapping of these fields to their types.

Note that the server needs to derive the type codepoints solely from the version number. It cannot extract the token, and the token extension, until the packet is identified as an Initial packet.

A straightforward implementation might take arbitrary bits from a hash of the version number. The first two bits it reads are the codepoint for Initial packets. The next pair of bits that is not a duplicate of the first is the codepoint for 0RTT packets. The next pair that does not duplicate the first two is the codepoint for Handshake packets, and the remaining codepoint is the Retry packet.

3.5. Standard Version Number

Servers also specify the Standard version that the client should use to guide the wire formats and behaviors of the aliased version. This version **MUST** meet the criteria to support version aliasing, and **MUST** either be included as a supported version in the client's `version_information` transport parameter (see [I-D.ietf-quic-version-negotiation]) or be the standard version of the current connection.

Note that servers **MUST NOT** accept resumption tickets or `NEW_TOKEN` tokens from different standard versions. Therefore, the choice of standard version might impact the performance of the connection that uses an aliased version. The standard version that generated tickets and/or tokens is typically encoded in those tickets or tokens.

There are several possible techniques for the server securely recovering the standard version in use for an aliased connection:

- * the server could store a mapping of aliased versions to standard version;
- * the server could encrypt the standard version in use in the aliased version number (note that the ITE cannot be extracted until the standard version in use is known);
- * the server only accepts one standard version for aliased versions; or
- * the standard version is included as an input to the parameter generation algorithm, and the server tries all supported standard versions and tests each resulting Packet Length Offset for validity.

3.6. Expiration Time

Servers should select an expiration time in seconds, measured from the instant the transport parameter is first sent. This time **SHOULD** be less than the time until the server expects to support new QUIC versions, rotate the keys used to encode information in the version number, or rotate the keys used in salt generation.

Furthermore, the expiration time SHOULD be short enough to frustrate a salt polling attack (Section 8.7)

Conversely, an extremely short expiration time will often force the client to use standard QUIC version numbers and salts.

3.7. Format

This document defines a new transport parameter extension for QUIC with provisional identifier 0x5641. The contents of the value field are indicated below.

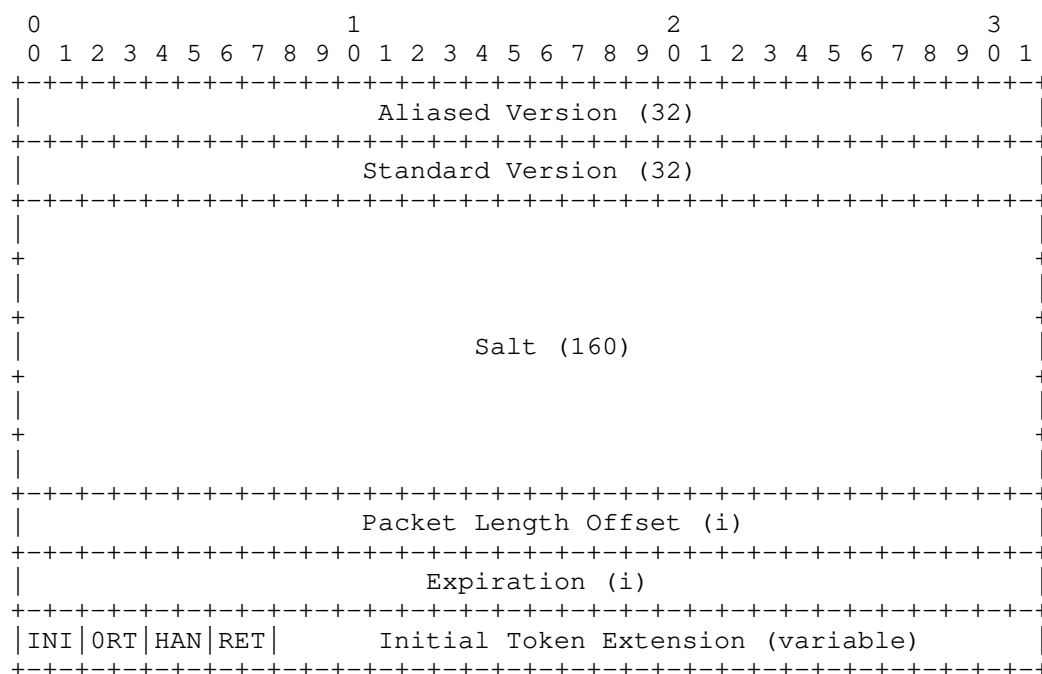


Figure 1: Version Alias Transport Parameter value

The definition of the fields is described above. Note that the "Expiration" field is in seconds, and its length is encoded using the Variable Length Integer encoding from Section 16 of [RFC9000].

The Packet Length Offset is also encoded as a Variable Length Integer.

INI, ORT, HAN, and RET are the codepoints for each long header packet type. If any two packet types have the same codepoint, the transport parameter is invalid.

Clients can compute the length of the Initial Token Extension by subtracting known and encoded field lengths from the overall transport parameter length.

Note that servers that support version aliasing need not send the transport parameter on every connection. Therefore, a client MAY attempt to connect with an unexpired aliased version, even if in its most recent connection it did not receive the transport parameter.

Clients MAY remember the values in this transport parameter for future connections. Servers MUST either store the contents of the transport parameter, or preserve the state to compute the full contents based on what the client provides.

A server that receives this transport parameter MUST close the connection with a `TRANSPORT_PARAMETER_ERROR`.

3.8. Multiple Servers for One Domain

If multiple servers serve the same entity behind a load balancer, all such servers SHOULD either have a common configuration for encoding standard versions and computing salts, or share a common database of mappings. They MUST NOT generate version numbers that any of them would advertise in a Version Negotiation Packet or Transport Parameter.

3.9. Multiple Entities With One Load Balancer

If mutually mistrustful entities share the same IP address and port, incoming packets are usually routed by examining the SNI at a load balancer server that routes the traffic. This use case makes concealing the contents of the Client Initial especially attractive, as the IP address reveals less information. There are several solutions to solve this problem.

- * All entities have a common cryptographic context for deriving salts and Packet Length Offsets from the version number and ITE. This is straightforward but also increases the risk that the keys will leak to an attacker which could then decode Initial packets from point where the packets are observable. This is therefore NOT RECOMMENDED.
- * Each entity has its own cryptographic context, shared with the load balancer. This requires the load balancer to trial decrypt each incoming Initial with each context. As there is no standard algorithm for encoding information in the Version and ITE, this involves synchronizing the method, not just the key material.

- * Each entity reports its Version Aliasing Transport Parameters to the load balancer out-of-band.
- * Each entity is assigned certain version numbers for use. This assignment SHOULD NOT follow observable patterns (e.g., assigning ranges to each entity), as this would allow observers to obtain the target server based on the version. The scheme SHOULD assign all available version numbers to maximize the entropy of the encoding.

Note that [ECHO] and [QUIC-PI] solve this problem elegantly by only holding the private key at the load balancer, which decodes the sensitive information on behalf of the back-end server.

4. Client Behavior

When a client receives the Version Alias Transport Parameter, it MAY cache the version number, ITE, salt, Packet Length Offset, packet type codepoints, and the expiration of these values. It MAY use the version number and ITE in a subsequent connection and compute the initial keys using the provided salt.

The Client MUST NOT use the contents of a Version Alias transport parameter if the handshake does not (1) later authenticate the server name or (2) result in both endpoints computing the same 1-RTT keys. See Section 8.1. The authenticated server name MAY be a "public name" distributed as described in [ECHO] rather than the true target domain.

Clients MUST NOT advertise aliased versions in the Version Negotiation Transport Parameter unless they support a standard version with the same number. Including that number signals support for the standard version, not the aliased version.

Clients SHOULD NOT attempt to use the provided version number and salt after the provided Expiration time has elapsed.

Clients MAY decline to use the provided version number or salt in more than one connection. It SHOULD do so if its IP address has changed between two connection attempts. Using a consistent version number can link the client across connection attempts.

Clients MUST use the same standard version to format the Initial Packet as the standard version used in the connection that provided the aliased version.

Clients MUST use the provided codepoints to encode the packet type.

If the server provided an ITE, the client MUST append it to any Initial Packet token it is including from a Retry packet or NEW_TOKEN frame, if it is using the associated aliased version. If there is no such token, it simply includes the ITE as the entire token.

When using an aliased version, the client MUST include a `aliasing_parameters` transport parameter in its Client Hello.

The QUIC Token Length field MUST include the length of both any Retry or NEW_TOKEN token and the ITE.

The Length fields of all Initial, Handshake, and 0-RTT packets in the connection are set to the value described in [RFC9000] plus the provided Packet Length Offset, modulo 2^{62} .

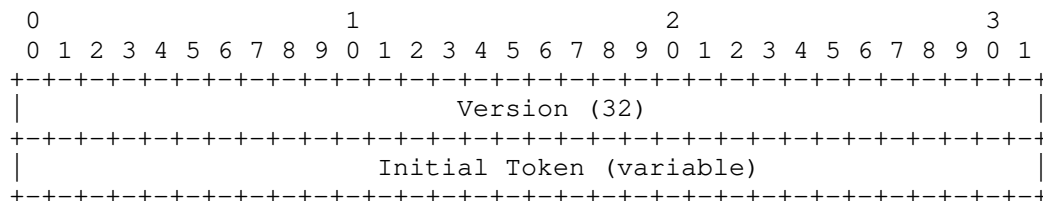
If a client receives an aliased version number that matches a standard version that the client supports, it SHOULD assume the server does not support the standard version and MUST use aliased version behaviors in any connection with the server using that version number.

If the response to an Initial packet using the provided version is a Version Negotiation Packet, the client SHOULD assume that the server no longer supports version aliasing and attempt to connect with one of the advertised versions (while observing the considerations in Section 8.3).

If the response to an Initial packet is a Bad Salt packet, the client follows the procedures in Section 6.

4.1. The `aliasing_parameters` Transport Parameter

This transport parameter has the following format. Its provisional type is 0x4150.



The Version field matches the one in the packet header.

The Initial Token field matches the Initial Token in the packet header, including any Retry token, NEW_TOKEN token, and Initial Token Extension. Its length is inferred from the specified length of the parameter.

The purpose of this parameter is to validate the contents of these header fields by including it in the TLS handshake transcript.

A client that receives this transport parameter MUST close the connection with a `TRANSPORT_PARAMETER_ERROR`.

5. Server Actions on Aliased Version Numbers

When a server receives a packet with an unsupported version number, it SHOULD send a Version Negotiation Packet if it is configured not to generate that version number at random.

Otherwise, when a server receives the first long header packet with an unsupported version number, it hashes that version number to obtain the packet type mapping. If the packet is Handshake or Retry, there may have been a loss of relevant server state; the server discards the packet and SHOULD follow the procedure in Section 6. If ORTT, the server MAY either buffer it in anticipation of a later Initial, or immediately follow the procedure in Section 6. If buffering, and an Initial packet never arrives, the server SHOULD follow the procedure in Section 6 when discarding any ORTT packets.

For an Initial packet, it extracts the ITE, if any, and either looks up the corresponding salt in its database or computes it using the technique originally used to derive the salt from the version number and ITE.

The server similarly obtains the Packet Length Offset and subtracts it from the provided Length field, modulo 2^{62} . If the resulting value is larger than the entire UDP datagram, the server discards the packet and SHOULD follow the procedure in Section 6. The server MAY apply further checks (e.g. against the minimum QUIC packet length) to further reduce the very small probability of a false positive.

If the server supports multiple standard versions, it uses the standard version extracted by the ITE or stored in the mapping to parse the decrypted packet.

In all packets with long headers, the server uses the aliased version number and adds the Packet Length Offset to the length field.

In the extremely unlikely event that the Packet Length Offset resulted in a legal value but the salt is incorrect, the packet may fail authentication. The server should drop these packets in case this is the result of packet corruption along the path.

To reduce linkability for the client, servers SHOULD provide a new Version Alias transport parameter, with a new version number, ITE, salt, and Packet Length Offset, each time a client connects. However, issuing version numbers to a client SHOULD be rate-limited to mitigate the salt polling attack Section 8.7 and MAY cease to clients that are consistently connecting with standard versions.

If there is no `aliasing_parameters` transport parameter, or the contents do not match the fields in the Initial header, the server MUST terminate the connection with a `TRANSPORT_PARAMETER_ERROR`.

6. Fallback

If the server has lost its encryption state, it may not be able to generate the correct salts from previously provided versions and ITEs. The fallback mechanism provides a means of recovering from this state while protecting against injection of messages by attackers.

When the packet length computation in Section 5 fails, it signals either that the packet has been corrupted in transit, or the client is using a transport parameter issued before a server failure. In either case, the server sends a Bad Salt packet.

6.1. Bad Salt Packets

The Bad Salt packet has a long header and a reserved version number, because it must not be confused with a legitimate packet in any standard version. They are not encrypted, not authenticated, and have the following format:

```
Bad Salt Packet {  
    Header Form (1) = 1,  
    Unused (7),  
    Version (32) = TBD (provisional value = 0x56415641),  
    Destination Connection ID Length (8),  
    Destination Connection ID (0..2040),  
    Source Connection ID Length (8),  
    Source Connection ID (0..2040),  
    Supported Version (32) ...,  
    Integrity Tag (128),  
}
```

Unused: The unused field is filled randomly by the sender and ignored on receipt.

Version: The version field is reserved for use by the Bad Salt packet.

Destination and Source Connection IDs and Lengths: These fields are copied from the client packet, with the source fields from the client packet written into the destination fields of the Bad Salt, and vice versa.

Supported Version: A list of standard QUIC version numbers which the server supports. The number of versions is inferred from the length of the datagram.

Integrity Tag: To compute the integrity tag, the server creates a pseudo-packet by contents of the entire client Initial UDP payload, including any coalesced packets, with the Bad Salt packet:

```
Bad Salt Pseudo-Packet {  
  Client UDP Payload (9600...),  
  Header Form (1) = 1,  
  Unused (7),  
  Version (32) = TBD (provisional value = 0x56415641),  
  Destination Connection ID Length (8),  
  Destination Connection ID (0..2040),  
  Source Connection ID Length (8),  
  Source Connection ID (0..2040),  
  Supported Version (32) ...,  
}
```

In a process similar to the Retry Integrity Tag, the Bad Salt Integrity Tag is computed as the output of AEAD_AES_128_GCM with the following inputs:

- * The secret key, K, is 0xbe0c690b9f66575a1d766b54e368c84e.
- * The nonce, N, is 0x461599d35d632bf2239825bb.
- * The plaintext, P, is empty.
- * The associated data, A, is the Bad Salt pseudo-packet.

These values are derived using HKDF-Expand-Label from the secret 0x767fedaff519a2aad117d8fd3ce0a04178ed205ab0d43425723e436853c4b3e2 and labels "quicva key" and "quicva iv".

The integrity tag serves to validate the integrity of both the Bad Salt packet itself and the Initial packet that triggered it.

6.2. Client Response to Bad Salt

Upon receipt of a Bad Salt packet, the client SHOULD wait for a Probe Timeout (PTO) to check if the Bad Salt packet was injected by an attacker, and a valid response arrives from the actual server.

After waiting, the client checks the Integrity Tag using its record of the Initial it sent. If this fails, the client SHOULD assume packet corruption and resend the Initial packet.

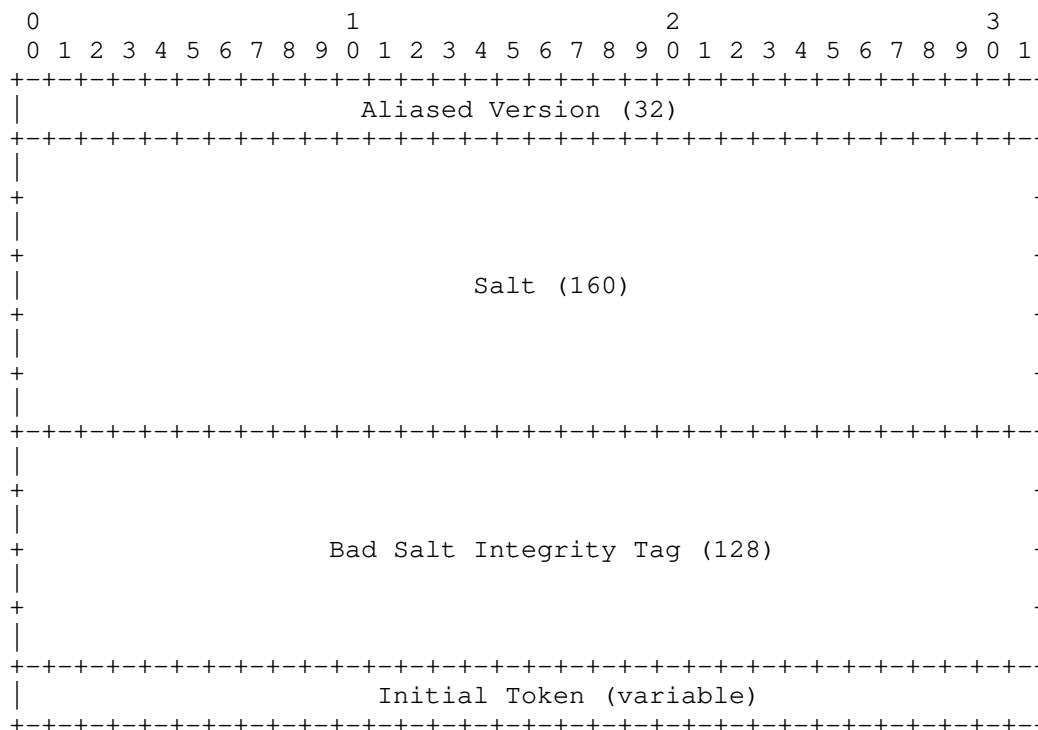
If the verification succeeds, the client SHOULD attempt to connect with one of the listed standard versions. It SHOULD observe the privacy considerations in Section 8.2. It MUST include a `version_aliasing_fallback` Transport Parameter in the Client Hello.

Once it sends this transport parameter, the client MUST NOT attempt to connect with that aliased version again.

The original Client Initial is not part of the new connection. Therefore, the Connection IDs can change, and the original client hello is not part of the transcript for TLS key derivation.

6.3. `version_aliasing_fallback` Transport Parameter

The client sends this transport parameter in a TLS Client Hello generated in response to a Bad Salt packet:



The Aliased Version, Salt, and Initial Token fields are taken from the connection attempt that triggered this fallback. The length of the Initial Token is inferred from the Transport Parameter's overall length.

The Bad Salt Integrity Tag comes from is taken from the Bad Salt packet that triggered this fallback. Its purpose is to include the Bad Salt packet contents in the TLS handshake hash.

6.4. Server Response to `version_aliasing_fallback` Transport Parameter

A client `version_aliasing_fallback` transport parameter tells the server that the client received a Bad Salt packet. The server checks if using the version and ITE as inputs results in the same salt.

If the salt does not match, the server **SHOULD** continue with the connection and **SHOULD** issue a new `version_aliasing` transport parameter.

If the salt and Packet Length Offset are valid, the server **MUST** terminate the connection with the error code `INVALID_BAD_SALT`.

Note that the client never sends this transport parameter with an aliased version. A server that receives such a packet MUST terminate the connection with a `TRANSPORT_PARAMETER_ERROR`.

7. Considerations for Retry Packets

QUIC Retry packets reduce the load on servers during periods of stress by forcing the client to prove it possesses the IP address before the server decrypts any Initial Packets or establishes any connection state. Version aliasing substantially complicates the process.

If a server has to send a Retry packet, the required format is ambiguous without understanding which standard version to use. If all supported standard versions use the same Retry format, it simply uses that format with the client-provided version number.

If the supported standard versions use different Retry formats, the server obtains the standard version via lookup or decoding and formats a Retry containing the aliased version number accordingly.

Servers generate the Retry Integrity Tag of a Retry Packet using the procedure in Section 5.8 of [RFC9001]. However, for aliased versions, the secret key *K* uses the first 16 octets of the aliased salt instead of the key provided in the specification.

Clients MUST ignore Retry packets that contain a QUIC version other than the version it used in its Initial Packet.

Servers MUST NOT reply to a packet with an incorrect Length field in its long header with a Retry packet; it SHOULD reply with Bad Salt as described above.

8. Security and Privacy Considerations

This document intends to improve the existing security and privacy properties of QUIC by dramatically improving the secrecy of QUIC Initial Packets. However, there are new attacks against this mechanism.

8.1. Endpoint Impersonation

An on-path attacker might respond to an Initial Packet with a standard version with a Version Aliasing Transport Parameter that then caused the client to reveal sensitive information in a subsequent Initial.

As described in Section 4, clients cannot use the contents of a Version Aliasing transport parameter until they have authenticated the source as a trusted domain, and have verified that the 1RTT key derivation is identical at both endpoints.

8.2. First-Connection Privacy

As version aliasing requires one connection over a standard QUIC version to acquire initial state, this initial connection leaks some information about the true target.

The client MAY alter its Initial Packet to sanitize sensitive information and obtain another aliased version before proceeding with its true request. However, the client Initial must lead to the authentication of a domain name the client trusts to provide accurate Version Aliasing information (possibly the `public_name` from an Encrypted Client Hello configuration from [ECHO]). Advice for the Outer ClientHello in Section 10.5 of [ECHO] applies here.

Endpoints are encouraged to instead use [ECHO] or [QUIC-PI] to increase privacy on the first connection between a client and server.

8.3. Forcing Downgrade

An attacker can attempt to force a client to send an Initial that uses a standard version by injecting a Version Negotiation packet (which implies the server no longer supports aliasing) or a Bad Salt packet (which implies the server has a new cryptographic context).

The weak form of this attack observes the Initial and injects the Version Negotiation or Bad Salt packet, but cannot drop the Initial. To counteract this, a client SHOULD NOT respond to these packets until they have waited for Probe Timeout (PTO) for a valid server Initial to arrive.

The strong form features an attacker that can drop Initial packets. In this case, the client can either abandon the connection attempt or connect with a standard version.

If it connects with a standard version, it should consider the privacy advice in Section 8.2.

Furthermore, if it received a Bad Salt packet, the client sends a Version Aliasing transport parameter to detect the downgrade attack, and the server will terminate the connection if the Bad Salt packet was an attack.

If the client received a Version Negotiation packet, it MUST implement a downgrade detection mechanism such as [I-D.ietf-quic-version-negotiation] or abandon the connection attempt. If it subsequently detects a downgrade detection, or discovers that the server does not support the same mechanism, it terminates the connection attempt.

8.4. Initial Packet Injection

QUIC version 1 handshakes are vulnerable to DoS from observers for the short interval that endpoints keep Initial keys (usually ~1.5 RTTS), since Initial Packets are not authenticated. With version aliasing, attackers do not have the necessary keys to launch such an attack.

8.5. Retry Injection

QUIC Version 1 Retry packets are spoofable, as they follow a fixed format, are sent in plaintext, and the integrity protection uses a widely known key. As a result, QUIC Version 1 has verification mechanisms in subsequent packets of the connection to validate the origin of the Retry.

Version aliasing largely frustrates this attack. As the integrity check key is derived from the secret salt, packets from attackers will fail their integrity check and the client will ignore them.

The Packet Length Offset is important in this framework. Without this mechanism, servers would have to perform trial decryption to verify the client was using the correct salt. As this does not occur before sending Retry Packets, servers would not detect disagreement on the salt beforehand and would send a Retry packet signed with a different salt than the client expects. Therefore, a client that received a Retry packet with an invalid integrity check would not be able to distinguish between the following possibilities:

- * a Retry packet corrupted in the network, which should be ignored;
- * a Retry packet generated by an attacker, which should be ignored;
or
- * a Retry packet from a server that lost its cryptographic state, meaning that further communication with aliased versions is impossible and the client should revert to using a standard version.

The Packet Length Offset introduces sufficient entropy to make the third possibility exceedingly unlikely.

8.6. Increased Linkability

As each version number and ITE is unique to each client, if a client uses one twice, those two connections are extremely likely to be from the same host. If the client has changed IP address, this is a significant increase in linkability relative to QUIC with a standard version numbers.

8.7. Salt Polling

Observers that wish to decode Initial Packets might open a large number of connections to the server in an effort to obtain part of the mapping of version numbers and ITEs to salts for a server. While storage-intensive, this attack could increase the probability that at least some version-aliased connections are observable. There are three mitigations servers can execute against this attack:

- * use a longer ITE to increase the entropy of the salt,
- * rate-limit transport parameters sent to a particular client, and/or
- * set a low expiration time to reduce the lifetime of the attacker's database.

Segmenting the version number space based on client information, i.e. using only a subset of version numbers for a certain IP address range, would significantly amplify an attack. Observers will generally be on the path to the client and be able to mimic having an identical IP address. Segmentation in this way would dramatically reduce the search space for attackers. Thus, servers are prohibited from using this mechanism.

8.8. Server Fingerprinting

The server chooses its own ITE length, and the length of this ITE is likely to be discoverable to an observer. Therefore, the destination server of a client Initial packet might be decipherable with an ITE length along with other observables. A four-octet ITE is RECOMMENDED. Deviations from this value should be carefully considered in light of this property.

Servers with acute needs for higher or lower entropy than provided by a four-octet ITE are RECOMMENDED to converge on common lengths to reduce the uniqueness of their signatures.

8.9. Increased Processing of Garbage UDP Packets

As QUIC shares the UDP protocol number with other UDP applications, in some deployments it may be possible for traffic intended for other UDP applications to arrive at a QUIC server endpoint. When servers support a finite set of version numbers, a valid version number field is a strong indicator the packet is, in fact, QUIC. If the version number is invalid, a QUIC Version Negotiation is a low-cost response that triggers very early in packet processing.

However, a server that provides version aliasing is prepared to accept almost any version number. As a result, many more sufficiently sized UDP payloads with the first bit set to '1' are potential QUIC Initial Packets that require computation of a salt and Packet Length Offset.

Note that a nonzero Packet Length Offset will allow the server to drop all but approximately 1 in every 2^{49} packets, so trial decryption is unnecessary.

While not a more potent attack than simply sending valid Initial Packets, servers may have to provision additional resources to address this possibility.

8.10. Increased Retry Overhead

This document requires two small cryptographic operations to build a Retry packet instead of one, placing more load on servers when already under load.

8.11. Request Forgery

Section 21.4 of [RFC9000] describes the request forgery attack, where a QUIC endpoint can cause its peer to deliver packets to a victim with specific content.

Version aliasing allows the server to specify the contents of the version field and part of the token field in Initial packets sent by the client, potentially increasing the potency of this attack.

9. IANA Considerations

9.1. QUIC Version Registry

This document request that IANA add the following entry to the QUIC version registry:

Value: TBD

Status: permanent

Specification: This document

Change Controller: IETF

Contact: QUIC WG

9.2. QUIC Transport Parameter Registry

This document requests that IANA add the following entries to the QUIC Transport Parameters Registry:

Value	Parameter Name	Specification
TBD	version_aliasing	This Document
TBD	aliasing_parameters	This Document
TBD	version_aliasing_fallback	This Document

Table 1

9.3. QUIC Transport Error Codes Registry

This document requests that IANA add the following entry to the QUIC Transport Error Codes registry:

Value: TBD (provisional: 0x4942)

Code: INVALID_BAD_SALT

10. References

10.1. Normative References

[I-D.ietf-quic-version-negotiation]
 Schinazi, D. and E. Rescorla, "Compatible Version Negotiation for QUIC", Work in Progress, Internet-Draft, draft-ietf-quic-version-negotiation-07, 5 April 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-version-negotiation-07>>.

- [RFC9000] Iyengar, J., Ed. and M. Thomson, Ed., "QUIC: A UDP-Based Multiplexed and Secure Transport", RFC 9000, DOI 10.17487/RFC9000, May 2021, <<https://www.rfc-editor.org/rfc/rfc9000>>.
- [RFC9001] Thomson, M., Ed. and S. Turner, Ed., "Using TLS to Secure QUIC", RFC 9001, DOI 10.17487/RFC9001, May 2021, <<https://www.rfc-editor.org/rfc/rfc9001>>.

10.2. Informative References

- [ECHO] Rescorla, E., Oku, K., Sullivan, N., and C. A. Wood, "TLS Encrypted Client Hello", Work in Progress, Internet-Draft, draft-ietf-tls-esni-14, 13 February 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-tls-esni-14>>.
- [QUIC-PI] Duke, M. and D. Schinazi, "Protected QUIC Initial Packets", Work in Progress, Internet-Draft, draft-duke-quic-protected-initial-04, 27 April 2022, <<https://datatracker.ietf.org/doc/html/draft-duke-quic-protected-initial-04>>.
- [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/rfc/rfc2119>>.
- [RFC8446] Rescorla, E., "The Transport Layer Security (TLS) Protocol Version 1.3", RFC 8446, DOI 10.17487/RFC8446, August 2018, <<https://www.rfc-editor.org/rfc/rfc8446>>.

Appendix A. Acknowledgments

Marten Seemann was the original creator of the version aliasing approach.

Appendix B. Change Log

RFC Editor's Note: Please remove this section prior to publication of a final version of this document.

B.1. since draft-duke-quic-version-aliasing-07

- * Added the Bad Salt Integrity Tag to the transport parameter
- * Greased packet types

- * Allowed the server to specify the standard version to connect with
- B.2. since draft-duke-quic-version-aliasing-05
- * Revised security considerations
 - * Discussed multiple SNIs behind one load balancer
 - * Removed VN from the fallback mechanism
- B.3. since draft-duke-quic-version-aliasing-04
- * Relationship with Encrypted Client Hello (ECH) and QUIC Protected Initials
 - * Corrected statement about version negotiation
- B.4. since draft-duke-quic-version-aliasing-03
- * Discussed request forgery attacks
- B.5. since draft-duke-quic-version-aliasing-02
- * Specified 0RTT status of the transport parameter
- B.6. since draft-duke-quic-version-aliasing-01
- * Fixed all references to "seed" where I meant "salt."
 - * Added the Packet Length Offset, which eliminates Retry Injection Attacks
- B.7. since draft-duke-quic-version-aliasing-00
- * Added "Initial Token Extensions" to increase salt entropy and make salt polling attacks impractical.
 - * Allowed servers to store a mapping of version number and ITE to salt instead.
 - * Made standard version encoding mandatory. This dramatically simplifies the new Retry logic and changes the security model.
 - * Added references to Version Negotiation Transport Parameters.
 - * Extensive readability edit.

Author's Address

Martin Duke
Google
Email: martin.h.duke@gmail.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: 8 October 2022

M. Kuehlewind
Ericsson
B. Trammell
Google
6 April 2022

Applicability of the QUIC Transport Protocol
draft-ietf-quic-applicability-16

Abstract

This document discusses the applicability of the QUIC transport protocol, focusing on caveats impacting application protocol development and deployment over QUIC. Its intended audience is designers of application protocol mappings to QUIC, and implementors of these application protocols.

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 October 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this 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	2
2. The Necessity of Fallback	3
3. Zero RTT	4
3.1. Replay Attacks	4
3.2. Session resumption versus Keep-alive	5
4. Use of Streams	7
4.1. Stream versus Flow Multiplexing	8
4.2. Prioritization	9
4.3. Ordered and Reliable Delivery	9
4.4. Flow Control Deadlocks	10
4.5. Stream Limit Commitments	11
5. Packetization and Latency	12
6. Error Handling	13
7. Acknowledgment Efficiency	14
8. Port Selection and Application Endpoint Discovery	14
8.1. Source Port Selection	15
9. Connection Migration	16
10. Connection Termination	16
11. Information Exposure and the Connection ID	17
11.1. Server-Generated Connection ID	18
11.2. Mitigating Timing Linkability with Connection ID Migration	18
11.3. Using Server Retry for Redirection	19
12. Quality of Service (QoS) and DSCP	19
13. Use of Versions and Cryptographic Handshake	20
14. Enabling New Versions	20
15. Unreliable Datagram Service over QUIC	21
16. IANA Considerations	22
17. Security Considerations	22
18. Contributors	22
19. Acknowledgments	23
20. References	23
20.1. Normative References	23
20.2. Informative References	23
Authors' Addresses	27

1. Introduction

QUIC [QUIC] is a new transport protocol providing a number of advanced features. While initially designed for the HTTP use case, it provides capabilities that can be used with a much wider variety of applications. QUIC is encapsulated in UDP. QUIC version 1 integrates TLS 1.3 [TLS13] to encrypt all payload data and most control information. The version of HTTP that uses QUIC is known as HTTP/3 [QUIC-HTTP].

This document provides guidance for application developers that want to use the QUIC protocol without implementing it on their own. This includes general guidance for applications operating over HTTP/3 or directly over QUIC.

In the following sections we discuss specific caveats to QUIC's applicability, and issues that application developers must consider when using QUIC as a transport for their application.

2. The Necessity of Fallback

QUIC uses UDP as a substrate. This enables userspace implementation and permits traversal of network middleboxes (including NAT) without requiring updates to existing network infrastructure.

Measurement studies have shown between three [Trammell16] and five [Swett16] percent of networks block all UDP traffic, though there is little evidence of other forms of systematic disadvantage to UDP traffic compared to TCP [Edeline16]. This blocking implies that all applications running on top of QUIC must either be prepared to accept connectivity failure on such networks, or be engineered to fall back to some other transport protocol. In the case of HTTP, this fallback is TLS over TCP.

The IETF TAPS specifications [I-D.ietf-taps-arch] describe a system with a common API for multiple protocols. This is particularly relevant for QUIC as it addresses the implications of fallback among multiple protocols.

Specifically, fallback to insecure protocols or to weaker versions of secure protocols needs to be avoided. In general, a application that implements fallback needs to consider the security consequences. A fallback to TCP and TLS exposes control information to modification and manipulation in the network. Further, downgrades to older TLS versions than 1.3, which is used in QUIC version 1, might result in significantly weaker cryptographic protection. For example, the results of protocol negotiation [RFC7301] only have confidentiality protection if TLS 1.3 is used.

These applications must operate, perhaps with impaired functionality, in the absence of features provided by QUIC not present in the fallback protocol. For fallback to TLS over TCP, the most obvious difference is that TCP does not provide stream multiplexing and therefore stream multiplexing would need to be implemented in the application layer if needed. Further, TCP implementations and network paths often do not support the Fast Open option [RFC7413], which enables sending of payload data together with the first control packet of a new connection as also provided by 0-RTT session

resumption in QUIC. Note that there is some evidence of middleboxes blocking SYN data even if TFO was successfully negotiated (see [PaaschNanog]). And even if Fast Open successfully operates end-to-end, it is limited to a single packet of TLS handshake and application data, unlike QUIC 0-RTT.

Moreover, while encryption (in this case TLS) is inseparably integrated with QUIC, TLS negotiation over TCP can be blocked. If TLS over TCP cannot be supported, the connection should be aborted, and the application then ought to present a suitable prompt to the user that secure communication is unavailable.

In summary, any fallback mechanism is likely to impose a degradation of performance and can degrade security; however, fallback must not silently violate the application's expectation of confidentiality or integrity of its payload data.

3. Zero RTT

QUIC provides for 0-RTT connection establishment. Though the same facility exists in TLS 1.3 with TCP, 0-RTT presents opportunities and challenges for applications using QUIC.

A transport protocol that provides 0-RTT connection establishment is qualitatively different than one that does not from the point of view of the application using it. Relative trade-offs between the cost of closing and reopening a connection and trying to keep it open are different; see Section 3.2.

An application needs to deliberately choose to use 0-RTT, as 0-RTT carries a risk of replay attack. Application protocols that use 0-RTT require a profile that describes the types of information that can be safely sent. For HTTP, this profile is described in [HTTP-REPLAY].

3.1. Replay Attacks

Retransmission or (malicious) replay of data contained in 0-RTT packets could cause the server side to receive multiple copies of the same data.

Application data sent by the client in 0-RTT packets could be processed more than once if it is replayed. Applications need to be aware of what is safe to send in 0-RTT. Application protocols that seek to enable the use of 0-RTT need a careful analysis and a description of what can be sent in 0-RTT; see Section 5.6 of [QUIC-TLS].

In some cases, it might be sufficient to limit application data sent in 0-RTT to that which only causes actions at a server that are known to be free of lasting effect. Initiating data retrieval or establishing configuration are examples of actions that could be safe. Idempotent operations - those for which repetition has the same net effect as a single operation - might be safe. However, it is also possible to combine individually idempotent operations into a non-idempotent sequence of operations.

Once a server accepts 0-RTT data there is no means of selectively discarding data that is received. However, protocols can define ways to reject individual actions that might be unsafe if replayed.

Some TLS implementations and deployments might be able to provide partial or even complete replay protection, which could be used to manage replay risk.

3.2. Session resumption versus Keep-alive

Because QUIC is encapsulated in UDP, applications using QUIC must deal with short network idle timeouts. Deployed stateful middleboxes will generally establish state for UDP flows on the first packet sent, and keep state for much shorter idle periods than for TCP. [RFC5382] suggests a TCP idle period of at least 124 minutes, though there is no evidence of widespread implementation of this guideline in the literature. Short network timeout for UDP, however, is well-documented. According to a 2010 study ([Hatonen10]), UDP applications can assume that any NAT binding or other state entry can expire after just thirty seconds of inactivity. Section 3.5 of [RFC8085] further discusses keep-alive intervals for UDP: it requires a minimum value of 15 seconds, but recommends larger values, or omitting keep-alive entirely.

By using a connection ID, QUIC is designed to be robust to NAT address rebinding after a timeout. However, this only helps if one endpoint maintains availability at the address its peer uses, and the peer is the one to send after the timeout occurs.

Some QUIC connections might not be robust to NAT rebinding because the routing infrastructure (in particular, load balancers) uses the address/port four-tuple to direct traffic. Furthermore, middleboxes with functions other than address translation could still affect the path. In particular, some firewalls do not admit server traffic for which the firewall has no recent state for a corresponding packet sent from the client.

QUIC applications can adjust idle periods to manage the risk of timeout. Idle periods and the network idle timeout are distinct from the connection idle timeout, which is defined as the minimum of either endpoint's idle timeout parameter; see Section 10.1 of [QUIC]). There are three options:

- * Ignore the issue, if the application-layer protocol consists only of interactions with no or very short idle periods, or the protocol's resistance to NAT rebinding is sufficient.
- * Ensure there are no long idle periods.
- * Resume the session after a long idle period, using 0-RTT resumption when appropriate.

The first strategy is the easiest, but it only applies to certain applications.

Either the server or the client in a QUIC application can send PING frames as keep-alives, to prevent the connection and any on-path state from timing out. Recommendations for the use of keep-alives are application-specific, mainly depending on the latency requirements and message frequency of the application. In this case, the application mapping must specify whether the client or server is responsible for keeping the application alive. While [Hatonen10] suggests that 30 seconds might be a suitable value for the public Internet when a NAT is on path, larger values are preferable if the deployment can consistently survive NAT rebinding or is known to be in a controlled environment (e.g. data centres) in order to lower network and computational load.

Sending PING frames more frequently than every 30 seconds over long idle periods may result in excessive unproductive traffic in some situations, and to unacceptable power usage for power-constrained (mobile) devices. Additionally, timeouts shorter than 30 seconds can make it harder to handle transient network interruptions, such as VM migration or coverage loss during mobility. See [RFC8085], especially Section 3.5.

Alternatively, the client (but not the server) can use session resumption instead of sending keepalive traffic. In this case, a client that wants to send data to a server over a connection that has been idle longer than the server's idle timeout (available from the `idle_timeout` transport parameter) can simply reconnect. When possible, this reconnection can use 0-RTT session resumption, reducing the latency involved with restarting the connection. Of course, this approach is only valid in cases in which it is safe to use 0-RTT and when the client is the restarting peer.

The tradeoffs between resumption and keep-alives need to be evaluated on a per-application basis. In general, applications should use keep-alives only in circumstances where continued communication is highly likely; [QUIC-HTTP], for instance, recommends using keep-alives only when a request is outstanding.

4. Use of Streams

QUIC's stream multiplexing feature allows applications to run multiple streams over a single connection, without head-of-line blocking between streams. Stream data is carried within frames, where one QUIC packet on the wire can carry one or multiple stream frames.

Streams can be unidirectional or bidirectional, and a stream may be initiated either by client or server. Only the initiator of a unidirectional stream can send data on it.

Streams and connections can each carry a maximum of $2^{62}-1$ bytes in each direction, due to encoding limitations on stream offsets and connection flow control limits. In the presently unlikely event that this limit is reached by an application, a new connection would need to be established.

Streams can be independently opened and closed, gracefully or abruptly. An application can gracefully close the egress direction of a stream by instructing QUIC to send a FIN bit in a STREAM frame. It cannot gracefully close the ingress direction without a peer-generated FIN, much like in TCP. However, an endpoint can abruptly close the egress direction or request that its peer abruptly close the ingress direction; these actions are fully independent of each other.

QUIC does not provide an interface for exceptional handling of any stream. If a stream that is critical for an application is closed, the application can generate error messages on the application layer to inform the other end and/or the higher layer, which can eventually terminate the QUIC connection.

Mapping of application data to streams is application-specific and described for HTTP/3 in [QUIC-HTTP]. There are a few general principles to apply when designing an application's use of streams:

- * A single stream provides ordering. If the application requires certain data to be received in order, that data should be sent on the same stream. There is no guarantee of transmission, reception, or delivery order across streams.

- * Multiple streams provide concurrency. Data that can be processed independently, and therefore would suffer from head of line blocking if forced to be received in order, should be transmitted over separate streams.
- * Streams can provide message orientation, and allow messages to be cancelled. If one message is mapped to a single stream, resetting the stream to expire an unacknowledged message can be used to emulate partial reliability for that message.

If a QUIC receiver has opened the maximum allowed concurrent streams, and the sender indicates that more streams are needed, it does not automatically lead to an increase of the maximum number of streams by the receiver. Therefore, an application can use the maximum number of allowed, currently open, and currently used streams when determining how to map data to streams.

QUIC assigns a numerical identifier to each stream, called the stream ID. While the relationship between these identifiers and stream types is clearly defined in version 1 of QUIC, future versions might change this relationship for various reasons. QUIC implementations should expose the properties of each stream (which endpoint initiated the stream, whether the stream is unidirectional or bidirectional, the stream ID used for the stream); applications should query for these properties rather than attempting to infer them from the stream ID.

The method of allocating stream identifiers to streams opened by the application might vary between transport implementations. Therefore, an application should not assume a particular stream ID will be assigned to a stream that has not yet been allocated. For example, HTTP/3 uses stream IDs to refer to streams that have already been opened, but makes no assumptions about future stream IDs or the way in which they are assigned Section 6 of [QUIC-HTTP]).

4.1. Stream versus Flow Multiplexing

Streams are meaningful only to the application; since stream information is carried inside QUIC's encryption boundary, a given packet exposes no information about which stream(s) are carried within the packet. Therefore, stream multiplexing is not intended to be used for differentiating streams in terms of network treatment. Application traffic requiring different network treatment should therefore be carried over different five-tuples (i.e. multiple QUIC connections). Given QUIC's ability to send application data in the first RTT of a connection (if a previous connection to the same host has been successfully established to provide the necessary credentials), the cost of establishing another connection is

extremely low.

4.2. Prioritization

Stream prioritization is not exposed to either the network or the receiver. Prioritization is managed by the sender, and the QUIC transport should provide an interface for applications to prioritize streams [QUIC]. Applications can implement their own prioritization scheme on top of QUIC: an application protocol that runs on top of QUIC can define explicit messages for signaling priority, such as those defined in [I-D.draft-ietf-httpbis-priority] for HTTP; it can define rules that allow an endpoint to determine priority based on context; or it can provide a higher level interface and leave the determination to the application on top.

Priority handling of retransmissions can be implemented by the sender in the transport layer. [QUIC] recommends retransmitting lost data before new data, unless indicated differently by the application. When a QUIC endpoint uses fully reliable streams for transmission, prioritization of retransmissions will be beneficial in most cases, filling in gaps and freeing up the flow control window. For partially reliable or unreliable streams, priority scheduling of retransmissions over data of higher-priority streams might not be desirable. For such streams, QUIC could either provide an explicit interface to control prioritization, or derive the prioritization decision from the reliability level of the stream.

4.3. Ordered and Reliable Delivery

QUIC streams enable ordered and reliable delivery. Though it is possible for an implementation to provide options that use streams for partial reliability or out-of-order delivery, most implementations will assume that data is reliably delivered in order.

Under this assumption, an endpoint that receives stream data might not make forward progress until data that is contiguous with the start of a stream is available. In particular, a receiver might withhold flow control credit until contiguous data is delivered to the application; see Section 2.2 of [QUIC]. To support this receive logic, an endpoint will send stream data until it is acknowledged, ensuring that data at the start of the stream is sent and acknowledged first.

An endpoint that uses a different sending behavior and does not negotiate that change with its peer might encounter performance issues or deadlocks.

4.4. Flow Control Deadlocks

QUIC flow control Section 4 of [QUIC] provides a means of managing access to the limited buffers endpoints have for incoming data. This mechanism limits the amount of data that can be in buffers in endpoints or in transit on the network. However, there are several ways in which limits can produce conditions that can cause a connection to either perform suboptimally or deadlock.

Deadlocks in flow control are possible for any protocol that uses QUIC, though whether they become a problem depends on how implementations consume data and provide flow control credit. Understanding what causes deadlocking might help implementations avoid deadlocks.

The size and rate of transport flow control credit updates can affect performance. Applications that use QUIC often have a data consumer that reads data from transport buffers. Some implementations might have independent transport-layer and application-layer receive buffers. Consuming data does not always imply it is immediately processed. However, a common flow control implementation technique is to extend credit to the sender, by emitting `MAX_DATA` and/or `MAX_STREAM_DATA` frames, as data is consumed. Delivery of these frames is affected by the latency of the back channel from the receiver to the data sender. If credit is not extended in a timely manner, the sending application can be blocked, effectively throttling the sender.

Large application messages can produce deadlocking if the recipient does not read data from the transport incrementally. If the message is larger than the flow control credit available and the recipient does not release additional flow control credit until the entire message is received and delivered, a deadlock can occur. This is possible even where stream flow control limits are not reached because connection flow control limits can be consumed by other streams.

A length-prefixed message format makes it easier for a data consumer to leave data unread in the transport buffer and thereby withhold flow control credit. If flow control limits prevent the remainder of a message from being sent, a deadlock will result. A length prefix might also enable the detection of this sort of deadlock. Where application protocols have messages that might be processed as a single unit, reserving flow control credit for the entire message atomically makes this style of deadlock less likely.

A data consumer can eagerly read all data as it becomes available, in order to make the receiver extend flow control credit and reduce the chances of a deadlock. However, such a data consumer might need other means for holding a peer accountable for the additional state it keeps for partially processed messages.

Deadlocking can also occur if data on different streams is interdependent. Suppose that data on one stream arrives before the data on a second stream on which it depends. A deadlock can occur if the first stream is left unread, preventing the receiver from extending flow control credit for the second stream. To reduce the likelihood of deadlock for interdependent data, the sender should ensure that dependent data is not sent until the data it depends on has been accounted for in both stream- and connection- level flow control credit.

Some deadlocking scenarios might be resolved by cancelling affected streams with `STOP_SENDING` or `RESET_STREAM`. Cancelling some streams results in the connection being terminated in some protocols.

4.5. Stream Limit Commitments

QUIC endpoints are responsible for communicating the cumulative limit of streams they would allow to be opened by their peer. Initial limits are advertised using the `initial_max_streams_bidi` and `initial_max_streams_uni` transport parameters. As streams are opened and closed they are consumed and the cumulative total is incremented. Limits can be increased using the `MAX_STREAMS` frame but there is no mechanism to reduce limits. Once stream limits are reached, no more streams can be opened, which prevents applications using QUIC from making further progress. At this stage connections can be terminated via idle timeout or explicit close; see Section 10).

An application that uses QUIC and communicated a cumulative stream limit might require the connection to be closed before the limit is reached. For example, to stop the server to perform scheduled maintenance. Immediate connection close causes abrupt closure of actively used streams. Depending on how an application uses QUIC streams, this could be undesirable or detrimental to behavior or performance.

A more graceful closure technique is to stop sending increases to stream limits and allow the connection to naturally terminate once remaining streams are consumed. However, the period of time it takes to do so is dependent on the peer and an unpredictable closing period might not fit application or operational needs. Applications using QUIC can be conservative with open stream limits in order to reduce the commitment and indeterminism. However, being overly conservative with stream limits affects stream concurrency. Balancing these aspects can be specific to applications and their deployments.

Instead of relying on stream limits to avoid abrupt closure, an application-layer graceful close mechanism can be used to communicate the intention to explicitly close the connection at some future point. HTTP/3 provides such a mechanism using the GOAWAY frame. In HTTP/3, when the GOAWAY frame is received by a client, it stops opening new streams even if the cumulative stream limit would allow. Instead, the client would create a new connection on which to open further streams. Once all streams are closed on the old connection, it can be terminated safely by a connection close or after expiration of the idle time out (see also Section 10).

5. Packetization and Latency

QUIC exposes an interface that provides multiple streams to the application; however, the application usually cannot control how data transmitted over those streams is mapped into frames or how those frames are bundled into packets.

By default, many implementations will try to maximally pack QUIC packets DATA frames from one or more streams to minimize bandwidth consumption and computational costs (see Section 13 of [QUIC]). If there is not enough data available to fill a packet, an implementation might wait for a short time, to optimize bandwidth efficiency instead of latency. This delay can either be pre-configured or dynamically adjusted based on the observed sending pattern of the application.

If the application requires low latency, with only small chunks of data to send, it may be valuable to indicate to QUIC that all data should be sent out immediately. Alternatively, if the application expects to use a specific sending pattern, it can also provide a suggested delay to QUIC for how long to wait before bundle frames into a packet.

Similarly, an application has usually no control about the length of a QUIC packet on the wire. QUIC provides the ability to add a PADDING frame to arbitrarily increase the size of packets. Padding is used by QUIC to ensure that the path is capable of transferring

datagrams of at least a certain size, during the handshake (see Sections 8.1 and 14.1 of [QUIC]) and for path validation after connection migration (see Section 8.2 of [QUIC]) as well as for Datagram Packetization Layer PMTU Discovery (DPLMTUD) (see Section 14.3 of [QUIC]).

Padding can also be used by an application to reduce leakage of information about the data that is sent. A QUIC implementation can expose an interface that allows an application layer to specify how to apply padding.

6. Error Handling

QUIC recommends that endpoints signal any detected errors to the peer. Errors can occur at the transport level and the application level. Transport errors, such as a protocol violation, affect the entire connection. Applications that use QUIC can define their own error detection and signaling (see, for example, Section 8 of [QUIC-HTTP]). Application errors can affect an entire connection or a single stream.

QUIC defines an error code space that is used for error handling at the transport layer. QUIC encourages endpoints to use the most specific code, although any applicable code is permitted, including generic ones.

Applications using QUIC define an error code space that is independent from QUIC or other applications (see, for example, Section 8.1 of [QUIC-HTTP]). The values in an application error code space can be reused across connection-level and stream-level errors.

Connection errors lead to connection termination. They are signaled using a CONNECTION_CLOSE frame, which contains an error code and a reason field that can be zero length. Different types of CONNECTION_CLOSE frame are used to signal transport and application errors.

Stream errors lead to stream termination. These are signaled using STOP_SENDING or RESET_STREAM frames, which contain only an error code.

7. Acknowledgment Efficiency

QUIC version 1 without extensions uses an acknowledgment strategy adopted from TCP Section 13.2 of [QUIC]). That is, it recommends every other packet is acknowledged. However, generating and processing QUIC acknowledgments consumes resources at a sender and receiver. Acknowledgments also incur forwarding costs and contribute to link utilization, which can impact performance over some types of network. Applications might be able to improve overall performance by using alternative strategies that reduce the rate of acknowledgments.

8. Port Selection and Application Endpoint Discovery

In general, port numbers serve two purposes: "first, they provide a demultiplexing identifier to differentiate transport sessions between the same pair of endpoints, and second, they may also identify the application protocol and associated service to which processes connect" [RFC6335]. The assumption that an application can be identified in the network based on the port number is less true today due to encapsulation, mechanisms for dynamic port assignments, and NATs.

As QUIC is a general-purpose transport protocol, there are no requirements that servers use a particular UDP port for QUIC. For applications with a fallback to TCP that do not already have an alternate mapping to UDP, usually the registration (if necessary) and use of the UDP port number corresponding to the TCP port already registered for the application is appropriate. For example, the default port for HTTP/3 [QUIC-HTTP] is UDP port 443, analogous to HTTP/1.1 or HTTP/2 over TLS over TCP.

Given the prevalence of the assumption in network management practice that a port number maps unambiguously to an application, the use of ports that cannot easily be mapped to a registered service name might lead to blocking or other changes to the forwarding behavior by network elements such as firewalls that use the port number for application identification.

Applications could define an alternate endpoint discovery mechanism to allow the usage of ports other than the default. For example, HTTP/3 (Sections 3.2 and 3.3 of [QUIC-HTTP]) specifies the use of HTTP Alternative Services [RFC7838] for an HTTP origin to advertise the availability of an equivalent HTTP/3 endpoint on a certain UDP port by using the "h3" Application-Layer Protocol Negotiation (ALPN) [RFC7301] token.

ALPN permits the client and server to negotiate which of several protocols will be used on a given connection. Therefore, multiple applications might be supported on a single UDP port based on the ALPN token offered. Applications using QUIC are required to register an ALPN token for use in the TLS handshake.

As QUIC version 1 deferred defining a complete version negotiation mechanism, HTTP/3 requires QUIC version 1 and defines the ALPN token ("h3") to only apply to that version. So far no single approach has been selected for managing the use of different QUIC versions, neither in HTTP/3 nor in general. Application protocols that use QUIC need to consider how the protocol will manage different QUIC versions. Decisions for those protocols might be informed by choices made by other protocols, like HTTP/3.

8.1. Source Port Selection

Some UDP protocols are vulnerable to reflection attacks, where an attacker is able to direct traffic to a third party as a denial of service. For example, these source ports are associated with applications known to be vulnerable to reflection attacks, often due to server misconfiguration:

- * port 53 - DNS [RFC1034]
- * port 123 - NTP [RFC5905]
- * port 1900 - SSDP [SSDP]
- * port 5353 - mDNS [RFC6762]
- * port 11211 - memcached

Services might block source ports associated with protocols known to be vulnerable to reflection attacks, to avoid the overhead of processing large numbers of packets. However, this practice has negative effects on clients: not only does it require establishment of a new connection, but in some instances, might cause the client to avoid using QUIC for that service for a period of time, downgrading to a non-UDP protocol (see Section 2).

As a result, client implementations are encouraged to avoid using source ports associated with protocols known to be vulnerable to reflection attacks. Note that the list above is not exhaustive; other source ports might be considered reflection vectors as well.

9. Connection Migration

QUIC supports connection migration by the client. If an IP address changes, a QUIC endpoint can still associate packets with an existing transport connection using the Destination Connection ID field (see also Section 11) in the QUIC header. This supports cases where address information changes, such as NAT rebinding, intentional change of the local interface, or based on an indication in the handshake of the server for a preferred address to be used.

Use of a non-zero-length connection ID for the server is strongly recommended if any clients are behind a NAT or could be. A non-zero-length connection ID is also strongly recommended when active migration is supported. If a connection is intentionally migrated to new path, a new connection ID is used to minimize linkability by network observers. The other QUIC endpoint uses the connection ID to link different addresses to the same connection and entity if a non-zero-length connection ID is provided.

The base specification of QUIC version 1 only supports the use of a single network path at a time, which enables failover use cases. Path validation is required so that endpoints validate paths before use to avoid address spoofing attacks. Path validation takes at least one RTT and congestion control will also be reset after path migration. Therefore, migration usually has a performance impact.

QUIC probing packets, which can be sent on multiple paths at once, are used to perform address validation as well as measure path characteristics. Probing packets cannot carry application data but likely contain padding frames. Endpoints can use information about their receipt as input to congestion control for that path. Applications could use information learned from probing to inform a decision to switch paths.

Only the client can actively migrate in version 1 of QUIC. However, servers can indicate during the handshake that they prefer to transfer the connection to a different address after the handshake. For instance, this could be used to move from an address that is shared by multiple servers to an address that is unique to the server instance. The server can provide an IPv4 and an IPv6 address in a transport parameter during the TLS handshake and the client can select between the two if both are provided. See also Section 9.6 of [QUIC].

10. Connection Termination

QUIC connections are terminated in one of three ways: implicit idle timeout, explicit immediate close, or explicit stateless reset.

QUIC does not provide any mechanism for graceful connection termination; applications using QUIC can define their own graceful termination process (see, for example, Section 5.2 of [QUIC-HTTP]).

QUIC idle timeout is enabled via transport parameters. Client and server announce a timeout period and the effective value for the connection is the minimum of the two values. After the timeout period elapses, the connection is silently closed. An application therefore should be able to configure its own maximum value, as well as have access to the computed minimum value for this connection. An application may adjust the maximum idle timeout for new connections based on the number of open or expected connections, since shorter timeout values may free-up resources more quickly.

Application data exchanged on streams or in datagrams defers the QUIC idle timeout. Applications that provide their own keep-alive mechanisms will therefore keep a QUIC connection alive. Applications that do not provide their own keep-alive can use transport-layer mechanisms (see Section 10.1.2 of [QUIC], and Section 3.2). However, QUIC implementation interfaces for controlling such transport behavior can vary, affecting the robustness of such approaches.

An immediate close is signaled by a CONNECTION_CLOSE frame (see Section 6). Immediate close causes all streams to become immediately closed, which may affect applications; see Section 4.5.

A stateless reset is an option of last resort for an endpoint that does not have access to connection state. Receiving a stateless reset is an indication of an unrecoverable error distinct from connection errors in that there is no application-layer information provided.

11. Information Exposure and the Connection ID

QUIC exposes some information to the network in the unencrypted part of the header, either before the encryption context is established or because the information is intended to be used by the network. For more information on manageability of QUIC see also [I-D.ietf-quic-manageability]. QUIC has a long header that exposes some additional information (the version and the source connection ID), while the short header exposes only the destination connection ID. In QUIC version 1, the long header is used during connection establishment, while the short header is used for data transmission in an established connection.

The connection ID can be zero length. Zero length connection IDs can be chosen on each endpoint individually, on any packet except the first packets sent by clients during connection establishment.

An endpoint that selects a zero-length connection ID will receive packets with a zero-length destination connection ID. The endpoint needs to use other information, such as the source and destination IP address and port number to identify which connection is referred to. This could mean that the endpoint is unable to match datagrams to connections successfully if these values change, making the connection effectively unable to survive NAT rebinding or migrate to a new path.

11.1. Server-Generated Connection ID

QUIC supports a server-generated connection ID, transmitted to the client during connection establishment (see Section 7.2 of [QUIC]). Servers behind load balancers may need to change the connection ID during the handshake, encoding the identity of the server or information about its load balancing pool, in order to support stateless load balancing.

Server deployments with load balancers and other routing infrastructure need to ensure that this infrastructure consistently routes packets to the server instance that has the connection state, even if addresses, ports, and/or connection IDs change. This might require coordination between servers and infrastructure. One method of achieving this involves encoding routing information into the connection ID. For an example of this technique, see [QUIC-LB].

11.2. Mitigating Timing Linkability with Connection ID Migration

QUIC requires that endpoints generate fresh connection IDs for use on new network paths. Choosing values that are unlinkable to an outside observer ensures that activity on different paths cannot be trivially correlated using the connection ID.

While sufficiently robust connection ID generation schemes will mitigate linkability issues, they do not provide full protection. Analysis of the lifetimes of six-tuples (source and destination addresses as well as the migrated CID) may expose these links anyway.

In the limit where connection migration in a server pool is rare, it is trivial for an observer to associate two connection IDs. Conversely, in the opposite limit where every server handles multiple simultaneous migrations, even an exposed server mapping may be insufficient information.

The most efficient mitigations for these attacks are through network design and/or operational practice, by using a load balancing architecture that loads more flows onto a single server-side address, by coordinating the timing of migrations in an attempt to increase the number of simultaneous migrations at a given time, or through other means.

11.3. Using Server Retry for Redirection

QUIC provides a Retry packet that can be sent by a server in response to the client Initial packet. The server may choose a new connection ID in that packet and the client will retry by sending another client Initial packet with the server-selected connection ID. This mechanism can be used to redirect a connection to a different server, e.g., due to performance reasons or when servers in a server pool are upgraded gradually, and therefore may support different versions of QUIC.

In this case, it is assumed that all servers belonging to a certain pool are served in cooperation with load balancers that forward the traffic based on the connection ID. A server can choose the connection ID in the Retry packet such that the load balancer will redirect the next Initial packet to a different server in that pool. Alternatively the load balancer can directly offer a Retry service as further described in [QUIC-LB].

Section 4 of [RFC5077] describes an example approach for constructing TLS resumption tickets that can be also applied for validation tokens, however, the use of more modern cryptographic algorithms is highly recommended.

12. Quality of Service (QoS) and DSCP

QUIC, as defined in [QUIC], has a single congestion controller and recovery handler. This design assumes that all packets of a QUIC connection, or at least with the same 5-tuple {dest addr, source addr, protocol, dest port, source port}, that have the same DiffServ Code Point (DSCP) [RFC2475] will receive similar network treatment since feedback about loss or delay of each packet is used as input to the congestion controller. Therefore, packets belonging to the same connection should use a single DSCP. Section 5.1 of [RFC7657] provides a discussion of DiffServ interactions with datagram transport protocols [RFC7657] (in this respect the interactions with QUIC resemble those of SCTP).

When multiplexing multiple flows over a single QUIC connection, the selected DSCP value should be the one associated with the highest priority requested for all multiplexed flows.

If differential network treatment is desired, e.g., by the use of different DSCPs, multiple QUIC connections to the same server may be used. However, in general it is recommended to minimize the number of QUIC connections to the same server, to avoid increased overhead and, more importantly, competing congestion control.

As in other uses of DiffServ, when a packet enters a network segment that does not support the DSCP value, this could result in the connection not receiving the network treatment it expects. The DSCP value in this packet could also be remarked as the packet travels along the network path, changing the requested treatment.

13. Use of Versions and Cryptographic Handshake

Versioning in QUIC may change the protocol's behavior completely, except for the meaning of a few header fields that have been declared to be invariant [QUIC-INVARIANTS]. A version of QUIC with a higher version number will not necessarily provide a better service, but might simply provide a different feature set. As such, an application needs to be able to select which versions of QUIC it wants to use.

A new version could use an encryption scheme other than TLS 1.3 or higher. [QUIC] specifies requirements for the cryptographic handshake as currently realized by TLS 1.3 and described in a separate specification [QUIC-TLS]. This split is performed to enable light-weight versioning with different cryptographic handshakes.

14. Enabling New Versions

QUIC version 1 does not specify a version negotiation mechanism in the base spec but [I-D.draft-ietf-quic-version-negotiation] proposes an extension. This process assumes that the set of versions that a server supports is fixed. This complicates the process for deploying new QUIC versions or disabling old versions when servers operate in clusters.

A server that rolls out a new version of QUIC can do so in three stages. Each stage is completed across all server instances before moving to the next stage.

In the first stage of deployment, all server instances start accepting new connections with the new version. The new version can be enabled progressively across a deployment, which allows for selective testing. This is especially useful when the new version is compatible with an old version, because the new version is more likely to be used.

While enabling the new version, servers do not advertise the new version in any Version Negotiation packets they send. This prevents clients that receive a Version Negotiation packet from attempting to connect to server instances that might not have the new version enabled.

During the initial deployment, some clients will have received Version Negotiation packets that indicate that the server does not support the new version. Other clients might have successfully connected with the new version and so will believe that the server supports the new version. Therefore, servers need to allow for this ambiguity when validating the negotiated version.

The second stage of deployment commences once all server instances are able to accept new connections with the new version. At this point, all servers can start sending the new version in Version Negotiation packets.

During the second stage, the server still allows for the possibility that some clients believe the new version to be available and some do not. This state will persist only for as long as any Version Negotiation packets take to be transmitted and responded to. So the third stage can follow after a relatively short delay.

The third stage completes the process by enabling authentication of the negotiated version with the assumption that the new version is fully available.

The process for disabling an old version or rolling back the introduction of a new version uses the same process in reverse. Servers disable validation of the old version, stop sending the old version in Version Negotiation packets, then the old version is no longer accepted.

15. Unreliable Datagram Service over QUIC

[I-D.ietf-quic-datagram] specifies a QUIC extension to enable sending and receiving unreliable datagrams over QUIC. Unlike operating directly over UDP, applications that use the QUIC datagram service do not need to implement their own congestion control, per [RFC8085], as QUIC datagrams are congestion controlled.

QUIC datagrams are not flow-controlled, and as such data chunks may be dropped if the receiver is overloaded. While the reliable transmission service of QUIC provides a stream-based interface to send and receive data in order over multiple QUIC streams, the datagram service has an unordered message-based interface. If needed, an application layer framing can be used on top to allow separate flows of unreliable datagrams to be multiplexed on one QUIC connection.

16. IANA Considerations

This document has no actions for IANA; however, note that Section 8 recommends that application bindings to QUIC for applications using TCP register UDP ports analogous to their existing TCP registrations.

17. Security Considerations

See the security considerations in [QUIC] and [QUIC-TLS]; the security considerations for the underlying transport protocol are relevant for applications using QUIC, as well. Considerations on linkability, replay attacks, and randomness discussed in [QUIC-TLS] should be taken into account when deploying and using QUIC.

Further, migration to a new address exposes a linkage between client addresses to the server and may expose this linkage also to the path if the connection ID cannot be changed or flows can otherwise be correlated. When migration is supported, this needs to be considered with respect to user privacy.

Application developers should note that any fallback they use when QUIC cannot be used due to network blocking of UDP should guarantee the same security properties as QUIC; if this is not possible, the connection should fail to allow the application to explicitly handle fallback to a less-secure alternative. See Section 2.

Further, [QUIC-HTTP] provides security considerations specific to HTTP. However, discussions such as on cross-protocol attacks, traffic analysis and padding, or migration might be relevant for other applications using QUIC as well.

18. Contributors

The following people have contributed significant text to and/or feedback on this document:

- * Gorrry Fairhurst
- * Ian Swett

- * Igor Lubashev
- * Lucas Pardue
- * Mike Bishop
- * Mark Nottingham
- * Martin Duke
- * Martin Thomson
- * Sean Turner
- * Tommy Pauly

19. Acknowledgments

Special thanks to last-call reviewers Chris Lonvick and Ines Robles.

This work was 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.

20. References

20.1. Normative References

[QUIC] Iyengar, J., Ed. and M. Thomson, Ed., "QUIC: A UDP-Based Multiplexed and Secure Transport", RFC 9000, DOI 10.17487/RFC9000, May 2021, <<https://www.rfc-editor.org/rfc/rfc9000>>.

[QUIC-INVARIANTS] Thomson, M., "Version-Independent Properties of QUIC", RFC 8999, DOI 10.17487/RFC8999, May 2021, <<https://www.rfc-editor.org/rfc/rfc8999>>.

[QUIC-TLS] Thomson, M., Ed. and S. Turner, Ed., "Using TLS to Secure QUIC", RFC 9001, DOI 10.17487/RFC9001, May 2021, <<https://www.rfc-editor.org/rfc/rfc9001>>.

20.2. Informative References

[Edeline16]

Edeline, K., Kuehlewind, M., Trammell, B., Aben, E., and B. Donnet, "Using UDP for Internet Transport Evolution (arXiv preprint 1612.07816)", 22 December 2016, <<https://arxiv.org/abs/1612.07816>>.

[Hatonen10]

Hatonen, S., Nyrhinen, A., Eggert, L., Strowes, S., Sarolahti, P., and M. Kojo, "An experimental study of home gateway characteristics (Proc. ACM IMC 2010)", October 2010.

[HTTP-REPLAY]

Thomson, M., Nottingham, M., and W. Tarreau, "Using Early Data in HTTP", RFC 8470, DOI 10.17487/RFC8470, September 2018, <<https://www.rfc-editor.org/rfc/rfc8470>>.

[I-D.draft-ietf-httpbis-priority]

Oku, K. and L. Pardue, "Extensible Prioritization Scheme for HTTP", Work in Progress, Internet-Draft, draft-ietf-httpbis-priority-12, 17 January 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-httpbis-priority-12>>.

[I-D.draft-ietf-quic-version-negotiation]

Schinazi, D. and E. Rescorla, "Compatible Version Negotiation for QUIC", Work in Progress, Internet-Draft, draft-ietf-quic-version-negotiation-07, 5 April 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-version-negotiation-07>>.

[I-D.ietf-quic-datagram]

Pauly, T., Kinnear, E., and D. Schinazi, "An Unreliable Datagram Extension to QUIC", Work in Progress, Internet-Draft, draft-ietf-quic-datagram-10, 4 February 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-datagram-10>>.

[I-D.ietf-quic-manageability]

Kuehlewind, M. and B. Trammell, "Manageability of the QUIC Transport Protocol", Work in Progress, Internet-Draft, draft-ietf-quic-manageability-15, 7 March 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-manageability-15>>.

[I-D.ietf-taps-arch]

Pauly, T., Trammell, B., Brunstrom, A., Fairhurst, G., and C. Perkins, "An Architecture for Transport Services", Work

in Progress, Internet-Draft, draft-ietf-taps-arch-12, 3 January 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-taps-arch-12>>.

[PaaschNanog]

Paasch, C., "Network Support for TCP Fast Open (NANOG 67 presentation)", 13 June 2016, <https://www.nanog.org/sites/default/files/Paasch_Network_Support.pdf>.

[QUIC-HTTP]

Bishop, M., "Hypertext Transfer Protocol Version 3 (HTTP/3)", Work in Progress, Internet-Draft, draft-ietf-quic-http-34, 2 February 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-http-34>>.

[QUIC-LB]

Duke, M., Banks, N., and C. Huitema, "QUIC-LB: Generating Routable QUIC Connection IDs", Work in Progress, Internet-Draft, draft-ietf-quic-load-balancers-13, 28 March 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-load-balancers-13>>.

[RFC1034]

Mockapetris, P., "Domain names - concepts and facilities", STD 13, RFC 1034, DOI 10.17487/RFC1034, November 1987, <<https://www.rfc-editor.org/rfc/rfc1034>>.

[RFC2475]

Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, DOI 10.17487/RFC2475, December 1998, <<https://www.rfc-editor.org/rfc/rfc2475>>.

[RFC5077]

Salowey, J., Zhou, H., Eronen, P., and H. Tschofenig, "Transport Layer Security (TLS) Session Resumption without Server-Side State", RFC 5077, DOI 10.17487/RFC5077, January 2008, <<https://www.rfc-editor.org/rfc/rfc5077>>.

[RFC5382]

Guha, S., Ed., Biswas, K., Ford, B., Sivakumar, S., and P. Srisuresh, "NAT Behavioral Requirements for TCP", BCP 142, RFC 5382, DOI 10.17487/RFC5382, October 2008, <<https://www.rfc-editor.org/rfc/rfc5382>>.

[RFC5905]

Mills, D., Martin, J., Ed., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, DOI 10.17487/RFC5905, June 2010, <<https://www.rfc-editor.org/rfc/rfc5905>>.

- [RFC6335] Cotton, M., Eggert, L., Touch, J., Westerlund, M., and S. Cheshire, "Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Port Number Registry", BCP 165, RFC 6335, DOI 10.17487/RFC6335, August 2011, <<https://www.rfc-editor.org/rfc/rfc6335>>.
- [RFC6762] Cheshire, S. and M. Krochmal, "Multicast DNS", RFC 6762, DOI 10.17487/RFC6762, February 2013, <<https://www.rfc-editor.org/rfc/rfc6762>>.
- [RFC7301] Friedl, S., Popov, A., Langley, A., and E. Stephan, "Transport Layer Security (TLS) Application-Layer Protocol Negotiation Extension", RFC 7301, DOI 10.17487/RFC7301, July 2014, <<https://www.rfc-editor.org/rfc/rfc7301>>.
- [RFC7413] Cheng, Y., Chu, J., Radhakrishnan, S., and A. Jain, "TCP Fast Open", RFC 7413, DOI 10.17487/RFC7413, December 2014, <<https://www.rfc-editor.org/rfc/rfc7413>>.
- [RFC7657] Black, D., Ed. and P. Jones, "Differentiated Services (Diffserv) and Real-Time Communication", RFC 7657, DOI 10.17487/RFC7657, November 2015, <<https://www.rfc-editor.org/rfc/rfc7657>>.
- [RFC7838] Nottingham, M., McManus, P., and J. Reschke, "HTTP Alternative Services", RFC 7838, DOI 10.17487/RFC7838, April 2016, <<https://www.rfc-editor.org/rfc/rfc7838>>.
- [RFC8085] Eggert, L., Fairhurst, G., and G. Shepherd, "UDP Usage Guidelines", BCP 145, RFC 8085, DOI 10.17487/RFC8085, March 2017, <<https://www.rfc-editor.org/rfc/rfc8085>>.
- [SSDP] Donoho, A., Roe, B., Bodlaender, M., Gildred, J., Messer, A., Kim, Y., Fairman, B., and J. Tourzan, "UPnP Device Architecture 2.0", 17 April 2020, <<https://openconnectivity.org/upnp-specs/UPnP-arch-DeviceArchitecture-v2.0-20200417.pdf>>.
- [Swett16] Swett, I., "QUIC Deployment Experience at Google (IETF96 QUIC BoF presentation)", 20 July 2016, <<https://www.ietf.org/proceedings/96/slides/slides-96-quic-3.pdf>>.
- [TLS13] Rescorla, E., "The Transport Layer Security (TLS) Protocol Version 1.3", RFC 8446, DOI 10.17487/RFC8446, August 2018, <<https://www.rfc-editor.org/rfc/rfc8446>>.

[Trammell16]

Trammell, B. and M. Kuehlewind, "Internet Path
Transparency Measurements using RIPE Atlas (RIPE72 MAT
presentation)", 25 May 2016, <<https://ripe72.ripe.net/wp-content/uploads/presentations/86-atlas-udpdiff.pdf>>.

Authors' Addresses

Mirja Kuehlewind
Ericsson
Email: mirja.kuehlewind@ericsson.com

Brian Trammell
Google
Gustav-Gull-Platz 1
CH- 8004 Zurich
Switzerland
Email: ietf@trammell.ch

QUIC
Internet-Draft
Intended status: Standards Track
Expires: 29 September 2022

M. Duke
Google
N. Banks
Microsoft
C. Huitema
Private Octopus Inc.
28 March 2022

QUIC-LB: Generating Routable QUIC Connection IDs
draft-ietf-quic-load-balancers-13

Abstract

QUIC address migration allows clients to change their IP address while maintaining connection state. To reduce the ability of an observer to link two IP addresses, clients and servers use new connection IDs when they communicate via different client addresses. This poses a problem for traditional "layer-4" load balancers that route packets via the IP address and port 4-tuple. This specification provides a standardized means of securely encoding routing information in the server's connection IDs so that a properly configured load balancer can route packets with migrated addresses correctly. As it proposes a structured connection ID format, it also provides a means of connection IDs self-encoding their length to aid some hardware offloads.

Note to Readers

Discussion of this document takes place on the QUIC Working Group mailing list (quic@ietf.org), which is archived at <https://mailarchive.ietf.org/arch/browse/quic/> (<https://mailarchive.ietf.org/arch/browse/quic/>).

Source for this draft and an issue tracker can be found at <https://github.com/quicwg/load-balancers> (<https://github.com/quicwg/load-balancers>).

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 29 September 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this 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	4
1.1. Terminology	5
1.2. Notation	5
2. First CID octet	6
2.1. Config Rotation	6
2.2. Configuration Failover	7
2.3. Length Self-Description	7
2.4. Format	7
3. Load Balancing Preliminaries	8
3.1. Unroutable Connection IDs	8
3.2. Fallback Algorithms	10
3.3. Server ID Allocation	10
4. Server ID Encoding in Connection IDs	11
4.1. CID format	11
4.2. Configuration Agent Actions	11
4.3. Server Actions	11
4.3.1. Special Case: Single Pass Encryption	12
4.3.2. General Case: Four-Pass Encryption	12
4.4. Load Balancer Actions	14
4.4.1. Special Case: Single Pass Encryption	15
4.4.2. General Case: Four-Pass Encryption	15
5. Per-connection state	15
6. Additional Use Cases	16
6.1. Load balancer chains	16
6.2. Moving connections between servers	17

7. Version Invariance of QUIC-LB	17
8. Security Considerations	18
8.1. Attackers not between the load balancer and server	19
8.2. Attackers between the load balancer and server	19
8.3. Multiple Configuration IDs	19
8.4. Limited configuration scope	19
8.5. Stateless Reset Oracle	20
8.6. Connection ID Entropy	21
9. IANA Considerations	21
10. References	22
10.1. Normative References	22
10.2. Informative References	22
Appendix A. QUIC-LB YANG Model	23
A.1. Tree Diagram	29
Appendix B. Load Balancer Test Vectors	29
B.1. Unencrypted CIDs	30
B.2. Encrypted CIDs	30
Appendix C. Interoperability with DTLS over UDP	30
C.1. DTLS 1.0 and 1.2	30
C.2. DTLS 1.3	31
C.3. Future Versions of DTLS	32
Appendix D. Acknowledgments	32
Appendix E. Change Log	32
E.1. since draft-ietf-quic-load-balancers-12	32
E.2. since draft-ietf-quic-load-balancers-11	32
E.3. since draft-ietf-quic-load-balancers-10	32
E.4. since draft-ietf-quic-load-balancers-09	33
E.5. since draft-ietf-quic-load-balancers-08	33
E.6. since draft-ietf-quic-load-balancers-07	33
E.7. since draft-ietf-quic-load-balancers-06	33
E.8. since draft-ietf-quic-load-balancers-05	33
E.9. since draft-ietf-quic-load-balancers-04	34
E.10. since draft-ietf-quic-load-balancers-03	34
E.11. since draft-ietf-quic-load-balancers-02	34
E.12. since draft-ietf-quic-load-balancers-01	34
E.13. since draft-ietf-quic-load-balancers-00	35
E.14. Since draft-duke-quic-load-balancers-06	35
E.15. Since draft-duke-quic-load-balancers-05	35
E.16. Since draft-duke-quic-load-balancers-04	35
E.17. Since draft-duke-quic-load-balancers-03	35
E.18. Since draft-duke-quic-load-balancers-02	35
E.19. Since draft-duke-quic-load-balancers-01	36
E.20. Since draft-duke-quic-load-balancers-00	36
Authors' Addresses	36

1. Introduction

QUIC packets [RFC9000] usually contain a connection ID to allow endpoints to associate packets with different address/port 4-tuples to the same connection context. This feature makes connections robust in the event of NAT rebinding. QUIC endpoints usually designate the connection ID which peers use to address packets. Server-generated connection IDs create a potential need for out-of-band communication to support QUIC.

QUIC allows servers (or load balancers) to designate an initial connection ID to encode useful routing information for load balancers. It also encourages servers, in packets protected by cryptography, to provide additional connection IDs to the client. This allows clients that know they are going to change IP address or port to use a separate connection ID on the new path, thus reducing linkability as clients move through the world.

There is a tension between the requirements to provide routing information and mitigate linkability. Ultimately, because new connection IDs are in protected packets, they must be generated at the server if the load balancer does not have access to the connection keys. However, it is the load balancer that has the context necessary to generate a connection ID that encodes useful routing information. In the absence of any shared state between load balancer and server, the load balancer must maintain a relatively expensive table of server-generated connection IDs, and will not route packets correctly if they use a connection ID that was originally communicated in a protected NEW_CONNECTION_ID frame.

This specification provides common algorithms for encoding the server mapping in a connection ID given some shared parameters. The mapping is generally only discoverable by observers that have the parameters, preserving unlinkability as much as possible.

As this document proposes a structured QUIC Connection ID, it also proposes a system for self-encoding connection ID length in all packets, so that crypto offload can efficiently obtain key information.

While this document describes a small set of configuration parameters to make the server mapping intelligible, the means of distributing these parameters between load balancers, servers, and other trusted intermediaries is out of its scope. There are numerous well-known infrastructures for distribution of configuration.

1.1. 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].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying significance described in RFC 2119.

In this document, "client" and "server" refer to the endpoints of a QUIC connection unless otherwise indicated. A "load balancer" is an intermediary for that connection that does not possess QUIC connection keys, but it may rewrite IP addresses or conduct other IP or UDP processing. A "configuration agent" is the entity that determines the QUIC-LB configuration parameters for the network and leverages some system to distribute that configuration.

Note that stateful load balancers that act as proxies, by terminating a QUIC connection with the client and then retrieving data from the server using QUIC or another protocol, are treated as a server with respect to this specification.

For brevity, "Connection ID" will often be abbreviated as "CID".

1.2. Notation

All wire formats will be depicted using the notation defined in Section 1.3 of [RFC9000]. There is one addition: the function `len()` refers to the length of a field which can serve as a limit on a different field, so that the lengths of two fields can be concisely defined as limited to a sum, for example:

`x(A..B) y(C..B-len(x))`

indicates that `x` can be of any length between `A` and `B`, and `y` can be of any length between `C` and `B` provided that `(len(x) + len(y))` does not exceed `B`.

The example below illustrates the basic framework:

```
Example Structure {  
  One-bit Field (1),  
  7-bit Field with Fixed Value (7) = 61,  
  Field with Variable-Length Integer (i),  
  Arbitrary-Length Field (..),  
  Variable-Length Field (8..24),  
  Variable-Length Field with Dynamic Limit (8..24-len(Variable-Length Field)),  
  Field With Minimum Length (16..),  
  Field With Maximum Length (..128),  
  [Optional Field (64)],  
  Repeated Field (8) ...,  
}
```

Figure 1: Example Format

2. First CID octet

The first octet of a Connection ID is reserved for two special purposes, one mandatory (config rotation) and one optional (length self-description).

Subsequent sections of this document refer to the contents of this octet as the "first octet."

2.1. Config Rotation

The first two bits of any connection ID MUST encode an identifier for the configuration that the connection ID uses. This enables incremental deployment of new QUIC-LB settings (e.g., keys).

When new configuration is distributed to servers, there will be a transition period when connection IDs reflecting old and new configuration coexist in the network. The rotation bits allow load balancers to apply the correct routing algorithm and parameters to incoming packets.

Configuration Agents SHOULD deliver new configurations to load balancers before doing so to servers, so that load balancers are ready to process CIDs using the new parameters when they arrive.

A Configuration Agent SHOULD NOT use a codepoint to represent a new configuration until it takes precautions to make sure that all connections using CIDs with an old configuration at that codepoint have closed or transitioned.

Servers MUST NOT generate new connection IDs using an old configuration after receiving a new one from the configuration agent. Servers MUST send NEW_CONNECTION_ID frames that provide CIDs using the new configuration, and retire CIDs using the old configuration using the "Retire Prior To" field of that frame.

It is also possible to use these bits for more long-lived distinction of different configurations, but this has privacy implications (see Section 8.3).

2.2. Configuration Failover

If a server has not received a valid QUIC-LB configuration, and believes that low-state, Connection-ID aware load balancers are in the path, it SHOULD generate connection IDs with the config rotation bits set to '11' and SHOULD use the "disable_active_migration" transport parameter in all new QUIC connections. It SHOULD NOT send NEW_CONNECTION_ID frames with new values.

A load balancer that sees a connection ID with config rotation bits set to '11' MUST revert to 5-tuple routing. These connection IDs may be of any length; however, see Section 8.6 for limits on this length.

2.3. Length Self-Description

Local hardware cryptographic offload devices may accelerate QUIC servers by receiving keys from the QUIC implementation indexed to the connection ID. However, on physical devices operating multiple QUIC servers, it is impractical to efficiently lookup these keys if the connection ID does not self-encode its own length.

Note that this is a function of particular server devices and is irrelevant to load balancers. As such, load balancers MAY omit this from their configuration. However, the remaining 6 bits in the first octet of the Connection ID are reserved to express the length of the following connection ID, not including the first octet.

A server not using this functionality SHOULD make the six bits appear to be random.

2.4. Format

```
First Octet {  
    Config Rotation (2),  
    CID Len or Random Bits (6),  
}
```

Figure 2: First Octet Format

The first octet has the following fields:

Config Rotation: Indicates the configuration used to interpret the CID.

CID Len or Random Bits: Length Self-Description (if applicable), or random bits otherwise. Encodes the length of the Connection ID following the First Octet.

3. Load Balancing Preliminaries

In QUIC-LB, load balancers do not generate individual connection IDs for servers. Instead, they communicate the parameters of an algorithm to generate routable connection IDs.

The algorithms differ in the complexity of configuration at both load balancer and server. Increasing complexity improves obfuscation of the server mapping.

This section describes three participants: the configuration agent, the load balancer, and the server. For any given QUIC-LB configuration that enables connection-ID-aware load balancing, there must be a choice of (1) routing algorithm, (2) server ID allocation strategy, and (3) algorithm parameters.

Fundamentally, servers generate connection IDs that encode their server ID. Load balancers decode the server ID from the CID in incoming packets to route to the correct server.

There are situations where a server pool might be operating two or more routing algorithms or parameter sets simultaneously. The load balancer uses the first two bits of the connection ID to multiplex incoming DCIDs over these schemes (see Section 2.1).

3.1. Unroutable Connection IDs

QUIC-LB servers will generate Connection IDs that are decodable to extract a server ID in accordance with a specified algorithm and parameters. However, QUIC often uses client-generated Connection IDs prior to receiving a packet from the server.

These client-generated CIDs might not conform to the expectations of the routing algorithm and therefore not be routable by the load balancer. Those that are not routable are "unroutable DCIDs" and receive similar treatment regardless of why they're unroutable:

- * The config rotation bits (Section 2.1) may not correspond to an active configuration. Note: a packet with a DCID that indicates 5-tuple routing (see Section 2.2) is always routable.
- * The DCID might not be long enough for the decoder to process.
- * The extracted server mapping might not correspond to an active server.

All other DCIDs are routable.

Load balancers **MUST** forward packets with routable DCIDs to a server in accordance with the chosen routing algorithm. Exception: if the load balancer can parse the QUIC packet and makes a routing decision depending on the contents (e.g., the SNI in a TLS client hello), it **MAY** route in accordance with this instead. However, load balancers **MUST** always route long header packets it cannot parse in accordance with the DCID (see Section 7).

Load balancers **SHOULD** drop short header packets with unroutable DCIDs.

When forwarding a packet with a long header and unroutable DCID, load balancers **MUST** use a fallback algorithm as specified in Section 3.2.

Load balancers **MAY** drop packets with long headers and unroutable DCIDs if and only if it knows that the encoded QUIC version does not allow an unroutable DCID in a packet with that signature. For example, a load balancer can safely drop a QUIC version 1 Handshake packet with an unroutable DCID, as a version 1 Handshake packet sent to a QUIC-LB routable server will always have a server-generated routable CID. The prohibition against dropping packets with long headers remains for unknown QUIC versions.

Furthermore, while the load balancer function **MUST NOT** drop packets, the device might implement other security policies, outside the scope of this specification, that might force a drop.

Servers that receive packets with unroutable CIDs **MUST** use the available mechanisms to induce the client to use a routable CID in future packets. In QUIC version 1, this requires using a routable CID in the Source CID field of server-generated long headers.

3.2. Fallback Algorithms

There are conditions described below where a load balancer routes a packet using a "fallback algorithm." It can choose any algorithm, without coordination with the servers, but the algorithm SHOULD be deterministic over short time scales so that related packets go to the same server. The design of this algorithm SHOULD consider the version-invariant properties of QUIC described in [RFC8999] to maximize its robustness to future versions of QUIC.

A fallback algorithm MUST NOT make the routing behavior dependent on any bits in the first octet of the QUIC packet header, except the first bit, which indicates a long header. All other bits are QUIC version-dependent and intermediaries SHOULD NOT base their design on version-specific templates.

For example, one fallback algorithm might convert a unroutable DCID to an integer and divided by the number of servers, with the modulus used to forward the packet. The number of servers is usually consistent on the time scale of a QUIC connection handshake. Another might simply hash the address/port 4-tuple. See also Section 7.

3.3. Server ID Allocation

Load Balancer configurations include a mapping of server IDs to forwarding addresses. The corresponding server configurations contain one or more unique server IDs.

The configuration agent chooses a server ID length for each configuration that MUST be at least one octet.

A QUIC-LB configuration MAY significantly over-provision the server ID space (i.e., provide far more codepoints than there are servers) to increase the probability that a randomly generated Destination Connection ID is unroutable.

The configuration agent SHOULD provide a means for servers to express the number of server IDs it can usefully employ, because a single routing address actually corresponds to multiple server entities (see Section 6.1).

Conceptually, each configuration has its own set of server ID allocations, though two static configurations with identical server ID lengths MAY use a common allocation between them.

A server encodes one of its assigned server IDs in any CID it generates using the relevant configuration.

4. Server ID Encoding in Connection IDs

4.1. CID format

All connection IDs use the following format:

```
QUIC-LB Connection ID {  
    First Octet (8),  
    Server ID (8..152-len(Nonce)),  
    Nonce (32..152-len(Server ID)),  
}
```

Figure 3: CID Format

4.2. Configuration Agent Actions

The configuration agent assigns a server ID to every server in its pool in accordance with Section 3.3, and determines a server ID length (in octets) sufficiently large to encode all server IDs, including potential future servers.

Each configuration specifies the length of the Server ID and Nonce fields, with limits defined for each algorithm.

Optionally, it also defines a 16-octet key. Note that failure to define a key means that observers can determine the assigned server of any connection, significantly increasing the linkability of QUIC address migration.

The nonce length **MUST** be at least 4 octets. The server ID length **MUST** be at least 1 octet.

As QUIC version 1 limits connection IDs to 20 octets, the server ID and nonce lengths **MUST** sum to 19 octets or less.

4.3. Server Actions

The server writes the first octet and its server ID into their respective fields.

If there is no key in the configuration, the server **MUST** fill the Nonce field with bytes that appear to be random. If there is a key, the server fills the nonce field with a nonce of its choosing. See Section 8.6 for details.

The server **MAY** append additional bytes to the connection ID, up to the limit specified in that version of QUIC, for its own use. These bytes **MUST NOT** provide observers with any information that could link

two connection IDs to the same connection, client, or server. In particular, all servers using a configuration MUST consistently add the same length to each connection ID, to preserve the linkability objectives of QUIC-LB. Any additional bytes SHOULD appear random unless individual servers are not distinguishable (e.g. any server using that configuration appends identical bytes to every connection ID).

If there is no key in the configuration, the Connection ID is complete. Otherwise, there are further steps, as described in the two following subsections.

Encryption below uses the AES-128-ECB cipher. Future standards could add new algorithms that use other ciphers to provide cryptographic agility in accordance with [RFC7696]. QUIC-LB implementations SHOULD be extensible to support new algorithms.

4.3.1. Special Case: Single Pass Encryption

When the nonce length and server ID length sum to exactly 16 octets, the server MUST use a single-pass encryption algorithm. All connection ID octets except the first form an AES-ECB block. This block is encrypted once, and the result forms the second through seventeenth most significant bytes of the connection ID.

4.3.2. General Case: Four-Pass Encryption

Any other field length requires four passes for encryption and at least three for decryption. To understand this algorithm, it is useful to define four functions that minimize the amount of bit-shifting necessary in the event that there are an odd number of octets.

The `expand_left()` function outputs 16 octets, with its first argument in the most significant bits, its second argument in the least significant byte, and zeros in all other positions. Thus,

```
expand_left(0xaaba3c, 0x13) = 0xaaba3c00000000000000000000000013
```

`expand_right()` is similar, except that the second argument is in the most significant byte, and the first argument is in the least significant bits. Therefore,

```
expand_right(0xaaba3c, 0x13) = 0x13000000000000000000000000aaba3c
```

Similarly, `truncate_left()` and `truncate_right()` take the most significant and least significant bits, respectively, from a ciphertext. For example, to take 28 bits of a ciphertext:

```
truncate_left(0x2094842ca49256198c2deaa0ba53caa0, 28) = 0x2094842
truncate_right(0x2094842ca49256198c2deaa0ba53caa0, 28) = 0xa53caa0
```

The example at the end of this section helps to clarify the steps described below.

1. The server concatenates the server ID and nonce to create plaintext_CID.
2. The server splits plaintext_CID into components left_0 and right_0 of equal length, splitting an odd octet in half if necessary. For example, 0x7040b81b55ccf3 would split into a left_0 of 0x7040b81 and right_0 of 0xb55ccf3.
3. Encrypt the result of expand_left(left_0) to obtain a ciphertext.
4. XOR the least significant bits of the ciphertext with right_0 to form right_1.

```
Thus steps 3 and 4 can be expressed as right_1 = right_0 ^
truncate_right( AES_ECB(key, expand_left(left_0, 0x01)),
len(right_0))
```

5. Repeat steps 3 and 4, but use them to compute left_1 by expanding and encrypting right_1 with the most significant octet as 0x02 and XOR the results with left_0.

```
left_1 = left_0 ^ truncate_left( AES_ECB(key,
expand_right(right_1, 0x02)), len(left_0))
```

6. Repeat steps 3 and 4, but use them to compute right_2 by expanding and encrypting left_1 with the least significant octet as 0x03 and XOR the results with right_1.

```
right_2 = right_1 ^ truncate_right( AES_ECB(key,
expand_left(left_1, 0x03)), len(right_1))
```

7. Repeat steps 3 and 4, but use them to compute left_2 by expanding and encrypting right_2 with the most significant octet as 0x04 and XOR the results with left_1.

```
left_2 = left_1 ^ truncate_left( AES_ECB(key,
expand_right(right_2, 0x04)), len(left_1))
```

8. The server concatenates left_2 with right_2 to form the ciphertext CID, which it appends to the first octet.

The following example executes the steps for the provided inputs. Note that the plaintext is of odd octet length, so the middle octet will be split evenly left_0 and right_0.

```
server_id = 0x31441a
nonce = 0x9c69c275
key = 0xdf726a9893ec05c0632d3956680baf0

// step 1
plaintext_CID = 0x31441a9c69c275

// step 2
left_0 = 0x31441a9
right_0 = 0xc69c275

// step 3
aes_input = 0x31441a90000000000000000000000001
ciphertext = 0x4d140de42d0b85bdf554ba35c1d5c653

// step 4
right_1 = 0xc69c275 ^ 0x1d5c653 = 0xdbbc0426

// step 5
aes_input = 0x020000000000000000000000dbbc0426
aes_output = 0x7e99160f3cf5b89c70584ccd2c2cd24b
left_1 = 0x31441a9 ^ 0x7e99160 = 0x4fdd0c9

// step 6
AES input = 0x4fdd0c90000000000000000000000003
AES output = 0x26c1d5a3a5e31ff8e3ca505da6061ac6
right_2 = 0xdbbc0426 ^ 0x6061ac6 = 0xbba1ee0

// step 7
AES input = 0x04000000000000000000000000bba1ee0
AES output = 0xad1b8b25b436a94007d80cf3704377b
left_2 = 0x4fdd0c9 ^ 0xad1b8b = 0xe23cb42

// step 8
cid = first_octet || left_2 || right_2 = 0x07e23cb42bba1ee0
```

4.4. Load Balancer Actions

On each incoming packet, the load balancer extracts consecutive octets, beginning with the second octet. If there is no key, the first octets correspond to the server ID.

If there is a key, the load balancer takes one of two actions:

4.4.1. Special Case: Single Pass Encryption

If server ID length and nonce length sum to exactly 16 octets, they form a ciphertext block. The load balancer decrypts the block using the AES-ECB key and extracts the server ID from the most significant bytes of the resulting plaintext.

4.4.2. General Case: Four-Pass Encryption

First, split the ciphertext CID (excluding the first octet) into its equal-length components `left_2` and `right_2`. Then follow the process below:

```
left_1 = left_2 ^ truncate_left(AES_ECB(key, expand_right(right_2), 0x04))
right_1 = right_2 ^ truncate_right(AES_ECB(key, expand_left(left_1, 0x03))
left_0 = left_1 ^ truncate_left(AES_ECB(key, expand_right(right_1), 0x02))
```

As the load balancer has no need for the nonce, it can conclude after 3 passes as long as the server ID is entirely contained in `left_0` (i.e., the nonce is at least as large as the server ID). If the server ID is longer, a fourth pass is necessary:

```
right_0 = right_1 ^ truncate_right(AES_ECB(key, expand_left(left_0,
0x01)))
```

and the load balancer has to concatenate `left_0` and `right_0` to obtain the complete server ID.

5. Per-connection state

QUIC-LB requires no per-connection state at the load balancer. The load balancer can extract the server ID from the connection ID of each incoming packet and route that packet accordingly.

However, once the routing decision has been made, the load balancer MAY associate the 4-tuple with the decision. This has two advantages:

- * The load balancer only extracts the server ID once per incoming 4-tuple. When the CID is encrypted, this substantially reduces computational load.
- * Incoming Stateless Reset packets and ICMP messages are easily routed to the correct origin server.

In addition to the increased state requirements, however, load balancers cannot detect the CONNECTION_CLOSE frame to indicate the end of the connection, so they rely on a timeout to delete connection state. There are numerous considerations around setting such a timeout.

In the event a connection ends, freeing an IP and port, and a different connection migrates to that IP and port before the timeout, the load balancer will misroute the different connection's packets to the original server. A short timeout limits the likelihood of such a misrouting.

Furthermore, if a short timeout causes premature deletion of state, the routing is easily recoverable by decoding an incoming Connection ID. However, a short timeout also reduces the chance that an incoming Stateless Reset is correctly routed.

Servers MAY implement the technique described in Section 14.4.1 of [RFC9000] in case the load balancer is stateless, to increase the likelihood a Source Connection ID is included in ICMP responses to Path Maximum Transmission Unit (PMTU) probes. Load balancers MAY parse the echoed packet to extract the Source Connection ID, if it contains a QUIC long header, and extract the Server ID as if it were in a Destination CID.

6. Additional Use Cases

This section discusses considerations for some deployment scenarios not implied by the specification above.

6.1. Load balancer chains

Some network architectures may have multiple tiers of low-state load balancers, where a first tier of devices makes a routing decision to the next tier, and so on, until packets reach the server. Although QUIC-LB is not explicitly designed for this use case, it is possible to support it.

If each load balancer is assigned a range of server IDs that is a subset of the range of IDs assigned to devices that are closer to the client, then the first devices to process an incoming packet can extract the server ID and then map it to the correct forwarding address. Note that this solution is extensible to arbitrarily large numbers of load-balancing tiers, as the maximum server ID space is quite large.

If the number of necessary server IDs per next hop is uniform, a simple implementation would use successively longer server IDs at each tier of load balancing, and the server configuration would match the last tier. The forward load balancers would simply treat the least significant bits of the server ID as part of the nonce.

6.2. Moving connections between servers

Some deployments may transparently move a connection from one server to another. The means of transferring connection state between servers is out of scope of this document.

To support a handover, a server involved in the transition could issue CIDs that map to the new server via a `NEW_CONNECTION_ID` frame, and retire CIDs associated with the new server using the "Retire Prior To" field in that frame.

Alternately, if the old server is going offline, the load balancer could simply map its server ID to the new server's address.

7. Version Invariance of QUIC-LB

The server ID encodings, and requirements for their handling, are designed to be QUIC version independent (see [RFC8999]). A QUIC-LB load balancer will generally not require changes as servers deploy new versions of QUIC. However, there are several unlikely future design decisions that could impact the operation of QUIC-LB.

The maximum Connection ID length could be below the minimum necessary for one or more encoding algorithms.

Section 3.1 provides guidance about how load balancers should handle unroutable DCIDs. This guidance, and the implementation of an algorithm to handle these DCIDs, rests on some assumptions:

- * Incoming short headers do not contain DCIDs that are client-generated.
- * The use of client-generated incoming DCIDs does not persist beyond a few round trips in the connection.
- * While the client is using DCIDs it generated, some exposed fields (IP address, UDP port, client-generated destination Connection ID) remain constant for all packets sent on the same connection.

While this document does not update the commitments in [RFC8999], the additional assumptions are minimal and narrowly scoped, and provide a likely set of constants that load balancers can use with minimal risk of version- dependence.

If these assumptions are invalid, this specification is likely to lead to loss of packets that contain unroutable DCIDs, and in extreme cases connection failure.

Some load balancers might inspect elements of the Server Name Indication (SNI) extension in the TLS Client Hello to make a routing decision. Note that the format and cryptographic protection of this information may change in future versions or extensions of TLS or QUIC, and therefore this functionality is inherently not version-invariant. See also Section 3.1 for other considerations about this case. Note that an SNI-aware load balancer, faced with an unknown QUIC version, might misdirect initial packets to the wrong tenant. While inefficient, this preserves the ability for tenants to deploy new versions provided they have an out-of-band means of providing a connection ID for the client to use.

8. Security Considerations

QUIC-LB is intended to prevent linkability. Attacks would therefore attempt to subvert this purpose.

Note that without a key for the encoding, QUIC-LB makes no attempt to obscure the server mapping, and therefore does not address these concerns. Without a key, QUIC-LB merely allows consistent CID encoding for compatibility across a network infrastructure, which makes QUIC robust to NAT rebinding. Servers that are encoding their server ID without a key algorithm SHOULD only use it to generate new CIDs for the Server Initial Packet and SHOULD NOT send CIDs in QUIC NEW_CONNECTION_ID frames, except that it sends one new Connection ID in the event of config rotation Section 2.1. Doing so might falsely suggest to the client that said CIDs were generated in a secure fashion.

A linkability attack would find some means of determining that two connection IDs route to the same server. As described above, there is no scheme that strictly prevents linkability for all traffic patterns, and therefore efforts to frustrate any analysis of server ID encoding have diminishing returns.

8.1. Attackers not between the load balancer and server

Any attacker might open a connection to the server infrastructure and aggressively simulate migration to obtain a large sample of IDs that map to the same server. It could then apply analytical techniques to try to obtain the server encoding.

An encrypted encoding provides robust protection against this. An unencrypted one provides none.

Were this analysis to obtain the server encoding, then on-path observers might apply this analysis to correlating different client IP addresses.

8.2. Attackers between the load balancer and server

Attackers in this privileged position are intrinsically able to map two connection IDs to the same server. The QUIC-LB algorithms do prevent the linkage of two connection IDs to the same individual connection if servers make reasonable selections when generating new IDs for that connection.

8.3. Multiple Configuration IDs

During the period in which there are multiple deployed configuration IDs (see Section 2.1), there is a slight increase in linkability. The server space is effectively divided into segments with CIDs that have different config rotation bits. Entities that manage servers SHOULD strive to minimize these periods by quickly deploying new configurations across the server pool.

8.4. Limited configuration scope

A simple deployment of QUIC-LB in a cloud provider might use the same global QUIC-LB configuration across all its load balancers that route to customer servers. An attacker could then simply become a customer, obtain the configuration, and then extract server IDs of other customers' connections at will.

To avoid this, the configuration agent SHOULD issue QUIC-LB configurations to mutually distrustful servers that have different keys for encryption algorithms. In many cases, the load balancers can distinguish these configurations by external IP address.

However, assigning multiple entities to an IP address is complimentary with concealing DNS requests (e.g., DoH [RFC8484]) and the TLS Server Name Indicator (SNI) ([I-D.ietf-tls-esni]) to obscure the ultimate destination of traffic. While the load balancer's

fallback algorithm (Section 3.2) can use the SNI to make a routing decision on the first packet, there are three ways to route subsequent packets:

- * all co-tenants can use the same QUIC-LB configuration, leaking the server mapping to each other as described above;
- * co-tenants can be issued one of up to three configurations distinguished by the config rotation bits (Section 2.1), exposing information about the target domain to the entire network; or
- * tenants can use 4-tuple routing in their CIDs (in which case they SHOULD disable migration in their connections), which neutralizes the value of QUIC-LB but preserves privacy.

When configuring QUIC-LB, administrators must evaluate the privacy tradeoff considering the relative value of each of these properties, given the trust model between tenants, the presence of methods to obscure the domain name, and value of address migration in the tenant use cases.

As the plaintext algorithm makes no attempt to conceal the server mapping, these deployments SHOULD simply use a common configuration.

8.5. Stateless Reset Oracle

Section 21.9 of [RFC9000] discusses the Stateless Reset Oracle attack. For a server deployment to be vulnerable, an attacking client must be able to cause two packets with the same Destination CID to arrive at two different servers that share the same cryptographic context for Stateless Reset tokens. As QUIC-LB requires deterministic routing of DCIDs over the life of a connection, it is a sufficient means of avoiding an Oracle without additional measures.

Note also that when a server starts using a new QUIC-LB config rotation codepoint, new CIDs might not be unique with respect to previous configurations that occupied that codepoint, and therefore different clients may have observed the same CID and stateless reset token. A straightforward method of managing stateless reset keys is to maintain a separate key for each config rotation codepoint, and replace each key when the configuration for that codepoint changes. Thus, a server transitions from one config to another, it will be able to generate correct tokens for connections using either type of CID.

8.6. Connection ID Entropy

If a server ever reuses a nonce in generating a CID for a given configuration, it risks exposing sensitive information. Given the same server ID, the CID will be identical (aside from a possible difference in the first octet). This can risk exposure of the QUIC-LB key. If two clients receive the same connection ID, they also have each other's stateless reset token unless that key has changed in the interim.

The encrypt mode needs to generate different cipher text for each generated Connection ID instance to protect the Server ID. To do so, at least four octets of the CID are reserved for a nonce that, if used only once, will result in unique cipher text for each Connection ID.

If servers simply increment the nonce by one with each generated connection ID, then it is safe to use the existing keys until any server's nonce counter exhausts the allocated space and rolls over. To maximize entropy, servers SHOULD start with a random nonce value, in which case the configuration is usable until the nonce value wraps around to zero and then reaches the initial value again.

Whether or not it implements the counter method, the server MUST NOT reuse a nonce until it switches to a configuration with new keys.

If the nonce is sent in plaintext, servers MUST generate nonces so that they appear to be random. Observable correlations between plaintext nonces would provide trivial linkability between individual connections, rather than just to a common server.

For any algorithm, configuration agents SHOULD implement an out-of-band method to discover when servers are in danger of exhausting their nonce space, and SHOULD respond by issuing a new configuration. A server that has exhausted its nonces MUST either switch to a different configuration, or if none exists, use the 4-tuple routing config rotation codepoint.

When sizing a nonce that is to be randomly generated, the configuration agent SHOULD consider that a server generating a N-bit nonce will create a duplicate about every $2^{(N/2)}$ attempts, and therefore compare the expected rate at which servers will generate CIDs with the lifetime of a configuration.

9. IANA Considerations

There are no IANA requirements.

10. References

10.1. Normative References

- [RFC8999] Thomson, M., "Version-Independent Properties of QUIC", RFC 8999, DOI 10.17487/RFC8999, May 2021, <<https://www.rfc-editor.org/info/rfc8999>>.
- [RFC9000] Iyengar, J., Ed. and M. Thomson, Ed., "QUIC: A UDP-Based Multiplexed and Secure Transport", RFC 9000, DOI 10.17487/RFC9000, May 2021, <<https://www.rfc-editor.org/info/rfc9000>>.

10.2. Informative References

- [I-D.draft-ietf-tls-dtls13]
Rescorla, E., Tschofenig, H., and N. Modadugu, "The Datagram Transport Layer Security (DTLS) Protocol Version 1.3", Work in Progress, Internet-Draft, draft-ietf-tls-dtls13-43, 30 April 2021, <<https://www.ietf.org/archive/id/draft-ietf-tls-dtls13-43.txt>>.
- [I-D.ietf-tls-dtls-connection-id]
Rescorla, E., Tschofenig, H., Fossati, T., and A. Kraus, "Connection Identifier for DTLS 1.2", Work in Progress, Internet-Draft, draft-ietf-tls-dtls-connection-id-13, 22 June 2021, <<https://www.ietf.org/archive/id/draft-ietf-tls-dtls-connection-id-13.txt>>.
- [I-D.ietf-tls-esni]
Rescorla, E., Oku, K., Sullivan, N., and C. A. Wood, "TLS Encrypted Client Hello", Work in Progress, Internet-Draft, draft-ietf-tls-esni-14, 13 February 2022, <<https://www.ietf.org/archive/id/draft-ietf-tls-esni-14.txt>>.
- [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>>.
- [RFC4347] Rescorla, E. and N. Modadugu, "Datagram Transport Layer Security", RFC 4347, DOI 10.17487/RFC4347, April 2006, <<https://www.rfc-editor.org/info/rfc4347>>.

- [RFC6020] Bjorklund, M., Ed., "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, DOI 10.17487/RFC6020, October 2010, <<https://www.rfc-editor.org/info/rfc6020>>.
- [RFC6347] Rescorla, E. and N. Modadugu, "Datagram Transport Layer Security Version 1.2", RFC 6347, DOI 10.17487/RFC6347, January 2012, <<https://www.rfc-editor.org/info/rfc6347>>.
- [RFC7696] Housley, R., "Guidelines for Cryptographic Algorithm Agility and Selecting Mandatory-to-Implement Algorithms", BCP 201, RFC 7696, DOI 10.17487/RFC7696, November 2015, <<https://www.rfc-editor.org/info/rfc7696>>.
- [RFC7983] Petit-Huguenin, M. and G. Salgueiro, "Multiplexing Scheme Updates for Secure Real-time Transport Protocol (SRTP) Extension for Datagram Transport Layer Security (DTLS)", RFC 7983, DOI 10.17487/RFC7983, September 2016, <<https://www.rfc-editor.org/info/rfc7983>>.
- [RFC8340] Bjorklund, M. and L. Berger, Ed., "YANG Tree Diagrams", BCP 215, RFC 8340, DOI 10.17487/RFC8340, March 2018, <<https://www.rfc-editor.org/info/rfc8340>>.
- [RFC8484] Hoffman, P. and P. McManus, "DNS Queries over HTTPS (DoH)", RFC 8484, DOI 10.17487/RFC8484, October 2018, <<https://www.rfc-editor.org/info/rfc8484>>.

Appendix A. QUIC-LB YANG Model

These YANG models conform to [RFC6020] and express a complete QUIC-LB configuration. There is one model for the server and one for the middlebox (i.e the load balancer and/or Retry Service).

```
module ietf-quic-lb-server {
  yang-version "1.1";
  namespace "urn:ietf:params:xml:ns:yang:ietf-quic-lb";
  prefix "quic-lb";

  import ietf-yang-types {
    prefix yang;
    reference
      "RFC 6991: Common YANG Data Types.";
  }

  import ietf-inet-types {
    prefix inet;
    reference
```

```
"RFC 6991: Common YANG Data Types.";
}

organization
  "IETF QUIC Working Group";

contact
  "WG Web:  <http://datatracker.ietf.org/wg/quic>
  WG List:  <quic@ietf.org>

  Authors: Martin Duke (martin.h.duke at gmail dot com)
           Nick Banks (nibanks at microsoft dot com)
           Christian Huitema (huitema at huitema.net)";

description
  "This module enables the explicit cooperation of QUIC servers with
  trusted intermediaries without breaking important protocol
  features.

  Copyright (c) 2022 IETF Trust and the persons identified as
  authors of the code.  All rights reserved.

  Redistribution and use in source and binary forms, with or
  without modification, is permitted pursuant to, and subject to
  the license terms contained in, the Simplified BSD License set
  forth in Section 4.c of the IETF Trust's Legal Provisions
  Relating to IETF Documents
  (https://trustee.ietf.org/license-info).

  This version of this YANG module is part of RFC XXXX
  (https://www.rfc-editor.org/info/rfcXXXX); see the RFC itself
  for full legal notices.

  The key words 'MUST', 'MUST NOT', 'REQUIRED', 'SHALL', 'SHALL
  NOT', 'SHOULD', 'SHOULD NOT', 'RECOMMENDED', 'NOT RECOMMENDED',
  'MAY', and 'OPTIONAL' in this document are to be interpreted as
  described in BCP 14 (RFC 2119) (RFC 8174) when, and only when,
  they appear in all capitals, as shown here.";

revision "2022-02-11" {
  description
    "Updated to design in version 13 of the draft";
  reference
    "RFC XXXX, QUIC-LB: Generating Routable QUIC Connection IDs";
}

container quic-lb {
  presence "The container for QUIC-LB configuration.";
```

```
description
  "QUIC-LB container.";

typedef quic-lb-key {
  type yang:hex-string {
    length 47;
  }
  description
    "This is a 16-byte key, represented with 47 bytes";
}

leaf config-id {
  type uint8 {
    range "0..2";
  }
  mandatory true;
  description
    "Identifier for this CID configuration.";
}

leaf first-octet-encodes-cid-length {
  type boolean;
  default false;
  description
    "If true, the six least significant bits of the first CID
    octet encode the CID length minus one.";
}

leaf server-id-length {
  type uint8 {
    range "1..15";
  }
  must '. <= (19 - ../nonce-length)' {
    error-message
      "Server ID and nonce lengths must sum to no more than 19.";
  }
  mandatory true;
  description
    "Length (in octets) of a server ID. Further range-limited
    by nonce-length.";
}

leaf nonce-length {
  type uint8 {
    range "4..18";
  }
  mandatory true;
  description
```

```
        "Length, in octets, of the nonce. Short nonces mean there will
        be frequent configuration updates.";
    }

    leaf cid-key {
        type quic-lb-key;
        description
            "Key for encrypting the connection ID.";
    }

    leaf server-id {
        type yang:hex-string;
        must "string-length(.) = 3 * ../../server-id-length - 1";
        mandatory true;
        description
            "An allocated server ID";
    }
}

module ietf-quic-lb-middlebox {
    yang-version "1.1";
    namespace "urn:ietf:params:xml:ns:yang:ietf-quic-lb";
    prefix "quic-lb";

    import ietf-yang-types {
        prefix yang;
        reference
            "RFC 6991: Common YANG Data Types.";
    }

    import ietf-inet-types {
        prefix inet;
        reference
            "RFC 6991: Common YANG Data Types.";
    }

    organization
        "IETF QUIC Working Group";

    contact
        "WG Web:  <http://datatracker.ietf.org/wg/quic>
        WG List:  <quic@ietf.org>

        Authors: Martin Duke (martin.h.duke at gmail dot com)
                 Nick Banks (nibanks at microsoft dot com)
                 Christian Huitema (huitema at huitema.net)";
```


description

"This module enables the explicit cooperation of QUIC servers with trusted intermediaries without breaking important protocol features.

Copyright (c) 2021 IETF Trust and the persons identified as authors of the code. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, is permitted pursuant to, and subject to the license terms contained in, the Simplified BSD License set forth in Section 4.c of the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>).

This version of this YANG module is part of RFC XXXX (<https://www.rfc-editor.org/info/rfcXXXX>); see the RFC itself for full legal notices.

The key words 'MUST', 'MUST NOT', 'REQUIRED', 'SHALL', 'SHALL NOT', 'SHOULD', 'SHOULD NOT', 'RECOMMENDED', 'NOT RECOMMENDED', 'MAY', and 'OPTIONAL' in this document are to be interpreted as described in BCP 14 (RFC 2119) (RFC 8174) when, and only when, they appear in all capitals, as shown here.";

```
revision "2021-02-11" {  
  description  
    "Updated to design in version 13 of the draft";  
  reference  
    "RFC XXXX, QUIC-LB: Generating Routable QUIC Connection IDs";  
}
```

```
container quic-lb {  
  presence "The container for QUIC-LB configuration.";  
  
  description  
    "QUIC-LB container.";  
  
  typedef quic-lb-key {  
    type yang:hex-string {  
      length 47;  
    }  
    description  
      "This is a 16-byte key, represented with 47 bytes";  
  }  
  
  list cid-configs {  
    key "config-rotation-bits";
```

```
description
  "List up to three load balancer configurations";

leaf config-rotation-bits {
  type uint8 {
    range "0..2";
  }
  mandatory true;
  description
    "Identifier for this CID configuration.";
}

leaf server-id-length {
  type uint8 {
    range "1..15";
  }
  must '. <= (19 - ../nonce-length)' {
    error-message
      "Server ID and nonce lengths must sum to no more than 19.";
  }
  mandatory true;
  description
    "Length (in octets) of a server ID. Further range-limited
    by nonce-length.";
}

leaf cid-key {
  type quic-lb-key;
  description
    "Key for encrypting the connection ID.";
}

leaf nonce-length {
  type uint8 {
    range "4..18";
  }
  mandatory true;
  description
    "Length, in octets, of the nonce. Short nonces mean there
    will be frequent configuration updates.";
}

list server-id-mappings {
  key "server-id";
  description "Statically allocated Server IDs";

  leaf server-id {
    type yang:hex-string;
```

```

        must "string-length(.) = 3 * ../../server-id-length - 1";
        mandatory true;
        description
            "An allocated server ID";
    }

    leaf server-address {
        type inet:ip-address;
        mandatory true;
        description
            "Destination address corresponding to the server ID";
    }
}
}
}
}

```

A.1. Tree Diagram

This summary of the YANG models uses the notation in [RFC8340].

```

module: ietf-quic-lb-server
  +--rw quic-lb!
    +--rw config-id                               uint8
    +--rw first-octet-encodes-cid-length?         boolean
    +--rw server-id-length                       uint8
    +--rw nonce-length                           uint8
    +--rw cid-key?                               quic-lb-key
    +--rw server-id                             yang:hex-string

module: ietf-quic-lb-middlebox
  +--rw quic-lb!
    +--rw cid-configs* [config-rotation-bits]
      +--rw config-rotation-bits                 uint8
      +--rw server-id-length                     uint8
      +--rw cid-key?                             quic-lb-key
      +--rw nonce-length                         uint8
      +--rw server-id-mappings* [server-id]
        +--rw server-id                         yang:hex-string
        +--rw server-address                     inet:ip-address

```

Appendix B. Load Balancer Test Vectors

This section uses the following abbreviations:

cid Connection ID
cr_bits Config Rotation Bits
LB Load Balancer
sid Server ID

In all cases, the server is configured to encode the CID length.

B.1. Unencrypted CIDs

```
cr_bits sid nonce cid
0 c4605e 4504cc4f 07c4605e4504cc4f
1 350d28b420 3487d970b 40a350d28b4203487d970b
```

B.2. Encrypted CIDs

The key for all of these examples is 8f95f09245765f80256934e50c66207f. The test vectors include an example that uses the 16-octet single-pass special case, as well as an instance where the server ID length exceeds the nonce length, requiring a fourth decryption pass.

```
cr_bits sid nonce cid
0 ed793a ee080dbf 07fbfe05f731b425
1 ed793a51d49b8f5fab65 ee080dbf48 4f010956fb5c1d4d86e010183e0b7dle
2 ed793a51d49b8f5f ee080dbf48c0dle5 904dd2d05a7b0de9b2b9907afb5ecf8cc3
0 ed793a51d49b8f5fab ee080dbf48c0dle55d 127a285a09f85280f4fd6abb434a7159e4d3eb
```

Appendix C. Interoperability with DTLS over UDP

Some environments may contain DTLS traffic as well as QUIC operating over UDP, which may be hard to distinguish.

In most cases, the packet parsing rules above will cause a QUIC-LB load balancer to route DTLS traffic in an appropriate way. DTLS 1.3 implementations that use the connection_id extension [I-D.ietf-tls-dtls-connection-id] might use the techniques in this document to generate connection IDs and achieve robust routability for DTLS associations if they meet a few additional requirements. This non-normative appendix describes this interaction.

C.1. DTLS 1.0 and 1.2

DTLS 1.0 [RFC4347] and 1.2 [RFC6347] use packet formats that a QUIC-LB router will interpret as short header packets with CIDs that request 4-tuple routing. As such, they will route such packets consistently as long as the 4-tuple does not change. Note that DTLS 1.0 has been deprecated by the IETF.

The first octet of every DTLS 1.0 or 1.2 datagram contains the content type. A QUIC-LB load balancer will interpret any content type less than 128 as a short header packet, meaning that the subsequent octets should contain a connection ID.

Existing TLS content types comfortably fit in the range below 128. Assignment of codepoints greater than 64 would require coordination in accordance with [RFC7983], and anyway would likely create problems demultiplexing DTLS and version 1 of QUIC. Therefore, this document believes it is extremely unlikely that TLS content types of 128 or greater will be assigned. Nevertheless, such an assignment would cause a QUIC-LB load balancer to interpret the packet as a QUIC long header with an essentially random connection ID, which is likely to be routed irregularly.

The second octet of every DTLS 1.0 or 1.2 datagram is the bitwise complement of the DTLS Major version (i.e. version 1.x = 0xfe). A QUIC-LB load balancer will interpret this as a connection ID that requires 4-tuple based load balancing, meaning that the routing will be consistent as long as the 4-tuple remains the same.

[I-D.ietf-tls-dtls-connection-id] defines an extension to add connection IDs to DTLS 1.2. Unfortunately, a QUIC-LB load balancer will not correctly parse the connection ID and will continue 4-tuple routing. A modified QUIC-LB load balancer that correctly identifies DTLS and parses a DTLS 1.2 datagram for the connection ID is outside the scope of this document.

C.2. DTLS 1.3

DTLS 1.3 [I-D.draft-ietf-tls-dtls13] changes the structure of datagram headers in relevant ways.

Handshake packets continue to have a TLS content type in the first octet and 0xfe in the second octet, so they will be 4-tuple routed, which should not present problems for likely NAT rebinding or address change events.

Non-handshake packets always have zero in their most significant bit and will therefore always be treated as QUIC short headers. If the connection ID is present, it follows in the succeeding octets. Therefore, a DTLS 1.3 association where the server utilizes Connection IDs and the encodings in this document will be routed correctly in the presence of client address and port changes.

However, if the client does not include the `connection_id` extension in its ClientHello, the server is unable to use connection IDs. In this case, non-handshake packets will appear to contain random

connection IDs and be routed randomly. Thus, unmodified QUIC-LB load balancers will not work with DTLS 1.3 if the client does not advertise support for connection IDs, or the server does not request the use of a compliant connection ID.

A QUIC-LB load balancer might be modified to identify DTLS 1.3 packets and correctly parse the fields to identify when there is no connection ID and revert to 4-tuple routing, removing the server requirement above. However, such a modification is outside the scope of this document, and classifying some packets as DTLS might be incompatible with future versions of QUIC.

C.3. Future Versions of DTLS

As DTLS does not have an IETF consensus document that defines what parts of DTLS will be invariant in future versions, it is difficult to speculate about the applicability of this section to future versions of DTLS.

Appendix D. Acknowledgments

Manasi Deval, Erik Fuller, Toma Gavrichenkov, Jana Iyengar, Subodh Iyengar, Ladislav Lhotka, Jan Lindblad, Ling Tao Nju, Ilari Liusvaara, Kazuho Oku, Udip Pant, Ian Swett, Martin Thomson, Dmitri Tikhonov, Victor Vasiliev, and William Zeng Ke all provided useful input to this document.

Appendix E. Change Log

RFC Editor's Note: Please remove this section prior to publication of a final version of this document.

E.1. since draft-ietf-quic-load-balancers-12

- * Separated Retry Service design into a separate draft.

E.2. since draft-ietf-quic-load-balancers-11

- * Fixed mistakes in test vectors

E.3. since draft-ietf-quic-load-balancers-10

- * Refactored algorithm descriptions; made the 4-pass algorithm easier to implement
- * Revised test vectors
- * Split YANG model into a server and middlebox version

E.4. since draft-ietf-quic-load-balancers-09

- * Renamed "Stream Cipher" and "Block Cipher" to "Encrypted Short" and "Encrypted Long"
- * Added section on per-connection state
- * Changed "Encrypted Short" to a 4-pass algorithm.
- * Recommended a random initial nonce when incrementing.
- * Clarified what SNI LBs should do with unknown QUIC versions.

E.5. since draft-ietf-quic-load-balancers-08

- * Eliminate Dynamic SID allocation
- * Eliminated server use bytes

E.6. since draft-ietf-quic-load-balancers-07

- * Shortened SSCID nonce minimum length to 4 bytes
- * Removed RSCID from Retry token body
- * Simplified CID formats
- * Shrunk size of SID table

E.7. since draft-ietf-quic-load-balancers-06

- * Added interoperability with DTLS
- * Changed "non-compliant" to "unroutable"
- * Changed "arbitrary" algorithm to "fallback"
- * Revised security considerations for mistrustful tenants
- * Added retry service considerations for non-Initial packets

E.8. since draft-ietf-quic-load-balancers-05

- * Added low-config CID for further discussion
- * Complete revision of shared-state Retry Token
- * Added YANG model

- * Updated configuration limits to ensure CID entropy
 - * Switched to notation from quic-transport
- E.9. since draft-ietf-quic-load-balancers-04
- * Rearranged the shared-state retry token to simplify token processing
 - * More compact timestamp in shared-state retry token
 - * Revised server requirements for shared-state retries
 - * Eliminated zero padding from the test vectors
 - * Added server use bytes to the test vectors
 - * Additional compliant DCID criteria
- E.10. since-draft-ietf-quic-load-balancers-03
- * Improved Config Rotation text
 - * Added stream cipher test vectors
 - * Deleted the Obfuscated CID algorithm
- E.11. since-draft-ietf-quic-load-balancers-02
- * Replaced stream cipher algorithm with three-pass version
 - * Updated Retry format to encode info for required TPs
 - * Added discussion of version invariance
 - * Cleaned up text about config rotation
 - * Added Reset Oracle and limited configuration considerations
 - * Allow dropped long-header packets for known QUIC versions
- E.12. since-draft-ietf-quic-load-balancers-01
- * Test vectors for load balancer decoding
 - * Deleted remnants of in-band protocol
 - * Light edit of Retry Services section

- * Discussed load balancer chains
- E.13. since-draft-ietf-quic-load-balancers-00
- * Removed in-band protocol from the document
- E.14. Since draft-duke-quic-load-balancers-06
- * Switch to IETF WG draft.
- E.15. Since draft-duke-quic-load-balancers-05
- * Editorial changes
 - * Made load balancer behavior independent of QUIC version
 - * Got rid of token in stream cipher encoding, because server might not have it
 - * Defined "non-compliant DCID" and specified rules for handling them.
 - * Added psuedocode for config schema
- E.16. Since draft-duke-quic-load-balancers-04
- * Added standard for retry services
- E.17. Since draft-duke-quic-load-balancers-03
- * Renamed Plaintext CID algorithm as Obfuscated CID
 - * Added new Plaintext CID algorithm
 - * Updated to allow 20B CIDs
 - * Added self-encoding of CID length
- E.18. Since draft-duke-quic-load-balancers-02
- * Added Config Rotation
 - * Added failover mode
 - * Tweaks to existing CID algorithms
 - * Added Block Cipher CID algorithm

- * Reformatted QUIC-LB packets

E.19. Since draft-duke-quic-load-balancers-01

- * Complete rewrite
- * Supports multiple security levels
- * Lightweight messages

E.20. Since draft-duke-quic-load-balancers-00

- * Converted to markdown
- * Added variable length connection IDs

Authors' Addresses

Martin Duke
Google
Email: martin.h.duke@gmail.com

Nick Banks
Microsoft
Email: nibanks@microsoft.com

Christian Huitema
Private Octopus Inc.
Email: huitema@huitema.net

Network Working Group
Internet-Draft
Intended status: Informational
Expires: 8 October 2022

M. Kuehlewind
Ericsson
B. Trammell
Google Switzerland GmbH
6 April 2022

Manageability of the QUIC Transport Protocol
draft-ietf-quic-manageability-16

Abstract

This document discusses manageability of the QUIC transport protocol, focusing on the implications of QUIC's design and wire image on network operations involving QUIC traffic. It is intended as a "user's manual" for the wire image, providing guidance for network operators and equipment vendors who rely on the use of transport-aware network functions.

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 October 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this 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	3
2. Features of the QUIC Wire Image	4
2.1. QUIC Packet Header Structure	4
2.2. Coalesced Packets	6
2.3. Use of Port Numbers	7
2.4. The QUIC Handshake	7
2.5. Integrity Protection of the Wire Image	12
2.6. Connection ID and Rebinding	12
2.7. Packet Numbers	13
2.8. Version Negotiation and Greasing	13
3. Network-Visible Information about QUIC Flows	14
3.1. Identifying QUIC Traffic	14
3.1.1. Identifying Negotiated Version	15
3.1.2. First Packet Identification for Garbage Rejection	15
3.2. Connection Confirmation	15
3.3. Distinguishing Acknowledgment Traffic	16
3.4. Server Name Indication (SNI)	16
3.4.1. Extracting Server Name Indication (SNI) Information	16
3.5. Flow Association	18
3.6. Flow Teardown	18
3.7. Flow Symmetry Measurement	19
3.8. Round-Trip Time (RTT) Measurement	19
3.8.1. Measuring Initial RTT	19
3.8.2. Using the Spin Bit for Passive RTT Measurement	19
4. Specific Network Management Tasks	21
4.1. Passive Network Performance Measurement and Troubleshooting	21
4.2. Stateful Treatment of QUIC Traffic	22
4.3. Address Rewriting to Ensure Routing Stability	23
4.4. Server Cooperation with Load Balancers	24
4.5. Filtering Behavior	24
4.6. UDP Blocking, Throttling, and NAT Binding	24
4.7. DDoS Detection and Mitigation	25
4.8. Quality of Service Handling and ECMP Routing	27
4.9. Handling ICMP Messages	27
4.10. Guiding Path MTU	27

5. IANA Considerations	29
6. Security Considerations	29
7. Contributors	29
8. Acknowledgments	30
9. References	30
9.1. Normative References	30
9.2. Informative References	31
Authors' Addresses	34

1. Introduction

QUIC [QUIC-TRANSPORT] is a new transport protocol that is encapsulated in UDP. QUIC integrates TLS [QUIC-TLS] to encrypt all payload data and most control information. QUIC version 1 was designed primarily as a transport for HTTP, with the resulting protocol being known as HTTP/3 [QUIC-HTTP].

This document provides guidance for network operations that manage QUIC traffic. This includes guidance on how to interpret and utilize information that is exposed by QUIC to the network, requirements and assumptions of the QUIC design with respect to network treatment, and a description of how common network management practices will be impacted by QUIC.

QUIC is an end-to-end transport protocol. No information in the protocol header, even that which can be inspected, is mutable by the network. This is enforced through integrity protection of the wire image [WIRE-IMAGE]. Encryption of most transport-layer control signaling means that less information is visible to the network than is the case with TCP.

Integrity protection can also simplify troubleshooting at the end points as none of the nodes on the network path can modify transport layer information. However, it means in-network operations that depend on modification of data (for examples, see [RFC9065]) are not possible without the cooperation of a QUIC endpoint. Such cooperation might be possible with the introduction of a proxy which authenticates as an endpoint. Proxy operations are not in scope for this document.

Network management is not a one-size-fits-all endeavour: practices considered necessary or even mandatory within enterprise networks with certain compliance requirements, for example, would be impermissible on other networks without those requirements. The presence of a particular practice in this document should therefore not be construed as a recommendation to apply it. For each practice, this document describes what is and is not possible with the QUIC transport protocol as defined.

This document focuses solely on network management practices that observe traffic on the wire. Replacement of troubleshooting based on observation with active measurement techniques, for example, is therefore out of scope. A more generalized treatment of network management operations on encrypted transports is given in [RFC9065].

QUIC-specific terminology used in this document is defined in [QUIC-TRANSPORT].

2. Features of the QUIC Wire Image

This section discusses those aspects of the QUIC transport protocol that have an impact on the design and operation of devices that forward QUIC packets. This section is therefore primarily considering the unencrypted part of QUIC's wire image [WIRE-IMAGE], which is defined as the information available in the packet header in each QUIC packet, and the dynamics of that information. Since QUIC is a versioned protocol, the wire image of the header format can also change from version to version. However, the field that identifies the QUIC version in some packets, and the format of the Version Negotiation Packet, are both inspectable and invariant [QUIC-INVARIANTS].

This document addresses version 1 of the QUIC protocol, whose wire image is fully defined in [QUIC-TRANSPORT] and [QUIC-TLS]. Features of the wire image described herein may change in future versions of the protocol, except when specified as an invariant [QUIC-INVARIANTS], and cannot be used to identify QUIC as a protocol or to infer the behavior of future versions of QUIC.

2.1. QUIC Packet Header Structure

QUIC packets may have either a long header or a short header. The first bit of the QUIC header is the Header Form bit, and indicates which type of header is present. The purpose of this bit is invariant across QUIC versions.

The long header exposes more information. It contains a version number, as well as source and destination connection IDs for associating packets with a QUIC connection. The definition and location of these fields in the QUIC long header are invariant for future versions of QUIC, although future versions of QUIC may provide additional fields in the long header [QUIC-INVARIANTS].

In version 1 of QUIC, the long header is used during connection establishment to transmit crypto handshake data, perform version negotiation, retry, and send 0-RTT data.

Short headers contain only an optional destination connection ID and the spin bit for RTT measurement. In version 1 of QUIC, they are used after connection establishment.

The following information is exposed in QUIC packet headers in all versions of QUIC:

- * **version number:** the version number is present in the long header, and identifies the version used for that packet. During Version Negotiation (see Section 17.2.1 of [QUIC-TRANSPORT] and Section 2.8), the version number field has a special value (0x00000000) that identifies the packet as a Version Negotiation packet. QUIC version 1 uses version 0x00000001. Operators should expect to observe packets with other version numbers as a result of various Internet experiments, future standards, and greasing ([RFC7801]). All deployed versions are maintained in an IANA registry (see Section 22.2 of [QUIC-TRANSPORT]).
- * **source and destination connection ID:** short and long headers carry a destination connection ID, a variable-length field that can be used to identify the connection associated with a QUIC packet, for load-balancing and NAT rebinding purposes; see Section 4.4 and Section 2.6. Long packet headers additionally carry a source connection ID. The source connection ID corresponds to the destination connection ID the source would like to have on packets sent to it, and is only present on long headers. On long header packets, the length of the connection IDs is also present; on short header packets, the length of the destination connection ID is implicit.

In version 1 of QUIC, the following additional information is exposed:

- * **"fixed bit":** The second-most-significant bit of the first octet of most QUIC packets of the current version is set to 1, enabling endpoints to demultiplex with other UDP-encapsulated protocols. Even though this bit is fixed in the version 1 specification, endpoints might use an extension that varies the bit. Therefore, observers cannot reliably use it as an identifier for QUIC.
- * **latency spin bit:** The third-most-significant bit of the first octet in the short header for version 1. The spin bit is set by endpoints such that tracking edge transitions can be used to passively observe end-to-end RTT. See Section 3.8.2 for further details.

- * header type: The long header has a 2 bit packet type field following the Header Form and fixed bits. Header types correspond to stages of the handshake; see Section 17.2 of [QUIC-TRANSPORT] for details.
- * length: The length of the remaining QUIC packet after the length field, present on long headers. This field is used to implement coalesced packets during the handshake (see Section 2.2).
- * token: Initial packets may contain a token, a variable-length opaque value optionally sent from client to server, used for validating the client's address. Retry packets also contain a token, which can be used by the client in an Initial packet on a subsequent connection attempt. The length of the token is explicit in both cases.

Retry (Section 17.2.5 of [QUIC-TRANSPORT]) and Version Negotiation (Section 17.2.1 of [QUIC-TRANSPORT]) packets are not encrypted or protected in any way. For other kinds of packets, version 1 of QUIC cryptographically obfuscates other information in the packet headers:

- * packet number: All packets except Version Negotiation and Retry packets have an associated packet number; however, this packet number is encrypted, and therefore not of use to on-path observers. The offset of the packet number can be decoded in long headers, while it is implicit (depending on destination connection ID length) in short headers. The length of the packet number is cryptographically protected.
- * key phase: The Key Phase bit, present in short headers, specifies the keys used to encrypt the packet to support key rotation. The Key Phase bit is cryptographically protected.

2.2. Coalesced Packets

Multiple QUIC packets may be coalesced into a single UDP datagram, with a datagram carrying one or more long header packets followed by zero or one short header packets. When packets are coalesced, the Length fields in the long headers are used to separate QUIC packets; see Section 12.2 of [QUIC-TRANSPORT]. The Length field is variable length, and its position in the header is also variable depending on the length of the source and destination connection ID; see Section 17.2 of [QUIC-TRANSPORT].

2.3. Use of Port Numbers

Applications that have a mapping for TCP as well as QUIC are expected to use the same port number for both services. However, as for all other IETF transports [RFC7605], there is no guarantee that a specific application will use a given registered port, or that a given port carries traffic belonging to the respective registered service, especially when application layer information is encrypted. For example, [QUIC-HTTP] specifies the use of the HTTP Alternative Services mechanism [RFC7838] for discovery of HTTP/3 services on other ports.

Further, as QUIC has a connection ID, it is also possible to maintain multiple QUIC connections over one 5-tuple (protocol, source and destination IP address, and source and destination port). However, if the connection ID is zero-length, all packets of the 5-tuple likely belong to the same QUIC connection.

2.4. The QUIC Handshake

New QUIC connections are established using a handshake, which is distinguishable on the wire and contains some information that can be passively observed.

To illustrate the information visible in the QUIC wire image during the handshake, we first show the general communication pattern visible in the UDP datagrams containing the QUIC handshake, then examine each of the datagrams in detail.

The QUIC handshake can normally be recognized on the wire through four flights of datagrams labelled "Client Initial", "Server Initial", "Client Completion", and "Server Completion", as illustrated in Figure 1.

A handshake starts with the client sending one or more datagrams containing Initial packets, detailed in Figure 2, which elicits the Server Initial response detailed in Figure 3 typically containing three types of packets: Initial packet(s) with the beginning of the server's side of the TLS handshake, Handshake packet(s) with the rest of the server's portion of the TLS handshake, and 1-RTT packet(s), if present.

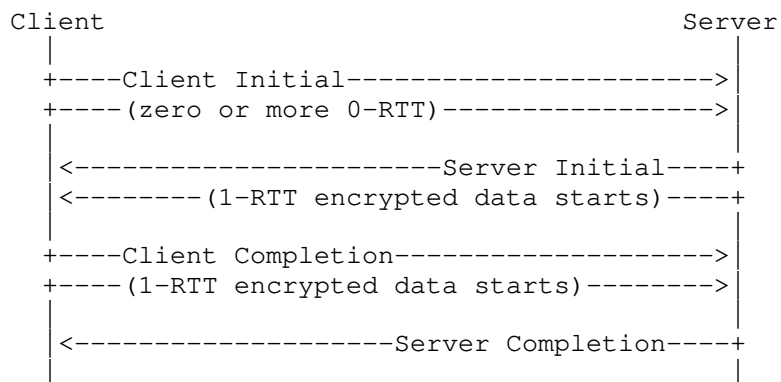


Figure 1: General communication pattern visible in the QUIC handshake

As shown here, the client can send 0-RTT data as soon as it has sent its Client Hello, and the server can send 1-RTT data as soon as it has sent its Server Hello. The Client Completion flight contains at least one Handshake packet and could also include an Initial packet. QUIC packets in separate contexts during the handshake can be coalesced (see Section 2.2) in order to reduce the number of UDP datagrams sent during the handshake.

Handshake packets can arrive out-of-order without impacting the handshake as long as the reordering was not accompanied by extensive delays that trigger a spurious Probe Timeout (Section 6.2 of RFC9002). If QUIC packets get lost or reordered, packets belonging to the same flight might not be observed in close time succession, though the sequence of the flights will not change, because one flight depends upon the peer's previous flight.

Datagrams that contain an Initial packet (Client Initial, Server Initial, and some Client Completion) contain at least 1200 octets of UDP payload. This protects against amplification attacks and verifies that the network path meets the requirements for the minimum QUIC IP packet size; see Section 14 of [QUIC-TRANSPORT]. This is accomplished by either adding PADDING frames within the Initial packet, coalescing other packets with the Initial packet, or leaving unused payload in the UDP packet after the Initial packet. A network path needs to be able to forward at least this size of packet for QUIC to be used.

The content of Initial packets is encrypted using Initial Secrets, which are derived from a per-version constant and the client's destination connection ID. That content is therefore observable by any on-path device that knows the per-version constant and is considered visible in this illustration. The content of QUIC Handshake packets is encrypted using keys established during the initial handshake exchange, and is therefore not visible.

Initial, Handshake, and 1-RTT packets belong to different cryptographic and transport contexts. The Client Completion (Figure 4) and the Server Completion (Figure 5) flights conclude the Initial and Handshake contexts, by sending final acknowledgments and CRYPTO frames.

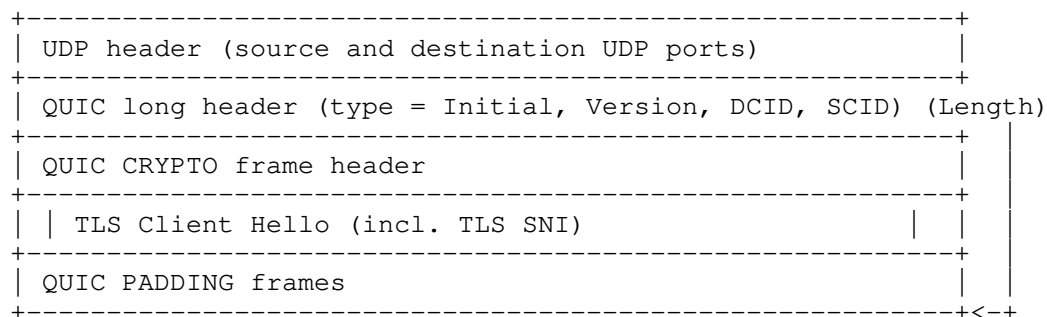


Figure 2: Example Client Initial datagram without 0-RTT

A Client Initial packet exposes the version, source and destination connection IDs without encryption. The payload of the Initial packet is protected using the Initial secret. The complete TLS Client Hello, including any TLS Server Name Indication (SNI) present, is sent in one or more CRYPTO frames across one or more QUIC Initial packets.

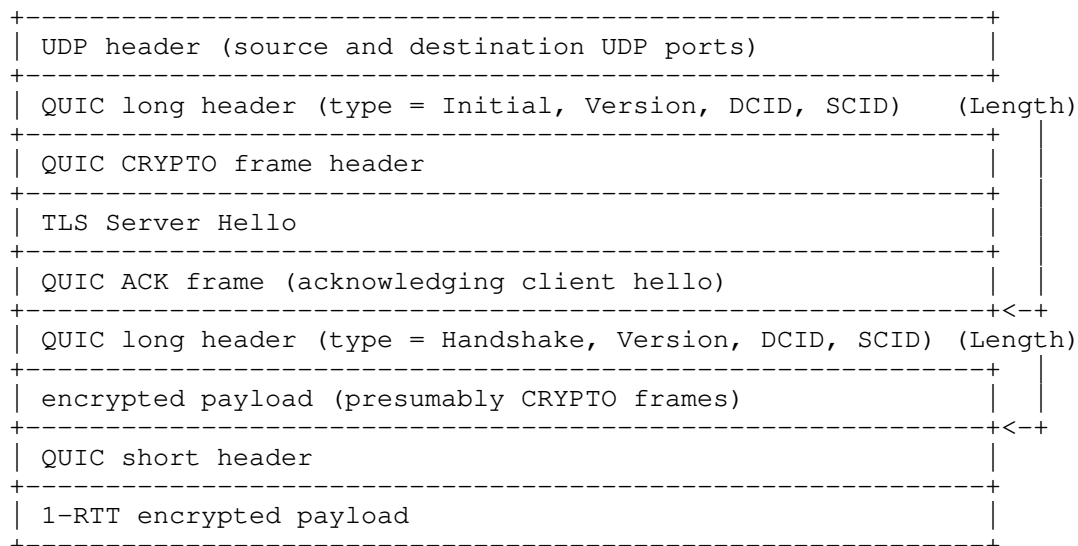


Figure 3: Coalesced Server Initial datagram pattern

The Server Initial datagram also exposes version number, source and destination connection IDs in the clear; the payload of the Initial packet(s) is protected using the Initial secret.

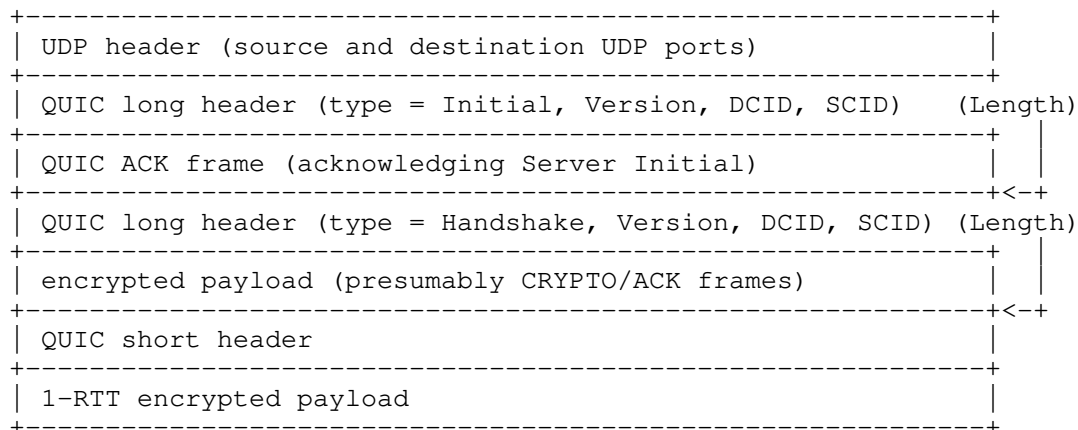


Figure 4: Coalesced Client Completion datagram pattern

The Client Completion flight does not expose any additional information; however, as the destination connection ID is server-selected, it usually is not the same ID that is sent in the Client Initial. Client Completion flights contain 1-RTT packets which indicate the handshake has completed (see Section 3.2) on the client, and for three-way handshake RTT estimation as in Section 3.8.

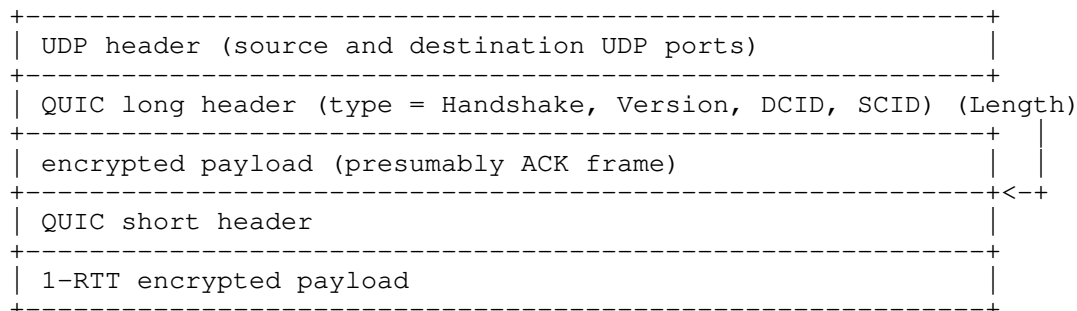


Figure 5: Coalesced Server Completion datagram pattern

Similar to Client Completion, Server Completion also exposes no additional information; observing it serves only to determine that the handshake has completed.

When the client uses 0-RTT data, the Client Initial flight can also include one or more 0-RTT packets, as shown in Figure 6.

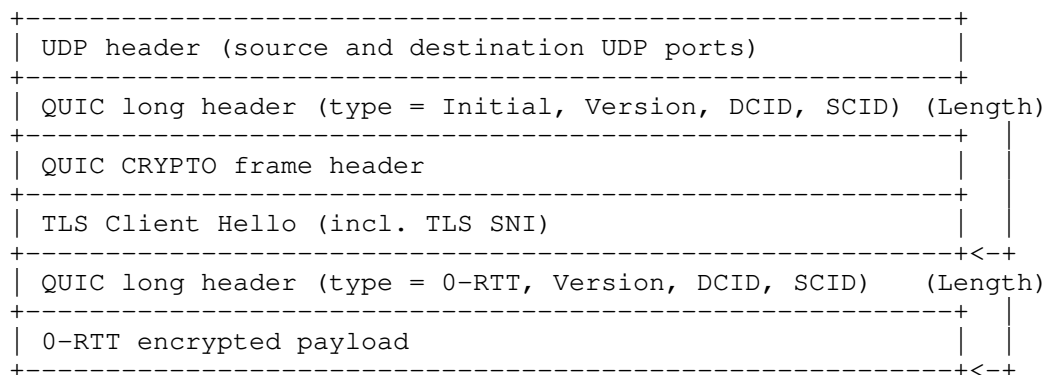


Figure 6: Coalesced 0-RTT Client Initial datagram

When a 0-RTT packet is coalesced with an Initial packet, the datagram will be padded to 1200 bytes. Additional datagrams containing only 0-RTT packets with long headers can be sent after the client Initial packet(s), containing more 0-RTT data. The amount of 0-RTT protected

data that can be sent in the first flight is limited by the initial congestion window, typically to around 10 packets (see Section 7.2 of [QUIC-RECOVERY]).

2.5. Integrity Protection of the Wire Image

As soon as the cryptographic context is established, all information in the QUIC header, including exposed information, is integrity protected. Further, information that was exposed in packets sent before the cryptographic context was established is validated during the cryptographic handshake. Therefore, devices on path cannot alter any information or bits in QUIC packets. Such alterations would cause the integrity check to fail, which results in the receiver discarding the packet. Some parts of Initial packets could be altered by removing and re-applying the authenticated encryption without immediate discard at the receiver. However, the cryptographic handshake validates most fields and any modifications in those fields will result in connection establishment failing later.

2.6. Connection ID and Rebinding

The connection ID in the QUIC packet headers allows association of QUIC packets using information independent of the 5-tuple. This allows rebinding of a connection after one of the endpoints - usually the client - has experienced an address change. Further it can be used by in-network devices to ensure that related 5-tuple flows are appropriately balanced together.

Client and server each choose a connection ID during the handshake; for example, a server might request that a client use a connection ID, whereas the client might choose a zero-length value. Connection IDs for either endpoint may change during the lifetime of a connection, with the new connection ID being supplied via encrypted frames (see Section 5.1 of [QUIC-TRANSPORT]). Therefore, observing a new connection ID does not necessarily indicate a new connection.

[QUIC_LB] specifies algorithms for encoding the server mapping in a connection ID in order to share this information with selected on-path devices such as load balancers. Server mappings should only be exposed to selected entities. Uncontrolled exposure would allow linkage of multiple IP addresses to the same host if the server also supports migration that opens an attack vector on specific servers or pools. The best way to obscure an encoding is to appear random to any other observers, which is most rigorously achieved with encryption. As a result, any attempt to infer information from specific parts of a connection ID is unlikely to be useful.

2.7. Packet Numbers

The Packet Number field is always present in the QUIC packet header in version 1; however, it is always encrypted. The encryption key for packet number protection on Initial packets -- which are sent before cryptographic context establishment -- is specific to the QUIC version, while packet number protection on subsequent packets uses secrets derived from the end-to-end cryptographic context. Packet numbers are therefore not part of the wire image that is visible to on-path observers.

2.8. Version Negotiation and Greasing

Version Negotiation packets are used by the server to indicate that a requested version from the client is not supported (see Section 6 of [QUIC-TRANSPORT]). Version Negotiation packets are not intrinsically protected, but future QUIC versions could use later encrypted messages to verify that they were authentic. Therefore, any modification of this list will be detected and may cause the endpoints to terminate the connection attempt.

Also note that the list of versions in the Version Negotiation packet may contain reserved versions. This mechanism is used to avoid ossification in the implementation on the selection mechanism. Further, a client may send an Initial packet with a reserved version number to trigger version negotiation. In the Version Negotiation packet, the connection IDs of the client's Initial packet are reflected to provide a proof of return-routability. Therefore, changing this information will also cause the connection to fail.

QUIC is expected to evolve rapidly, so new versions, both experimental and IETF standard versions, will be deployed on the Internet more often than with other commonly deployed Internet- and transport-layer protocols. Use of the version number field for traffic recognition will therefore behave differently than with these protocols. Using a particular version number to recognize valid QUIC traffic is likely to persistently miss a fraction of QUIC flows, and completely fail in the near future. Reliance on the version number field for the purposes of admission control is similarly likely to rapidly lead to unintended failure modes. Admission of QUIC traffic regardless of version avoids these failure modes, avoids unnecessary deployment delays, and supports continuous version-based evolution.

3. Network-Visible Information about QUIC Flows

This section addresses the different kinds of observations and inferences that can be made about QUIC flows by a passive observer in the network based on the wire image in Section 2. Here we assume a bidirectional observer (one that can see packets in both directions in the sequence in which they are carried on the wire) unless noted, but typically without access to any keying information.

3.1. Identifying QUIC Traffic

The QUIC wire image is not specifically designed to be distinguishable from other UDP traffic by a passive observer in the network. While certain QUIC applications may be heuristically identifiable on a per-application basis, there is no general method for distinguishing QUIC traffic from otherwise-unclassifiable UDP traffic on a given link. Any unrecognized UDP traffic may therefore be QUIC traffic.

At the time of writing, two application bindings for QUIC have been published or adopted by the IETF: HTTP/3 [QUIC-HTTP] and DNS over Dedicated QUIC Connections [I-D.ietf-dprive-dnssoquic]. These are both known at the time of writing to have active Internet deployments, so an assumption that all QUIC traffic is HTTP/3 is not valid. HTTP/3 uses UDP port 443 by convention but various methods can be used to specify alternate port numbers. Simple assumptions about whether a given flow is using QUIC based upon a UDP port number may therefore not hold; see also Section 5 of [RFC7605].

While the second-most-significant bit (0x40) of the first octet is set to 1 in most QUIC packets of the current version (see Section 2.1 and Section 17 of [QUIC-TRANSPORT]), this method of recognizing QUIC traffic is not reliable. First, it only provides one bit of information and is prone to collision with UDP-based protocols other than those considered in [RFC7983]. Second, this feature of the wire image is not invariant [QUIC-INVARIANTS] and may change in future versions of the protocol, or even be negotiated during the handshake via the use of an extension [QUIC-GREASE].

Even though transport parameters transmitted in the client's Initial packet are observable by the network, they cannot be modified by the network without causing connection failure. Further, the reply from the server cannot be observed, so observers on the network cannot know which parameters are actually in use.

3.1.1. Identifying Negotiated Version

An in-network observer assuming that a set of packets belongs to a QUIC flow might infer the version number in use by observing the handshake: for QUIC version 1, if the version number in the Initial packet from a client is the same as the version number in the Initial packet of the server response, that version has been accepted by both endpoints to be used for the rest of the connection.

The negotiated version cannot be identified for flows for which a handshake is not observed, such as in the case of connection migration; however, it might be possible to associate a flow with a flow for which a version has been identified; see Section 3.5.

3.1.2. First Packet Identification for Garbage Rejection

A related question is whether the first packet of a given flow on a port known to be associated with QUIC is a valid QUIC packet. This determination supports in-network filtering of garbage UDP packets (reflection attacks, random backscatter, etc.). While heuristics based on the first byte of the packet (packet type) could be used to separate valid from invalid first packet types, the deployment of such heuristics is not recommended, as bits in the first byte may have different meanings in future versions of the protocol.

3.2. Connection Confirmation

This document focuses on QUIC version 1, and this Connection Confirmation section applies only to packets belonging to QUIC version 1 flows; for purposes of on-path observation, it assumes that these packets have been identified as such through the observation of a version number exchange as described above.

Connection establishment uses Initial and Handshake packets containing a TLS handshake, and Retry packets that do not contain parts of the handshake. Connection establishment can therefore be detected using heuristics similar to those used to detect TLS over TCP. A client initiating a connection may also send data in 0-RTT packets directly after the Initial packet containing the TLS Client Hello. Since packets may be reordered or lost in the network, 0-RTT packets could be seen before the Initial packet.

Note that in this version of QUIC, clients send Initial packets before servers do, servers send Handshake packets before clients do, and only clients send Initial packets with tokens. Therefore, an endpoint can be identified as a client or server by an on-path observer. An attempted connection after Retry can be detected by correlating the contents of the Retry packet with the Token and the Destination Connection ID fields of the new Initial packet.

3.3. Distinguishing Acknowledgment Traffic

Some deployed in-network functions distinguish packets that carry only acknowledgment (ACK-only) information from packets carrying upper-layer data in order to attempt to enhance performance, for example by queueing ACKs differently or manipulating ACK signaling [RFC3449]. Distinguishing ACK packets is possible in TCP, but is not supported by QUIC, since acknowledgment signaling is carried inside QUIC's encrypted payload, and ACK manipulation is impossible. Specifically, heuristics attempting to distinguish ACK-only packets from payload-carrying packets based on packet size are likely to fail, and are not recommended to use as a way to construe internals of QUIC's operation as those mechanisms can change, e.g., due to the use of extensions.

3.4. Server Name Indication (SNI)

The client's TLS ClientHello may contain a Server Name Indication (SNI) [RFC6066] extension, by which the client reveals the name of the server it intends to connect to, in order to allow the server to present a certificate based on that name. It may also contain an Application-Layer Protocol Negotiation (ALPN) [RFC7301] extension, by which the client exposes the names of application-layer protocols it supports; an observer can deduce that one of those protocols will be used if the connection continues.

Work is currently underway in the TLS working group to encrypt the contents of the ClientHello in TLS 1.3 [TLS-ECH]. This would make SNI-based application identification impossible by on-path observation for QUIC and other protocols that use TLS.

3.4.1. Extracting Server Name Indication (SNI) Information

If the ClientHello is not encrypted, SNI can be derived from the client's Initial packet(s) by calculating the Initial secret to decrypt the packet payload and parsing the QUIC CRYPTO frame(s) containing the TLS ClientHello.

As both the derivation of the Initial secret and the structure of the Initial packet itself are version-specific, the first step is always to parse the version number (the second through fifth bytes of the long header). Note that only long header packets carry the version number, so it is necessary to also check if the first bit of the QUIC packet is set to 1, indicating a long header.

Note that proprietary QUIC versions, that have been deployed before standardization, might not set the first bit in a QUIC long header packet to 1. However, it is expected that these versions will gradually disappear over time and therefore do not require any special consideration or treatment.

When the version has been identified as QUIC version 1, the packet type needs to be verified as an Initial packet by checking that the third and fourth bits of the header are both set to 0. Then the Destination Connection ID needs to be extracted from the packet. The Initial secret is calculated using the version-specific Initial salt, as described in Section 5.2 of [QUIC-TLS]. The length of the connection ID is indicated in the 6th byte of the header followed by the connection ID itself.

Note that subsequent Initial packets might contain a Destination Connection ID other than the one used to generate the Initial secret. Therefore, attempts to decrypt these packets using the procedure above might fail unless the Initial secret is retained by the observer.

To determine the end of the packet header and find the start of the payload, the packet number length, the source connection ID length, and the token length need to be extracted. The packet number length is defined by the seventh and eighth bits of the header as described in Section 17.2 of [QUIC-TRANSPORT], but is protected as described in Section 5.4 of [QUIC-TLS]. The source connection ID length is specified in the byte after the destination connection ID. The token length, which follows the source connection ID, is a variable-length integer as specified in Section 16 of [QUIC-TRANSPORT].

After decryption, the client's Initial packet(s) can be parsed to detect the CRYPTO frame(s) that contains the TLS ClientHello, which then can be parsed similarly to TLS over TCP connections. Note that there can be multiple CRYPTO frames spread out over one or more Initial packets, and they might not be in order, so reassembling the CRYPTO stream by parsing offsets and lengths is required. Further, the client's Initial packet(s) may contain other frames, so the first bytes of each frame need to be checked to identify the frame type and determine whether the frame can be skipped over. Note that the length of the frames is dependent on the frame type; see Section 18

of [QUIC-TRANSPORT]. E.g., PADDING frames, each consisting of a single zero byte, may occur before, after, or between CRYPTO frames. However, extensions might define additional frame types. If an unknown frame type is encountered, it is impossible to know the length of that frame which prevents skipping over it, and therefore parsing fails.

3.5. Flow Association

The QUIC connection ID (see Section 2.6) is designed to allow a coordinating on-path device, such as a load-balancer, to associate two flows when one of the endpoints changes address. This change can be due to NAT rebinding or address migration.

The connection ID must change upon intentional address change by an endpoint, and connection ID negotiation is encrypted, so it is not possible for a passive observer to link intended changes of address using the connection ID.

When one endpoint's address unintentionally changes, as is the case with NAT rebinding, an on-path observer may be able to use the connection ID to associate the flow on the new address with the flow on the old address.

A network function that attempts to use the connection ID to associate flows must be robust to the failure of this technique. Since the connection ID may change multiple times during the lifetime of a connection, packets with the same 5-tuple but different connection IDs might or might not belong to the same connection. Likewise, packets with the same connection ID but different 5-tuples might not belong to the same connection, either.

Connection IDs should be treated as opaque; see Section 4.4 for caveats regarding connection ID selection at servers.

3.6. Flow Teardown

QUIC does not expose the end of a connection; the only indication to on-path devices that a flow has ended is that packets are no longer observed. Stateful devices on path such as NATs and firewalls must therefore use idle timeouts to determine when to drop state for QUIC flows; see Section 4.2.

3.7. Flow Symmetry Measurement

QUIC explicitly exposes which side of a connection is a client and which side is a server during the handshake. In addition, the symmetry of a flow (whether primarily client-to-server, primarily server-to-client, or roughly bidirectional, as input to basic traffic classification techniques) can be inferred through the measurement of data rate in each direction. Note that QUIC packets containing only control frames (such as ACK-only packets) may be padded. Padding, though optional, may conceal connection roles or flow symmetry information.

3.8. Round-Trip Time (RTT) Measurement

The round-trip time (RTT) of QUIC flows can be inferred by observation once per flow, during the handshake, as in passive TCP measurement; this requires parsing of the QUIC packet header and recognition of the handshake, as illustrated in Section 2.4. It can also be inferred during the flow's lifetime, if the endpoints use the spin bit facility described below and in Section 17.3.1 of [QUIC-TRANSPORT].

3.8.1. Measuring Initial RTT

In the common case, the delay between the client's Initial packet (containing the TLS ClientHello) and the server's Initial packet (containing the TLS ServerHello) represents the RTT component on the path between the observer and the server. The delay between the server's first Handshake packet and the Handshake packet sent by the client represents the RTT component on the path between the observer and the client. While the client may send 0-RTT packets after the Initial packet during connection re-establishment, these can be ignored for RTT measurement purposes.

Handshake RTT can be measured by adding the client-to-observer and observer-to-server RTT components together. This measurement necessarily includes all transport- and application-layer delay at both endpoints.

3.8.2. Using the Spin Bit for Passive RTT Measurement

The spin bit provides a version-specific method to measure per-flow RTT from observation points on the network path throughout the duration of a connection. See Section 17.4 of [QUIC-TRANSPORT] for the definition of the spin bit in Version 1 of QUIC. Endpoint participation in spin bit signaling is optional. That is, while its location is fixed in this version of QUIC, an endpoint can unilaterally choose to not support "spinning" the bit.

Use of the spin bit for RTT measurement by devices on path is only possible when both endpoints enable it. Some endpoints may disable use of the spin bit by default, others only in specific deployment scenarios, e.g., for servers and clients where the RTT would reveal the presence of a VPN or proxy. To avoid making these connections identifiable based on the usage of the spin bit, all endpoints randomly disable "spinning" for at least one eighth of connections, even if otherwise enabled by default. An endpoint not participating in spin bit signaling for a given connection can use a fixed spin value for the duration of the connection, or can set the bit randomly on each packet sent.

When in use, the latency spin bit in each direction changes value once per RTT any time that both endpoints are sending packets continuously. An on-path observer can observe the time difference between edges (changes from 1 to 0 or 0 to 1) in the spin bit signal in a single direction to measure one sample of end-to-end RTT. This mechanism follows the principles of protocol measurability laid out in [IPIM].

Note that this measurement, as with passive RTT measurement for TCP, includes all transport protocol delay (e.g., delayed sending of acknowledgments) and/or application layer delay (e.g., waiting for a response to be generated). It therefore provides devices on path a good instantaneous estimate of the RTT as experienced by the application.

However, application-limited and flow-control-limited senders can have application and transport layer delay, respectively, that are much greater than network RTT. When the sender is application-limited and e.g., only sends small amount of periodic application traffic, where that period is longer than the RTT, measuring the spin bit provides information about the application period, not the network RTT.

Since the spin bit logic at each endpoint considers only samples from packets that advance the largest packet number, signal generation itself is resistant to reordering. However, reordering can cause problems at an observer by causing spurious edge detection and therefore inaccurate (i.e., lower) RTT estimates, if reordering occurs across a spin-bit flip in the stream.

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 lost or reordered edges in the spin signal, as well as application or flow control limitation; for example, QoF [TMA-QOF] rejects component RTTs significantly higher than RTTs over the history of the flow. These heuristics may use the handshake RTT as an initial RTT estimate for a given flow. Usually such heuristics would also detect if the spin is either constant or randomly set for a connection.

An on-path observer that can see traffic in both directions (from client to server and from server to client) can also use the spin bit to measure "upstream" and "downstream" component RTT; i.e, the component of the end-to-end RTT attributable to the paths between the observer and the server and the observer and the client, respectively. It does this by measuring the delay between a spin edge observed in the upstream direction and that observed in the downstream direction, and vice versa.

Raw RTT samples generated using these techniques can be processed in various ways to generate useful network performance metrics. A simple linear smoothing or moving minimum filter can be applied to the stream of RTT samples to get a more stable estimate of application-experienced RTT. RTT samples measured from the spin bit can also be used to generate RTT distribution information, including minimum RTT (which approximates network RTT over longer time windows) and RTT variance (which approximates one-way packet delay variance as seen by an application end-point).

4. Specific Network Management Tasks

In this section, we review specific network management and measurement techniques and how QUIC's design impacts them.

4.1. Passive Network Performance Measurement and Troubleshooting

Limited RTT measurement is possible by passive observation of QUIC traffic; see Section 3.8. No passive measurement of loss is possible with the present wire image. Limited observation of upstream congestion may be possible via the observation of CE markings in the IP header [RFC3168] on ECN-enabled QUIC traffic.

On-path devices can also make measurements of RTT, loss and other performance metrics when information is carried in an additional network-layer packet header (Section 6 of [I-D.ietf-tsvwg-transport-encrypt] describes use of operations, administration and management (OAM) information). Using network-layer approaches also has the advantage that common observation and analysis tools can be consistently used for multiple transport protocols, however, these techniques are often limited to measurements within one or multiple cooperating domains.

4.2. Stateful Treatment of QUIC Traffic

Stateful treatment of QUIC traffic (e.g., at a firewall or NAT middlebox) is possible through QUIC traffic and version identification (Section 3.1) and observation of the handshake for connection confirmation (Section 3.2). The lack of any visible end-of-flow signal (Section 3.6) means that this state must be purged either through timers or through least-recently-used eviction, depending on application requirements.

While QUIC has no clear network-visible end-of-flow signal and therefore does require timer-based state removal, the QUIC handshake indicates confirmation by both ends of a valid bidirectional transmission. As soon as the handshake completed, timers should be set long enough to also allow for short idle time during a valid transmission.

[RFC4787] requires a network state timeout that is not less than 2 minutes for most UDP traffic. However, in practice, a QUIC endpoint can experience lower timeouts, in the range of 30 to 60 seconds [QUIC-TIMEOUT].

In contrast, [RFC5382] recommends a state timeout of more than 2 hours for TCP, given that TCP is a connection-oriented protocol with well-defined closure semantics. Even though QUIC has explicitly been designed to tolerate NAT rebindings, decreasing the NAT timeout is not recommended, as it may negatively impact application performance or incentivize endpoints to send very frequent keep-alive packets.

The recommendation is therefore that, even when lower state timeouts are used for other UDP traffic, a state timeout of at least two minutes ought to be used for QUIC traffic.

If state is removed too early, this could lead to black-holing of incoming packets after a short idle period. To detect this situation, a timer at the client needs to expire before a re-establishment can happen (if at all), which would lead to unnecessarily long delays in an otherwise working connection.

Furthermore, not all endpoints use routing architectures where connections will survive a port or address change. So even when the client revives the connection, a NAT rebinding can cause a routing mismatch where a packet is not even delivered to the server that might support address migration. For these reasons, the limits in [RFC4787] are important to avoid black-holing of packets (and hence avoid interrupting the flow of data to the client), especially where devices are able to distinguish QUIC traffic from other UDP payloads.

The QUIC header optionally contains a connection ID which could provide additional entropy beyond the 5-tuple. The QUIC handshake needs to be observed in order to understand whether the connection ID is present and what length it has. However, connection IDs may be renegotiated after the handshake, and this renegotiation is not visible to the path. Therefore, using the connection ID as a flow key field for stateful treatment of flows is not recommended as connection ID changes will cause undetectable and unrecoverable loss of state in the middle of a connection. In particular, the use of the connection ID for functions that require state to make a forwarding decision is not viable as it will break connectivity, or at minimum cause long timeout-based delays before this problem is detected by the endpoints and the connection can potentially be re-established.

Use of connection IDs is specifically discouraged for NAT applications. If a NAT hits an operational limit, it is recommended to rather drop the initial packets of a flow (see also Section 4.5), which potentially triggers TCP fallback. Use of the connection ID to multiplex multiple connections on the same IP address/port pair is not a viable solution as it risks connectivity breakage, in case the connection ID changes.

4.3. Address Rewriting to Ensure Routing Stability

While QUIC's migration capability makes it possible for a connection to survive client address changes, this does not work if the routers or switches in the server infrastructure route using the address-port 4-tuple. If infrastructure routes on addresses only, NAT rebinding or address migration will cause packets to be delivered to the wrong server. [QUIC_LB] describes a way to address this problem by coordinating the selection and use of connection IDs between load-balancers and servers.

Applying address translation at a middlebox to maintain a stable address-port mapping for flows based on connection ID might seem like a solution to this problem. However, hiding information about the change of the IP address or port conceals important and security-relevant information from QUIC endpoints and as such would facilitate

amplification attacks (see Section 8 of [QUIC-TRANSPORT]). A NAT function that hides peer address changes prevents the other end from detecting and mitigating attacks as the endpoint cannot verify connectivity to the new address using QUIC PATH_CHALLENGE and PATH_RESPONSE frames.

In addition, a change of IP address or port is also an input signal to other internal mechanisms in QUIC. When a path change is detected, path-dependent variables like congestion control parameters will be reset protecting the new path from overload.

4.4. Server Cooperation with Load Balancers

In the case of networking architectures that include load balancers, the connection ID can be used as a way for the server to signal information about the desired treatment of a flow to the load balancers. Guidance on assigning connection IDs is given in [QUIC-APPLICABILITY]. [QUIC_LB] describes a system for coordinating selection and use of connection IDs between load-balancers and servers.

4.5. Filtering Behavior

[RFC4787] describes possible packet filtering behaviors that relate to NATs but is often also used in other scenarios where packet filtering is desired. Though the guidance there holds, a particularly unwise behavior admits a handful of UDP packets and then makes a decision to whether or not filter later packets in the same connection. QUIC applications are encouraged to fall back to TCP if early packets do not arrive at their destination [QUIC-APPLICABILITY], as QUIC is based on UDP and there are known blocks of UDP traffic (see Section 4.6). Admitting a few packets allows the QUIC endpoint to determine that the path accepts QUIC. Sudden drops afterwards will result in slow and costly timeouts before abandoning the connection.

4.6. UDP Blocking, Throttling, and NAT Binding

Today, UDP is the most prevalent DDoS vector, since it is easy for compromised non-admin applications to send a flood of large UDP packets (while with TCP the attacker gets throttled by the congestion controller) or to craft reflection and amplification attacks. Some networks therefore block UDP traffic. With increased deployment of QUIC, there is also an increased need to allow UDP traffic on ports used for QUIC. However, if UDP is generally enabled on these ports, UDP flood attacks may also use the same ports. One possible response to this threat is to throttle UDP traffic on the network, allocating a fixed portion of the network capacity to UDP and blocking UDP

datagrams over that cap. As the portion of QUIC traffic compared to TCP is also expected to increase over time, using such a limit is not recommended but if done, limits might need to be adapted dynamically.

Further, if UDP traffic is desired to be throttled, it is recommended to block individual QUIC flows entirely rather than dropping packets indiscriminately. When the handshake is blocked, QUIC-capable applications may fall back to TCP. However, blocking a random fraction of QUIC packets across 4-tuples will allow many QUIC handshakes to complete, preventing TCP fallback, but these connections will suffer from severe packet loss (see also Section 4.5). Therefore, UDP throttling should be realized by per-flow policing, as opposed to per-packet policing. Note that this per-flow policing should be stateless to avoid problems with stateful treatment of QUIC flows (see Section 4.2), for example blocking a portion of the space of values of a hash function over the addresses and ports in the UDP datagram. While QUIC endpoints are often able to survive address changes, e.g., by NAT rebindings, blocking a portion of the traffic based on 5-tuple hashing increases the risk of black-holing an active connection when the address changes.

Note that some source ports are assumed to be reflection attack vectors by some servers; see Section 8.1 of [QUIC-APPLICABILITY]. As a result, NAT binding to these source ports can result in that traffic being blocked.

4.7. DDoS Detection and Mitigation

On-path observation of the transport headers of packets can be used for various security functions. For example, Denial of Service (DOS) and Distributed DOS (DDoS) attacks against the infrastructure or against an endpoint can be detected and mitigated by characterising anomalous traffic. Other uses include support for security audits (e.g., verifying the compliance with ciphersuites); client and application fingerprinting for inventory; and to provide alerts for network intrusion detection and other next generation firewall functions.

Current practices in detection and mitigation of DDoS attacks generally involve classification of incoming traffic (as packets, flows, or some other aggregate) into "good" (productive) and "bad" (DDoS) traffic, and then differential treatment of this traffic to forward only good traffic. This operation is often done in a separate specialized mitigation environment through which all traffic is filtered; a generalized architecture for separation of concerns in mitigation is given in [DOTS-ARCH].

Efficient classification of this DDoS traffic in the mitigation environment is key to the success of this approach. Limited first-packet garbage detection as in Section 3.1.2 and stateful tracking of QUIC traffic as in Section 4.2 above may be useful during classification.

Note that the use of a connection ID to support connection migration renders 5-tuple based filtering insufficient to detect active flows and requires more state to be maintained by DDoS defense systems if support of migration of QUIC flows is desired. For the common case of NAT rebinding, where the client's address changes without the client's intent or knowledge, DDoS defense systems can detect a change in the client's endpoint address by linking flows based on the server's connection IDs. However, QUIC's linkability resistance ensures that a deliberate connection migration is accompanied by a change in the connection ID. In this case, the connection ID can not be used to distinguish valid, active traffic from new attack traffic.

It is also possible for endpoints to directly support security functions such as DoS classification and mitigation. Endpoints can cooperate with an in-network device directly by e.g., sharing information about connection IDs.

Another potential method could use an on-path network device that relies on pattern inferences in the traffic and heuristics or machine learning instead of processing observed header information.

However, it is questionable whether connection migrations must be supported during a DDoS attack. While unintended migration without a connection ID change can be more easily supported, it might be acceptable to not support migrations of active QUIC connections that are not visible to the network functions performing the DDoS detection. As soon as the connection blocking is detected by the client, the client may be able to rely on the 0-RTT data mechanism provided by QUIC. When clients migrate to a new path, they should be prepared for the migration to fail and attempt to reconnect quickly.

Beyond in-network DDoS protection mechanisms, TCP synccookies [RFC4937] are a well-established method of mitigating some kinds of TCP DDoS attacks. QUIC Retry packets are the functional analogue to synccookies, forcing clients to prove possession of their IP address before committing server state. However, there are safeguards in QUIC against unsolicited injection of these packets by intermediaries who do not have consent of the end server. See [QUIC_LB] for standard ways for intermediaries to send Retry packets on behalf of consenting servers.

4.8. Quality of Service Handling and ECMP Routing

It is expected that any QoS handling in the network, e.g., based on use of DiffServ Code Points (DSCPs) [RFC2475] as well as Equal-Cost Multi-Path (ECMP) routing, is applied on a per flow-basis (and not per-packet) and as such that all packets belonging to the same active QUIC connection get uniform treatment.

Using ECMP to distribute packets from a single flow across multiple network paths or any other non-uniform treatment of packets belong to the same connection could result in variations in order, delivery rate, and drop rate. As feedback about loss or delay of each packet is used as input to the congestion controller, these variations could adversely affect performance. Depending on the loss recovery mechanism implemented, QUIC may be more tolerant of packet re-ordering than traditional TCP traffic (see Section 2.7). However, the recovery mechanism used by a flow cannot be known by the network and therefore reordering tolerance should be considered as unknown.

Note that the 5-tuple of a QUIC connection can change due to migration. In this case different flows are observed by the path and maybe be treated differently, as congestion control is usually reset on migration (see also Section 3.5).

4.9. Handling ICMP Messages

Datagram Packetization Layer PMTU Discovery (PLPMTUD) can be used by QUIC to probe for the supported PMTU. PLPMTUD optionally uses ICMP messages (e.g., IPv6 Packet Too Big messages). Given known attacks with the use of ICMP messages, the use of PLPMTUD in QUIC has been designed to safely use but not rely on receiving ICMP feedback (see Section 14.2.1. of [QUIC-TRANSPORT]).

Networks are recommended to forward these ICMP messages and retain as much of the original packet as possible without exceeding the minimum MTU for the IP version when generating ICMP messages as recommended in [RFC1812] and [RFC4443].

4.10. Guiding Path MTU

Some network segments support 1500-byte packets, but can only do so by fragmenting at a lower layer before traversing a network segment with a smaller MTU, and then reassembling within the network segment. This is permissible even when the IP layer is IPv6 or IPv4 with the DF bit set, because fragmentation occurs below the IP layer. However, this process can add to compute and memory costs, leading to a bottleneck that limits network capacity. In such networks this generates a desire to influence a majority of senders to use smaller

packets, to avoid exceeding limited reassembly capacity.

For TCP, MSS clamping (Section 3.2 of [RFC4459]) is often used to change the sender's TCP maximum segment size, but QUIC requires a different approach. Section 14 of [QUIC-TRANSPORT] advises senders to probe larger sizes using Datagram Packetization Layer PMTU Discovery ([DPLPMTUD]) or Path Maximum Transmission Unit Discovery (PMTUD: [RFC1191] and [RFC8201]). This mechanism encourages senders to approach the maximum packet size, which could then cause fragmentation within a network segment of which they may not be aware.

If path performance is limited when forwarding larger packets, an on-path device should support a maximum packet size for a specific transport flow and then consistently drop all packets that exceed the configured size when the inner IPv4 packet has DF set, or IPv6 is used.

Networks with configurations that would lead to fragmentation of large packets within a network segment should drop such packets rather than fragmenting them. Network operators who plan to implement a more selective policy may start by focusing on QUIC.

QUIC flows cannot always be easily distinguished from other UDP traffic, but we assume at least some portion of QUIC traffic can be identified (see Section 3.1). For networks supporting QUIC, it is recommended that a path drops any packet larger than the fragmentation size. When a QUIC endpoint uses DPLPMTUD, it will use a QUIC probe packet to discover the PMTU. If this probe is lost, it will not impact the flow of QUIC data.

IPv4 routers generate an ICMP message when a packet is dropped because the link MTU was exceeded. [RFC8504] specifies how an IPv6 node generates an ICMPv6 Packet Too Big message (PTB) in this case. PMTUD relies upon an endpoint receiving such PTB messages [RFC8201], whereas DPLPMTUD does not reply upon these messages, but still can optionally use these to improve performance Section 4.6 of [DPLPMTUD].

A network cannot know in advance which discovery method is used by a QUIC endpoint, so it should send a PTB message in addition to dropping an oversized packet. A generated PTB message should be compliant with the validation requirements of Section 14.2.1 of [QUIC-TRANSPORT], otherwise it will be ignored for PMTU discovery. This provides a signal to the endpoint to prevent the packet size from growing too large, which can entirely avoid network segment fragmentation for that flow.

Endpoints can cache PMTU information, in the IP-layer cache. This short-term consistency between the PMTU for flows can help avoid an endpoint using a PMTU that is inefficient. The IP cache can also influence the PMTU value of other IP flows that use the same path [RFC8201][DPLPMTUD], including IP packets carrying protocols other than QUIC. The representation of an IP path is implementation-specific [RFC8201].

5. IANA Considerations

This document has no actions for IANA.

6. Security Considerations

QUIC is an encrypted and authenticated transport. That means, once the cryptographic handshake is complete, QUIC endpoints discard most packets that are not authenticated, greatly limiting the ability of an attacker to interfere with existing connections.

However, some information is still observerable, as supporting manageability of QUIC traffic inherently involves tradeoffs with the confidentiality of QUIC's control information; this entire document is therefore security-relevant.

More security considerations for QUIC are discussed in [QUIC-TRANSPORT] and [QUIC-TLS], generally considering active or passive attackers in the network as well as attacks on specific QUIC mechanism.

Version Negotiation packets do not contain any mechanism to prevent version downgrade attacks. However, future versions of QUIC that use Version Negotiation packets are required to define a mechanism that is robust against version downgrade attacks. Therefore, a network node should not attempt to impact version selection, as version downgrade may result in connection failure.

7. Contributors

The following people have contributed significant text to and/or feedback on this document:

- * Chris Box
- * Dan Druta
- * David Schinazi
- * Gorrry Fairhurst

- * Ian Swett
- * Igor Lubashev
- * Jana Iyengar
- * Jared Mauch
- * Lars Eggert
- * Lucas Purdue
- * Marcus Ihlar
- * Mark Nottingham
- * Martin Duke
- * Martin Thomson
- * Matt Joras
- * Mike Bishop
- * Nick Banks
- * Thomas Fossati
- * Sean Turner

8. Acknowledgments

Special thanks to last call reviewers Elwyn Davies, Barry Lieba, Al Morton, and Peter Saint-Andre.

This work was 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.

9. References

9.1. Normative References

- [QUIC-TLS] Thomson, M., Ed. and S. Turner, Ed., "Using TLS to Secure QUIC", RFC 9001, DOI 10.17487/RFC9001, May 2021, <<https://www.rfc-editor.org/rfc/rfc9001>>.

[QUIC-TRANSPORT]

Iyengar, J., Ed. and M. Thomson, Ed., "QUIC: A UDP-Based Multiplexed and Secure Transport", RFC 9000, DOI 10.17487/RFC9000, May 2021, <<https://www.rfc-editor.org/rfc/rfc9000>>.

9.2. Informative References

[DOTS-ARCH]

Mortensen, A., Ed., Reddy, K., T., Ed., Andreasen, F., Teague, N., and R. Compton, "DDoS Open Threat Signaling (DOTS) Architecture", RFC 8811, DOI 10.17487/RFC8811, August 2020, <<https://www.rfc-editor.org/rfc/rfc8811>>.

[DPLPMTUD] Fairhurst, G., Jones, T., Tüxen, M., Rüngeler, I., and T. Völker, "Packetization Layer Path MTU Discovery for Datagram Transports", RFC 8899, DOI 10.17487/RFC8899, September 2020, <<https://www.rfc-editor.org/rfc/rfc8899>>.

[I-D.ietf-dprive-dnsquic]

Huitema, C., Dickinson, S., and A. Mankin, "DNS over Dedicated QUIC Connections", Work in Progress, Internet-Draft, draft-ietf-dprive-dnsquic-11, 21 March 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-dprive-dnsquic-11>>.

[I-D.ietf-tsvwg-transport-encrypt]

Fairhurst, G. and C. Perkins, "Considerations around Transport Header Confidentiality, Network Operations, and the Evolution of Internet Transport Protocols", Work in Progress, Internet-Draft, draft-ietf-tsvwg-transport-encrypt-21, 20 April 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-tsvwg-transport-encrypt-21>>.

[IPIM]

Allman, M., Beverly, R., and B. Trammell, "In-Protocol Internet Measurement (arXiv preprint 1612.02902)", 9 December 2016, <<https://arxiv.org/abs/1612.02902>>.

[QUIC-APPLICABILITY]

Kuehlewind, M. and B. Trammell, "Applicability of the QUIC Transport Protocol", Work in Progress, Internet-Draft, draft-ietf-quic-applicability-15, 7 March 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-applicability-15>>.

[QUIC-GREASE]

Thomson, M., "Greasing the QUIC Bit", Work in Progress, Internet-Draft, draft-ietf-quic-bit-grease-02, 10 November 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-bit-grease-02>>.

[QUIC-HTTP]

Bishop, M., "Hypertext Transfer Protocol Version 3 (HTTP/3)", Work in Progress, Internet-Draft, draft-ietf-quic-http-34, 2 February 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-http-34>>.

[QUIC-INVARIANTS]

Thomson, M., "Version-Independent Properties of QUIC", RFC 8999, DOI 10.17487/RFC8999, May 2021, <<https://www.rfc-editor.org/rfc/rfc8999>>.

[QUIC-RECOVERY]

Iyengar, J., Ed. and I. Swett, Ed., "QUIC Loss Detection and Congestion Control", RFC 9002, DOI 10.17487/RFC9002, May 2021, <<https://www.rfc-editor.org/rfc/rfc9002>>.

[QUIC-TIMEOUT]

Roskind, J., "QUIC (IETF-88 TSV Area Presentation)", 7 November 2013, <<https://www.ietf.org/proceedings/88/slides/slides-88-tsvarea-10.pdf>>.

[QUIC_LB]

Duke, M., Banks, N., and C. Huitema, "QUIC-LB: Generating Routable QUIC Connection IDs", Work in Progress, Internet-Draft, draft-ietf-quic-load-balancers-13, 28 March 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-load-balancers-13>>.

[RFC1191]

Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, DOI 10.17487/RFC1191, November 1990, <<https://www.rfc-editor.org/rfc/rfc1191>>.

[RFC1812]

Baker, F., Ed., "Requirements for IP Version 4 Routers", RFC 1812, DOI 10.17487/RFC1812, June 1995, <<https://www.rfc-editor.org/rfc/rfc1812>>.

[RFC2475]

Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, DOI 10.17487/RFC2475, December 1998, <<https://www.rfc-editor.org/rfc/rfc2475>>.

- [RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <<https://www.rfc-editor.org/rfc/rfc3168>>.
- [RFC3449] Balakrishnan, H., Padmanabhan, V., Fairhurst, G., and M. Sooriyabandara, "TCP Performance Implications of Network Path Asymmetry", BCP 69, RFC 3449, DOI 10.17487/RFC3449, December 2002, <<https://www.rfc-editor.org/rfc/rfc3449>>.
- [RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <<https://www.rfc-editor.org/rfc/rfc4443>>.
- [RFC4459] Savola, P., "MTU and Fragmentation Issues with In-the-Network Tunneling", RFC 4459, DOI 10.17487/RFC4459, April 2006, <<https://www.rfc-editor.org/rfc/rfc4459>>.
- [RFC4787] Audet, F., Ed. and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, DOI 10.17487/RFC4787, January 2007, <<https://www.rfc-editor.org/rfc/rfc4787>>.
- [RFC4937] Arberg, P. and V. Mammoliti, "IANA Considerations for PPP over Ethernet (PPPoE)", RFC 4937, DOI 10.17487/RFC4937, June 2007, <<https://www.rfc-editor.org/rfc/rfc4937>>.
- [RFC5382] Guha, S., Ed., Biswas, K., Ford, B., Sivakumar, S., and P. Srisuresh, "NAT Behavioral Requirements for TCP", BCP 142, RFC 5382, DOI 10.17487/RFC5382, October 2008, <<https://www.rfc-editor.org/rfc/rfc5382>>.
- [RFC6066] Eastlake 3rd, D., "Transport Layer Security (TLS) Extensions: Extension Definitions", RFC 6066, DOI 10.17487/RFC6066, January 2011, <<https://www.rfc-editor.org/rfc/rfc6066>>.
- [RFC7301] Friedl, S., Popov, A., Langley, A., and E. Stephan, "Transport Layer Security (TLS) Application-Layer Protocol Negotiation Extension", RFC 7301, DOI 10.17487/RFC7301, July 2014, <<https://www.rfc-editor.org/rfc/rfc7301>>.
- [RFC7605] Touch, J., "Recommendations on Using Assigned Transport Port Numbers", BCP 165, RFC 7605, DOI 10.17487/RFC7605, August 2015, <<https://www.rfc-editor.org/rfc/rfc7605>>.

- [RFC7801] Dolmatov, V., Ed., "GOST R 34.12-2015: Block Cipher "Kuznyechik"", RFC 7801, DOI 10.17487/RFC7801, March 2016, <<https://www.rfc-editor.org/rfc/rfc7801>>.
- [RFC7838] Nottingham, M., McManus, P., and J. Reschke, "HTTP Alternative Services", RFC 7838, DOI 10.17487/RFC7838, April 2016, <<https://www.rfc-editor.org/rfc/rfc7838>>.
- [RFC7983] Petit-Huguenin, M. and G. Salgueiro, "Multiplexing Scheme Updates for Secure Real-time Transport Protocol (SRTP) Extension for Datagram Transport Layer Security (DTLS)", RFC 7983, DOI 10.17487/RFC7983, September 2016, <<https://www.rfc-editor.org/rfc/rfc7983>>.
- [RFC8201] McCann, J., Deering, S., Mogul, J., and R. Hinden, Ed., "Path MTU Discovery for IP version 6", STD 87, RFC 8201, DOI 10.17487/RFC8201, July 2017, <<https://www.rfc-editor.org/rfc/rfc8201>>.
- [RFC8504] Chown, T., Loughney, J., and T. Winters, "IPv6 Node Requirements", BCP 220, RFC 8504, DOI 10.17487/RFC8504, January 2019, <<https://www.rfc-editor.org/rfc/rfc8504>>.
- [RFC9065] Fairhurst, G. and C. Perkins, "Considerations around Transport Header Confidentiality, Network Operations, and the Evolution of Internet Transport Protocols", RFC 9065, DOI 10.17487/RFC9065, July 2021, <<https://www.rfc-editor.org/rfc/rfc9065>>.
- [TLS-ECH] Rescorla, E., Oku, K., Sullivan, N., and C. A. Wood, "TLS Encrypted Client Hello", Work in Progress, Internet-Draft, draft-ietf-tls-esni-14, 13 February 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-tls-esni-14>>.
- [TMA-QOF] Trammell, B., Gugelmann, D., and N. Brownlee, "Inline Data Integrity Signals for Passive Measurement (in Proc. TMA 2014)", April 2014.
- [WIRE-IMAGE]
Trammell, B. and M. Kuehlewind, "The Wire Image of a Network Protocol", RFC 8546, DOI 10.17487/RFC8546, April 2019, <<https://www.rfc-editor.org/rfc/rfc8546>>.

Authors' Addresses

Mirja Kuehlewind
Ericsson

Email: mirja.kuehlewind@ericsson.com

Brian Trammell
Google Switzerland GmbH
Gustav-Gull-Platz 1
CH- 8004 Zurich
Switzerland
Email: ietf@trammell.ch

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: 26 April 2022

N. Kuhn
CNES
E. Stephan
Orange
G. Fairhurst
T. Jones
University of Aberdeen
C. Huitema
Private Octopus Inc.
23 October 2021

Transport parameters for 0-RTT connections
draft-kuhn-quic-0rtt-bdp-11

Abstract

QUIC 0-RTT transport features currently focuses on egress traffic optimization. This draft describes a QUIC extension that can be used to improve the performance of ingress traffic.

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 26 April 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components

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

Table of Contents

1. Introduction	2
1.1. Notations and terms	4
1.2. Requirements Language	5
2. Safe jump start	5
2.1. Rationale behind the safety guidelines	5
2.2. Rationale #1: Variable network conditions	6
2.3. Rationale #2: Malicious clients	7
2.4. Trade-off between the different solutions	8
2.4.1. Security aspects	8
2.4.2. Interoperability and use-cases	8
2.4.3. Summary	9
3. Safety guidelines	10
4. Implementation considerations	12
4.1. Rationale behind the different implementation options	12
4.2. Independent local storage of values	12
4.3. Using NEW_TOKEN frames	13
4.4. BDP Frame	13
4.4.1. BDP Frame Format	13
4.4.2. Extension activation	14
5. Discussion	15
5.1. BDP extension protected as much as initial_max_data	15
5.2. Other use-cases	15
5.2.1. Optimizing client's requests	15
5.2.2. Sharing transport information across multiple connections	16
6. Acknowledgments	16
7. IANA Considerations	16
8. Security Considerations	16
9. References	16
9.1. Normative References	16
9.2. Informative References	17
Authors' Addresses	17

1. Introduction

QUIC 0-RTT transport features currently focus on egress traffic optimization. This draft describes a QUIC extension that can be used to improve the performance of ingress traffic.

[RFC9000] mentions that "Generally, implementations are advised to be cautious when using previous values on a new path." This draft proposes a discussion on how using previous values can be achieved in a interoperable manner and how it can be done safely.

When clients resume a session to download a large object, the congestion control algorithms will require time to ramp-up the packet rate as a sequence of Round-Trip Time (RTT)-based increases. This document specifies a method that can improve traffic delivery by allowing a QUIC connection to avoid a the slow process to discover key path parameters including a way to more rapidly grow the congestion window (cwnd):

1. During a previous session, current RTT (`current_rtt`), bottleneck bandwidth (`current_bb`) and current client IP (`current_client_ip`) are stored as `saved_rtt`, `saved_bb` and `saved_client_ip`;
2. When resuming a session to the same IP address, the server can then utilize the `current_rtt` and the `current_bb` to the `saved_rtt` and `saved_bb` of a previous connection.

This method applies to any resumed QUIC session: both `saved_session` and `recon_session` can be a 0-RTT QUIC connection or a 1-RTT QUIC connection.

The current version of this draft considers several possible solutions: (1) the saved parameters are stored at the server; they are not sent to the client; (2) the saved parameters are sent to the client as an encrypted opaque blob; although the client is unable to read the parameters can include this opaque blob in a subsequent request to the server; (3) the saved parameters are sent to the client and the client is notified of their value, but the parameters also include a cryptographic integrity check; the client can include both the parameters and the integrity check in a subsequent request to the server.

None of these possible solutions allow q client to modify the parameters that will be used by the server.

There are several cases where the parameters of a previous session are not appropriate. These include:

- (1) the network conditions have changed and the current capacity is less than the previously estimated bottleneck bandwidth. Using the saved congestion control state would increase congestion;

(2) the network path has changed and the new path is different. Using the saved congestion control state could increase congestion. This case might be accompanied by a change in the RTT or IP address.

(3) a client uses parameters that are no longer appropriate, e.g., to intentionally try to use a CWND larger than appropriate.

This document:

1. proposes guidelines for how to safely apply the previously computed parameters to new sessions;
2. describes different implementation considerations for the proposed method using QUIC;
3. discusses the trade-offs associated with the different implementation solutions.

1.1. Notations and terms

- * IW: Initial Window (e.g., from [RFC6928]);
- * current_iw: Current Initial Window
- * recom_iw: Recommended Initial Window
- * BDP: defined below
- * CWND: the congestion window used by server (maximum number of bytes allowed in flight by the CC)
- * current_bb : Current estimated bottleneck bandwidth
- * saved_bb: Estimated bottleneck bandwidth preserved from a previous connection
- * RTT: Round-Trip Time
- * current_rtt: Current RTT
- * saved_rtt: RTT preserved from a previous connection
- * client_ip : IP address of the client
- * current_client_ip : Current IP address of the client

- * `saved_client_ip` : IP address of the client preserved from a previous connection
- * remembered BDP parameters: a combination of `saved_rtt` and `saved_bb`
- * ITT : Interpacket Transmission Time
- * MSS : Maximum Message Size
- * AEAD : Authenticated Encryption with Associated Data
- * LRU : Least Recently Used

[RFC6349] defines the BDP as follows: "Derived from Round-Trip Time (RTT) and network Bottleneck Bandwidth (BB), the Bandwidth-Delay Product (BDP) determines the Send and Received Socket buffer sizes required to achieve the maximum TCP Throughput." This draft considers the BDP estimated by a server that includes all buffering along the network path. In that sense, the BDP estimated is related to the amount of bytes in flight.

A QUIC connection might not reproduce the procedure detailed in [RFC6349] to measure the BDP. A server might be able to exploit an internal evaluation of the Bottleneck Bandwidth to estimate the BDP.

This document refers to the `saved_bb` and `current_bb` for the previously estimated bottleneck bandwidth. This value can be easily estimated when using a rate-based congestion controller, such as BBR. Other congestion controllers, such as CUBIC or RENO, could estimate the bottleneck bandwidth by utilizing a combination of the `cwnd` and the minimum RTT. This approach could result in over estimating the bottleneck bandwidth and ought to be used with caution.

1.2. Requirements Language

The keywords "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "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.

2. Safe jump start

2.1. Rationale behind the safety guidelines

The previously measured `saved_rtt` and `saved_bb` SHOULD NOT be used as-is, to avoid potential congestion collapse:

- * Rationale #1: Internet path capacity can change at any time. An Internet method needs to be robust to network conditions that can differ from one session to the next.
- * Rationale #2: Information sent by a malicious client is not relevant. A client could try to convince a server to use a CWND higher than appropriate, to gain an unfair share of capacity for itself or to induce congestion for other flows.

2.2. Rationale #1: Variable network conditions

The server MUST check the validity of the `saved_rtt` and `saved_bb` parameters, whether these are sent by a client or are stored at the server. The following events indicates cases where use of these parameters is inappropriate:

- * IP address changed: If the client changes its IP address (i.e. the `saved_client_ip` is different from the `current_client_ip`), the different address is to be taken as an indication of a different network path. This new path does not necessarily exhibit the same characteristics as the old one. If the server changes its IP address after a migration, it would not be safe to exploit previously estimated parameters.
- * RTT changed: A significant change in RTT might be an indication that the network conditions changed. Since the CC information is directly impacted by the RTT, a significant change in RTT is a strong indication that the previously estimated BDP parameters are likely to not be valid for the current path.
- * Lifetime of the extension: The CC information is temporal. Frequent connections to the same IP address are likely to track changes, but long-term use of previous values are not appropriate.
- * BB over estimation: There are cases where using the `cwnd` would infringe the bottleneck bandwidth. However, at the end of a CC slow start, the value of `cwnd` can be significantly larger than the value, that the CC finally converges to (after a few more rounds). Directly exploiting such value for the bottleneck bandwidth estimation may be inappropriate. One mitigation could be to restrict to only a fraction (e.g., 1/2) of the previously used `cwnd`; another mitigation might be to calculate the bottleneck bandwidth based on the flight size.

There are different solutions for the variable network conditions:

- * Rationale #1 - Solution #1 : When resuming a session, restore the current_bb and current_rtt from the saved_bb and saved_rtt parameters estimated from a previous connection.
- * Rationale #1 - Solution #2 : When resuming a session, implement a safety check to measure avoid using the saved_bb and saved_rtt parameters to cause congestion over the path. In this case, the current_bb and current_rtt might not be set directly to the saved_bb and saved_rtt: the server might wait for the completion of the safety check before doing so.

Section 3 describes various approaches for Rationale #1 - Solution #2.

2.3. Rationale #2: Malicious clients

The server MUST check the integrity of the saved_rtt and saved_bb parameters received from a client.

There are several solutions to avoid attacks by malicious clients:

- * Rationale #2 - Solution #1 : The server stores a local estimate of the bottleneck bandwidth and RTT parameters as the saved_bb and saved_rtt.
- * Rationale #2 - Solution #2 : The server sends the estimate of the bottleneck bandwidth and RTT parameters to the client as the saved_bb and saved_rtt. This information is encrypted by the server. The client resends the same encrypted information when resuming a connection. The client can neither read nor modify the saved_rtt and saved_bb parameters.
- * Rationale #2 - Solution #3 : The server sends an estimate of the saved_rtt and saved_bb parameters to the client. The information includes an integrity protection check. The client can resend the information when resuming a connection. This allows a client to read, but not modify, the saved_rtt and saved_bb parameters. This might enable a client to decide whether the new parameters are appropriate, based on client-side information about the network conditions or connectivity.

Section 4 describes various implementation approaches for each of these solutions using local storage (Section 4.2 for Rationale #2 - Solution #1), NEW_TOKEN Frame (Section 4.3 for Rationale #2 - Solution #2), BDP extension Frame (Section 4.4 for Rationale #2 - Solution #3).

2.4. Trade-off between the different solutions

This section provides a description of different implementation options and discusses their respective advantages and drawbacks. While there are some discussions for the solutions regarding Rationale #2, the server **MUST** consider Rationale #1 - Solution #2 and avoid Rationale #1 - Solution #1: the server **MUST** implement a safety check to measure whether the saved BDP parameters (i.e. `saved_rtt` and `saved_bb`) are relevant or check that their usage would not cause excessive congestion over the path.

2.4.1. Security aspects

The client can send information related to the `saved_rtt` and `saved_bb` to the server with the BDP Frame extension using either Rationale #2 - Solution #2 or Rationale #2 - Solution #3. However, the server **SHOULD NOT** trust the client. Indeed, even if 0-RTT packets containing the BDP Frame are encrypted, a client could modify the values within the extension and encrypt the 0-RTT packet. Authentication mechanisms might not guarantee that the values are safe. It is not an easy operation for a client to modify authenticated or encrypted data without this being detected by a server. Modification could be realized by malicious clients. One way to avoid this is for a server to also store the `saved_rtt` and `saved_bb` parameters.

A malicious client might modify the `saved_bb` parameter to convince the server to use a larger CWND than appropriate. Using the algorithms proposed in Section 3, the server may reduce any intended harm and can check that part of the information provided by the client are valid.

Storing the BDP parameters locally at the server reduces the associated risks by allowing the client to transmit information related to the BDP of the path in the case of a malicious client trying to break the encryption mechanism that it had received.

2.4.2. Interoperability and use-cases

If the server stores a resumption ticket for each client to protect against replay on a third party IP, it could also store the IP address (i.e. `saved_client_ip`) and BDP parameters (i.e. `saved_rtt` and `saved_bb`) of the previous session of the client.

In cases where the BDP Frame extension is exploited, the approach of storing the BDP parameters locally at the server can provide a cross-check of the BDP parameters sent by a client. The server can anyway enable a safe jumpstart, but without the BDP Frame extension.

However, the client does not have the choice of accepting to use this or not, and is unable to utilize local knowledge of the network conditions or connectivity.

Storing local values related to the BDP would help in improving the ingress for 0-RTT connections, however, not using a BDP Frame extension could reduce the interest of the approach where (1) the client knows the BDP estimations done at the server, (2) the client decides to accept or reject ingress optimization, (3) the client tunes application level requests.

2.4.3. Summary

As a summary, the approach of local storage of values can be secure and the BDP Frame extension provides more information to the client and more interoperability. The Figure 1 provides a summary of the advantages and drawbacks of each approach.

Rationale	Solution	Advantage	Drawback	Comment
#1 Variable Network	#1 set current_* to saved_*	Ingress optim.	Risks of adding congestion	MUST NOT implement
	#2 Implement safety check	Reduce risks of adding congestion	Negative impact on ingress optim.	MUST implement Section 3
#2 Malicious client	#1 Local storage	Enforced security	Client unable to decide to reject Malicious server could fill client's buffer Limited use-cases	Section 4.2
	#2 NEW_TOKEN	Save resource at server Opaque token protected	Malicious client could change token even if protected	

			Malicious server could fill client's buffer Server may not trust client	Section 4.3
	#3 BDP extension	Extended use-cases Save resource at server Client can read and decide to reject BDP extension protected	Malicious client could change BDP even if protected Server may not trust client	Section 4.4

Figure 1: Comparing solutions

3. Safety guidelines

The safety guidelines are designed to avoid a server adding excessive congestion to an already congested path. The following mechanisms help in fulfilling this objective:

- * The server SHOULD compare the measured transport parameters (in particular `current_rtt`) of the 0-RTT connection with those of the 1-RTT connection (in particular `saved_rtt`);
- * The server SHOULD NOT consider the `saved_bb` parameter when there is any indicated congestion (e.g., loss of packet during the first transmission of data or ECN-CE mark);
- * The server MUST NOT send more than the recommended maximum IW (`recom_iw`) in the first transmission of data. This value could be based on a local understanding of the path characteristics. Knowing the congestion status of the network in closed environments may help in increasing the recommended maximum IW.
- * The server SHOULD NOT store and/or send information related to the previously estimated bottleneck bandwidth (`saved_bb`) (see Section 1.1 for more details on bottleneck bandwidth definition), if this estimation has not been computed after some rounds during the 1-RTT connection. At least, the 1-RTT connection should have reached the congestion avoidance phase.

The proposed mechanisms SHOULD be limited by any rate-limitation mechanisms of QUIC, such as flow control mechanisms or amplification attack prevention. In particular, it may be necessary to issue proactive MAX_DATA frames to increase the flow control limits of a connection. In particular, the maximum number of packets that can be sent without acknowledgment needs to be chosen to avoid the creation and the increase of congestion for the path.

This extension should not provide an opportunity for the current connection to be a vector of an amplification attack. The address validation process, used to prevent amplification attacks, SHOULD be performed [RFC9000].

The following mechanisms could be implemented:

* Exploit a standard IW:

1. The server sends the first data packet using the IW - this is a safe starting point for any path where there is no path information or where there is no congestion state. This avoids adding excessive congestion to the path;
2. If the reception of IW exhibits characteristics that resemble those of a recent previous session from the client (i.e. $\text{current_rtt} < 1.2 * \text{saved_rtt}$ and all data was acknowledged without reported congestion), the method permits the sender to consider the saved_bb as an input to adapt current_bb to rapidly determine a new safe rate;
3. The sender needs to avoid a burst of packets resulting from a step-increase in the congestion window [RFC9000]. Pacing the packets as a function of the current_rtt can provide this additional safety during the period in which the CWND is increased by the method.

* Identify a relevant pacing rhythm:

- The server estimates the pacing rhythm using saved_rtt and saved_bb. The Interpacket Transmission Time (ITT) is determined by the ratio between the current Maximum Message Size (MSS) for packets and the ratio between the saved_bb and saved_rtt. A tunable safety margin might be introduced to avoid sending more than a recommended maximum IW (recom_iw):
 - o $\text{current_iw} = \min(\text{recom_iw}, \text{saved_bb})$
 - o $\text{ITT} = \text{MSS} / (\text{current_iw} / \text{saved_rtt})$

- When the IW is acknowledged, the server falls back to a standard slow-start mechanism.
- * Tune slow-start mechanisms: After transport parameters are set to a previously estimated bottleneck bandwidth, if slow-start mechanisms continue, the sender can overshoot the bottleneck capacity. This can occur even if the safety check described in this section is implemented.
- For NewReno and CUBIC, it is recommended to exit slow-start and enter in congestion avoidance phase.
- For BBR, it is recommended to move to the "probe bandwidth" state.

This follows the idea of [RFC4782],
[I-D.irtf-iccr-g-sallantin-initial-spreading] and [CONEXT15].

4. Implementation considerations

4.1. Rationale behind the different implementation options

The NewSessionTickets messages of TLS offer a solution. The idea would have been to add a 'bdp_metada' field in the NewSessionTickets that the client could read. The sole extension currently defined in TLS1.3 that can be seen by the client is max_early_data_size (see section 4.6.1 of [RFC8446]). However, in the general design of QUIC, TLS sessions are managed by the TLS stacks.

Three distinct approaches are presented: sending an opaque blob to the client that it may return to the server for a future connection (see Section 4.3), enable a local storage of BDP related values (see Section 4.2) and a BDP Frame extension (see Section 4.4).

4.2. Independent local storage of values

This approach independently lets both a client and a server remember their BDP parameters:

- * During a 1-RTT session, the endpoint stores the RTT (as the saved_rtt) and bottleneck bandwidth (as the saved_bb) together with the session resume ticket. The client can also store the IP address of the server.
- * The server maintains a table of previously issued tickets, indexed by the random ticket identifier that is used to guarantee uniqueness of the Authenticated Encryption with Associated Data (AEAD) encryption. Old tokens are removed from the table using

the Least Recently Used (LRU) logic. For each ticket identifier, the table holds the RTT and bottleneck bandwidth (i.e. `saved_rtt` and `saved_bb`), and also the IP address of the client (i.e. `saved_client_ip`).

During the 0-RTT session, the endpoint waits for the first RTT measurement from the peer's IP address. This is used to verify that the `current_rtt` has not significantly changed from the `saved_rtt`, and hence is an indication that the BDP information is appropriate to the path that is currently being used.

If this RTT is confirmed (e.g. `current_rtt < 1.2*saved_rtt`, the endpoint also verifies that an initial window of data has been acknowledged without requiring retransmission. This second check detects a path with significant incipient congestion (i.e. where it would not be safe to update the CWND based on the `saved_bb`). In practice, this could be realized by a proportional increase in the CWND, where the increase is $(\text{saved_bb}/\text{IW}) * \text{proportion_of_IW_currently_ACKed}$.

This solution does not allow the client to refuse the exploitation of the BDP parameters. If the server does not want to store the metrics from previous connections, an equivalent of the `tcp_no_metrics_save` for QUIC may be necessary. This option could be negotiated that allows a client to choose whether to use the saved information.

4.3. Using NEW_TOKEN frames

Using NEW_TOKEN Frames, the server could send a token to the client through a NEW_TOKEN Frame. The token is an opaque blob and the client can not read its content (see section 19.7 of [RFC9000]). The client sends the received token in the header of an Initial packet for a later connection.

4.4. BDP Frame

This section describes the use of a new Frame, the BDP Frame. The BDP Frame MUST be contained in 0-RTT packets, if sent by the client. The BDP Frame MUST be contained in 1-RTT packets, if sent by the server. The BDP Frame MUST be considered by congestion control and its data is not be limited by flow control limits. The server MAY send multiple BDP Frames in both 1-RTT and 0-RTT connections. The client can send BDP Frames during 1-RTT and 0-RTT connections.

4.4.1. BDP Frame Format

A BDP Frame is formatted as shown in Figure 2.

```
BDP Frame {  
    Type (i) = 0xXXX,  
    Lifetime (i),  
    Saved BB (i),  
    Saved RTT (i),  
    Saved IP length (i),  
    Saved IP (...)  
}
```

Figure 2: BDP Frame Format

A BDP Frame contains the following fields:

- * Lifetime (extension_lifetime): The extension_lifetime is a value in milliseconds, encoded as a variable length integer. This follows the idea of NewSessionTicket of TLS [RFC8446]. This represents the validity in time of this extension.
- * Saved BB (saved_bb): The saved_bb is a value in bytes, encoded as a variable length integer. The bottleneck bandwidth estimated for the previous connection by the server. Using the previous values of bytes_in_flight defined in [RFC9002] can result in overshoot of the bottleneck capacity and is not advised.
- * Saved RTT (saved_rtt): The saved_rtt is a value in milliseconds, encoded as a variable length integer. This could be set to the minimum RTT (min_rtt). The saved_rtt can be set to min_rtt. NOTE: The min_rtt defined in [RFC9002], does not track a decreasing RTT: therefore min_rtt reported might be larger than the actual minimum RTT measured during the 1-RTT connection.
- * Saved IP length (saved_ip_length) : The length of the IP address set to either 4 (IPv4) or 16 (IPv6).
- * Saved IP (saved_client_ip) : The saved_client_ip could be set to the IP address of the client.

4.4.2. Extension activation

The client can accept the transmission of BDP Frames from the server by using the enable_bdp transport extension.

enable_bdp (0xTBD): in the 1-RTT connection, the client indicates to the server that it wishes to receive BDP extension Frames for improving ingress of 0-RTT connection. The default value is 0. Values strictly above 3 are invalid, and receipt of these values MUST be treated as a connection error of type TRANSPORT_PARAMETER_ERROR.

- * 0: Default value. If the client does not send this parameter, the server considers that the client does not support or does not wish to activate the BDP extension.
- * 1: The client indicates to the server that it wishes to receive BDP Frame and activates the ingress optimization for the 0-RTT connection.
- * 2: The client indicates that it does not wish to receive BDP Frames but activates ingress optimization.
- * 3: The client indicates that it wishes to receive BDP Frames but does not activate ingress optimization.

This Transport Parameter is encoded as per Section 18 of [RFC9000].

5. Discussion

5.1. BDP extension protected as much as initial_max_data

The BDP metadata parameters are measured by the server during a previous connection. The BDP extension is protected by the mechanism that protects the exchange of the 0-RTT transport parameters. For version 1 of QUIC, the BDP extension is protected using the mechanism that already protects the "initial_max_data" parameter. This is defined in sections 4.5 to 4.7 of [RFC9001]. This provides a way for the server to verify that the parameters proposed by the client are the same as those that the server sent to the client during the previous connection.

5.2. Other use-cases

5.2.1. Optimizing client's requests

When using Dynamic Adaptive Streaming over HTTPS (DASH), clients might encounter issues in knowing the available path capacity or DASH can encounter issues in reaching the best available video playback quality. The client requests could then be adapted and specific traffic could utilize information from the path characteristics (such as encouraging the client to increase the quality of video chunks, to fill the buffers and avoid video blocking or to send high quality adds).

In other cases, applications could provide additional services if clients can know the server estimation of the path characteristics.

5.2.2. Sharing transport information across multiple connections

There can be benefit in sharing transport information across multiple connections. [I-D.ietf-tcpm-2140bis] considers the sharing of transport parameters between TCP connections originating from the same host. The proposal in this document has the advantage of storing server-generated information at the client and not requiring the server to retain additional state for each client.

6. Acknowledgments

The authors would like to thank Gabriel Montenegro, Patrick McManus, Ian Swett, Igor Lubashev, Robin Marx, Roland Bless and Franklin Simo for their fruitful comments on earlier versions of this document.

7. IANA Considerations

TBD: Text is required to register the BDP Frame and the enable_bdp transport parameter. Parameters are registered using the procedure defined in [RFC9000].

8. Security Considerations

Security considerations are discussed in Section 5 and in Section 3.

9. References

9.1. Normative References

- [I-D.ietf-tcpm-2140bis]
Touch, J., Welzl, M., and S. Islam, "TCP Control Block Interdependence", Work in Progress, Internet-Draft, draft-ietf-tcpm-2140bis-11, 12 April 2021, <<https://www.ietf.org/archive/id/draft-ietf-tcpm-2140bis-11.txt>>.
- [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>>.
- [RFC4782] Floyd, S., Allman, M., Jain, A., and P. Sarolahti, "Quick-Start for TCP and IP", RFC 4782, DOI 10.17487/RFC4782, January 2007, <<https://www.rfc-editor.org/info/rfc4782>>.

- [RFC6349] Constantine, B., Forget, G., Geib, R., and R. Schrage, "Framework for TCP Throughput Testing", RFC 6349, DOI 10.17487/RFC6349, August 2011, <<https://www.rfc-editor.org/info/rfc6349>>.
- [RFC6928] Chu, J., Dukkupati, N., Cheng, Y., and M. Mathis, "Increasing TCP's Initial Window", RFC 6928, DOI 10.17487/RFC6928, April 2013, <<https://www.rfc-editor.org/info/rfc6928>>.
- [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>>.
- [RFC8446] Rescorla, E., "The Transport Layer Security (TLS) Protocol Version 1.3", RFC 8446, DOI 10.17487/RFC8446, August 2018, <<https://www.rfc-editor.org/info/rfc8446>>.
- [RFC9000] Iyengar, J., Ed. and M. Thomson, Ed., "QUIC: A UDP-Based Multiplexed and Secure Transport", RFC 9000, DOI 10.17487/RFC9000, May 2021, <<https://www.rfc-editor.org/info/rfc9000>>.
- [RFC9001] Thomson, M., Ed. and S. Turner, Ed., "Using TLS to Secure QUIC", RFC 9001, DOI 10.17487/RFC9001, May 2021, <<https://www.rfc-editor.org/info/rfc9001>>.
- [RFC9002] Iyengar, J., Ed. and I. Swett, Ed., "QUIC Loss Detection and Congestion Control", RFC 9002, DOI 10.17487/RFC9002, May 2021, <<https://www.rfc-editor.org/info/rfc9002>>.

9.2. Informative References

- [CONEXT15] Li, Q., Dong, M., and P B. Godfrey, "Halfback: Running Short Flows Quickly and Safely", ACM CoNEXT , 2015.
- [I-D.irtf-iccr-g-sallantin-initial-spreading] Sallantin, R., Baudoin, C., Arnal, F., Dubois, E., Chaput, E., and A. Beylot, "Safe increase of the TCP's Initial Window Using Initial Spreading", Work in Progress, Internet-Draft, draft-irtf-iccr-g-sallantin-initial-spreading-00, 15 January 2014, <<https://www.ietf.org/archive/id/draft-irtf-iccr-g-sallantin-initial-spreading-00.txt>>.

Authors' Addresses

Nicolas Kuhn
CNES

Email: nicolas.kuhn.ietf@gmail.com

Emile Stephan
Orange

Email: emile.stephan@orange.com

Godred Fairhurst
University of Aberdeen
Department of Engineering
Fraser Noble Building
Aberdeen

Email: gorry@erg.abdn.ac.uk

Tom Jones
University of Aberdeen
Department of Engineering
Fraser Noble Building
Aberdeen

Email: tom@erg.abdn.ac.uk

Christian Huitema
Private Octopus Inc.

Email: huitema@huitema.net

QUIC
Internet-Draft
Intended status: Standards Track
Expires: 6 May 2021

R. Marx
Hasselt University
2 November 2020

QUIC and HTTP/3 event definitions for qlog
draft-marx-qlog-event-definitions-quic-h3-02

Abstract

This document describes concrete qlog event definitions and their metadata for QUIC and HTTP/3-related events. These events can then be embedded in the higher level schema defined in [QLOG-MAIN].

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 6 May 2021.

Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	5
1.1. Notational Conventions	5
2. Overview	5
2.1. Importance	6
2.2. Custom fields	7
3. Events not belonging to a single connection	7
4. QUIC and HTTP/3 fields	8
4.1. Raw packet and frame information	8
5. QUIC event definitions	10
5.1. connectivity	10
5.1.1. server_listening	10
5.1.2. connection_started	10
5.1.3. connection_closed	11
5.1.4. connection_id_updated	12
5.1.5. spin_bit_updated	12
5.1.6. connection_retried	12
5.1.7. connection_state_updated	13
5.1.8. MIGRATION-related events	15
5.2. security	15
5.2.1. key_updated	15
5.2.2. key_retired	15
5.3. transport	16
5.3.1. version_information	16
5.3.2. alpn_information	17
5.3.3. parameters_set	18
5.3.4. parameters_restored	20
5.3.5. packet_sent	20
5.3.6. packet_received	21
5.3.7. packet_dropped	22
5.3.8. packet_buffered	23
5.3.9. packets_acked	24
5.3.10. datagrams_sent	24
5.3.11. datagrams_received	25
5.3.12. datagram_dropped	25
5.3.13. stream_state_updated	26
5.3.14. frames_processed	27
5.3.15. data_moved	28
5.4. recovery	30
5.4.1. parameters_set	30
5.4.2. metrics_updated	30
5.4.3. congestion_state_updated	31
5.4.4. loss_timer_updated	32
5.4.5. packet_lost	33
5.4.6. marked_for_retransmit	34
6. HTTP/3 event definitions	34
6.1. http	34

6.1.1.	parameters_set	34
6.1.2.	parameters_restored	35
6.1.3.	stream_type_set	36
6.1.4.	frame_created	36
6.1.5.	frame_parsed	37
6.1.6.	push_resolved	37
6.2.	qpack	38
6.2.1.	state_updated	38
6.2.2.	stream_state_updated	39
6.2.3.	dynamic_table_updated	39
6.2.4.	headers_encoded	39
6.2.5.	headers_decoded	40
6.2.6.	instruction_created	40
6.2.7.	instruction_parsed	41
7.	Generic events and Simulation indicators	41
7.1.	generic	41
7.1.1.	error	42
7.1.2.	warning	42
7.1.3.	info	42
7.1.4.	debug	42
7.1.5.	verbose	43
7.2.	simulation	43
7.2.1.	scenario	43
7.2.2.	marker	44
8.	Security Considerations	44
9.	IANA Considerations	44
10.	References	44
10.1.	Normative References	44
10.2.	Informative References	45
Appendix A.	QUIC data field definitions	45
A.1.	IPAddress	45
A.2.	PacketType	45
A.3.	PacketNumberSpace	45
A.4.	PacketHeader	45
A.5.	Token	46
A.6.	KeyType	46
A.7.	QUIC Frames	47
A.7.1.	PaddingFrame	47
A.7.2.	PingFrame	47
A.7.3.	AckFrame	47
A.7.4.	ResetStreamFrame	48
A.7.5.	StopSendingFrame	48
A.7.6.	CryptoFrame	49
A.7.7.	NewTokenFrame	49
A.7.8.	StreamFrame	49
A.7.9.	MaxDataFrame	50
A.7.10.	MaxStreamDataFrame	50
A.7.11.	MaxStreamsFrame	50

A.7.12.	DataBlockedFrame	50
A.7.13.	StreamDataBlockedFrame	50
A.7.14.	StreamsBlockedFrame	50
A.7.15.	NewConnectionIDFrame	51
A.7.16.	RetireConnectionIDFrame	51
A.7.17.	PathChallengeFrame	51
A.7.18.	PathResponseFrame	51
A.7.19.	ConnectionCloseFrame	52
A.7.20.	HandshakeDoneFrame	52
A.7.21.	UnknownFrame	52
A.7.22.	TransportError	52
A.7.23.	CryptoError	53
Appendix B.	HTTP/3 data field definitions	53
B.1.	HTTP/3 Frames	53
B.1.1.	DataFrame	53
B.1.2.	HeadersFrame	54
B.1.3.	CancelPushFrame	54
B.1.4.	SettingsFrame	54
B.1.5.	PushPromiseFrame	54
B.1.6.	GoAwayFrame	55
B.1.7.	MaxPushIDFrame	55
B.1.8.	DuplicatePushFrame	55
B.1.9.	ReservedFrame	55
B.1.10.	UnknownFrame	55
B.2.	ApplicationError	55
Appendix C.	QPACK DATA type definitions	56
C.1.	QPACK Instructions	56
C.1.1.	SetDynamicTableCapacityInstruction	56
C.1.2.	InsertWithNameReferenceInstruction	56
C.1.3.	InsertWithoutNameReferenceInstruction	57
C.1.4.	DuplicateInstruction	57
C.1.5.	HeaderAcknowledgementInstruction	57
C.1.6.	StreamCancellationInstruction	57
C.1.7.	InsertCountIncrementInstruction	58
C.2.	QPACK Header compression	58
C.2.1.	IndexedHeaderField	58
C.2.2.	LiteralHeaderFieldWithName	58
C.2.3.	LiteralHeaderFieldWithoutName	59
C.2.4.	QPackHeaderBlockPrefix	59
Appendix D.	Change Log	59
D.1.	Since draft-01:	59
D.2.	Since draft-00:	61
Appendix E.	Design Variations	61
Appendix F.	Acknowledgements	61
Author's Address	61

1. Introduction

This document describes the values of the qlog name ("category" + "event") and "data" fields and their semantics for the QUIC and HTTP/3 protocols. This document is based on draft-29 of the QUIC and HTTP/3 I-Ds QUIC-TRANSPORT [QUIC-HTTP] and draft-16 of the QPACK I-D [QUIC-QPACK].

Feedback and discussion welcome at <https://github.com/quiclog/internet-drafts> (<https://github.com/quiclog/internet-drafts>). Readers are advised to refer to the "editor's draft" at that URL for an up-to-date version of this document.

Concrete examples of integrations of this schema in various programming languages can be found at <https://github.com/quiclog/qlog/> (<https://github.com/quiclog/qlog/>).

1.1. Notational 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].

The examples and data definitions in this document are expressed in a custom data definition language, inspired by JSON and TypeScript, and described in [QLOG-MAIN].

2. Overview

This document describes the values of the qlog "name" ("category" + "event") and "data" fields and their semantics for the QUIC and HTTP/3 protocols.

This document assumes the usage of the encompassing main qlog schema defined in [QLOG-MAIN]. Each subsection below defines a separate category (for example connectivity, transport, http) and each subsubsection is an event type (for example "packet_received").

For each event type, its importance and data definition is laid out, often accompanied by possible values for the optional "trigger" field. For the definition and semantics of "trigger", see the main schema document.

Most of the complex datastructures, enums and re-usable definitions are grouped together on the bottom of this document for clarity.

2.1. Importance

Many of the events defined in this document map directly to concepts seen in the QUIC and HTTP/3 documents, while others act as aggregating events that combine data from several possible protocol behaviours or code paths into one. This is done to reduce the amount of unique event definitions, as reflecting each possible protocol event as a separate qlog entity would cause an explosion of event types. Similarly, we prevent logging duplicate packet data as much as possible. As such, especially packet header value updates are split out into separate events (for example `spin_bit_updated`, `connection_id_updated`), as they are expected to change sparingly.

Consequently, many events that can be directly inferred from data on the wire (for example flow control limit changes) if the implementation is bug-free, are currently not explicitly defined as stand-alone events. Exceptions can be made for common events that benefit from being easily identifiable or individually logged (for example the `"packets_acked"` event). This can in turn give rise to separate events logging similar data, where it is not always clear which event should be logged (for example the separate `"connection_started"` event, whereas the more general `"connection_state_updated"` event also allows indicating that a connection was started).

To aid in this decision making, each event has an "importance indicator" with one of three values, in decreasing order of importance and expected usage:

- * Core
- * Base
- * Extra

The "Core" events are the events that SHOULD be present in all qlog files. These are mostly tied to basic packet and frame parsing and creation, as well as listing basic internal metrics. Tool implementers SHOULD expect and add support for these events, though SHOULD NOT expect all Core events to be present in each qlog trace.

The "Base" events add additional debugging options and CAN be present in qlog files. Most of these can be implicitly inferred from data in Core events (if those contain all their properties), but for many it is better to log the events explicitly as well, making it clearer how the implementation behaves. These events are for example tied to passing data around in buffers, to how internal state machines change and help show when decisions are actually made based on received data. Tool implementers SHOULD at least add support for showing the contents of these events, if they do not handle them explicitly.

The "Extra" events are considered mostly useful for low-level debugging of the implementation, rather than the protocol. They allow more fine-grained tracking of internal behaviour. As such, they CAN be present in qlog files and tool implementers CAN add support for these, but they are not required to.

Note that in some cases, implementers might not want to log for example frame-level details in the "Core" events due to performance or privacy considerations. In this case, they SHOULD use (a subset of) relevant "Base" events instead to ensure usability of the qlog output. As an example, implementations that do not log "packet_received" events and thus also not which (if any) ACK frames the packet contain, SHOULD log "packets_acked" events instead.

Finally, for event types whose data (partially) overlap with other event types' definitions, where necessary this document includes guidance on which to use in specific situations.

2.2. Custom fields

Note that implementers are free to define new category and event types, as well as values for the "trigger" property within the "data" field, or other member fields of the "data" field, as they see fit. They SHOULD NOT however expect non-specialized tools to recognize or visualize this custom data. However, tools SHOULD make an effort to visualize even unknown data if possible in the specific tool's context.

3. Events not belonging to a single connection

For several types of events, it is sometimes impossible to tie them to a specific conceptual QUIC connection (e.g., a packet_dropped event triggered because the packet has an unknown connection_id in the header). Since qlog events in a trace are typically associated with a single connection, it is unclear how to log these events.

Ideally, implementers SHOULD create a separate, individual "endpoint-level" trace file (or group_id value), not associated with a specific connection (for example a "server.qlog" or group_id = "client"), and log all events that do not belong to a single connection to this grouping trace. However, this is not always practical, depending on the implementation. Because the semantics of most of these events are well-defined in the protocols and because they are difficult to mis-interpret as belonging to a connection, implementers MAY choose to log events not belonging to a particular connection in any other trace, even those strongly associated with a single connection.

Note that this can make it difficult to match logs from different vantage points with each other. For example, from the client side, it is easy to log connections with version negotiation or retry in the same trace, while on the server they would most likely be logged in separate traces. Servers can take extra efforts (and keep additional state) to keep these events combined in a single trace however (for example by also matching connections on their four-tuple instead of just the connection ID).

4. QUIC and HTTP/3 fields

This document re-uses all the fields defined in the main qlog schema (e.g., name, category, type, data, group_id, protocol_type, the time-related fields, etc.).

The value of the "protocol_type" qlog field MUST be "QUIC_HTTP3".

When the qlog "group_id" field is used, it is recommended to use QUIC's Original Destination Connection ID (ODCID, the CID chosen by the client when first contacting the server), as this is the only value that does not change over the course of the connection and can be used to link more advanced QUIC packets (e.g., Retry, Version Negotiation) to a given connection. Similarly, the ODCID should be used as the qlog filename or file identifier, potentially suffixed by the vantagepoint type (For example, abcd1234_server.qlog would contain the server-side trace of the connection with ODCID abcd1234).

4.1. Raw packet and frame information

While qlog is a more high-level logging format, it also allows the inclusion of most raw wire image information, such as byte lengths and even raw byte values. This can be useful when for example investigating or tuning packetization behaviour or determining encoding/framing overheads. However, these fields are not always necessary and can take up considerable space if logged for each packet or frame. As such, they are grouped in a separate optional field called "raw" of type RawInfo (where applicable).

```
class RawInfo {  
    length?:uint64; // full packet/frame length, including header and AEAD authentication tag lengths (where applicable)  
    payload_length?:uint64; // length of the packet/frame payload, excluding AEAD tag. For many control frames, this will have a value of zero  
  
    data?:bytes; // full packet/frame contents, including header and AEAD authentication tag (where applicable)  
}
```

Note: QUIC packets always include an AEAD authentication tag at the end. As this tag is always the same size for a given connection (it depends on the used TLS cipher), we do not have a separate "aead_tag_length" field here. Instead, this field is reflected in "transport:parameters_set" and can be logged only once.

Note: There is intentionally no explicit header_length field in RawInfo. QUIC and HTTP/3 use many Variable-Length Integer Encoded (VLIE) values in their packet and frame headers, which are of a dynamic length. Note too that because of this, we cannot deterministically reconstruct the header encoding/length from qlog data, as implementations might not necessarily employ the most efficient VLIE scheme for all values. As such, it is typically easier to log just the total packet/frame length and the payload length. The header length can be calculated by tools as:

For QUIC packets: $\text{header_length} = \text{length} - \text{payload_length} - \text{aead_tag_length}$

For QUIC and HTTP/3 frames: $\text{header_length} = \text{length} - \text{payload_length}$

For UDP datagrams: $\text{header_length} = \text{length} - \text{payload_length}$

Note: In some cases, the length fields are also explicitly reflected inside of frame/packet headers. For example, the QUIC STREAM frame has a "length" field indicating its payload size. Similarly, all HTTP/3 frames include their explicit payload lengths in the frame header. Finally, the QUIC Long Header has a "length" field which is equal to the payload length plus the packet number length. In these cases, those fields are intentionally preserved in the event definitions. Even though this can lead to duplicate data when the full RawInfo is logged, it allows a more direct mapping of the QUIC and HTTP/3 specifications to qlog, making it easier for users to interpret.

Note: as described in [QLOG-MAIN], the RawInfo:data field can be truncated for privacy or security purposes (for example excluding payload data). In this case, the length properties should still indicate the non-truncated lengths.

5. QUIC event definitions

Each subheading in this section is a qlog event category, while each sub-subheading is a qlog event type. Concretely, for the following two items, we have the category "connectivity" and event type "server_listening", resulting in a concatenated qlog "name" field value of "connectivity:server_listening".

5.1. connectivity

5.1.1. server_listening

Importance: Extra

Emitted when the server starts accepting connections.

Data:

```
{
  ip_v4?: IPAddress,
  ip_v6?: IPAddress,
  port_v4?: uint32,
  port_v6?: uint32,

  retry_required?:boolean // the server will always answer client initials with
  a retry (no 1-RTT connection setups by choice)
}
```

Note: some QUIC stacks do not handle sockets directly and are thus unable to log IP and/or port information.

5.1.2. connection_started

Importance: Base

Used for both attempting (client-perspective) and accepting (server-perspective) new connections. Note that this event has overlap with `connection_state_updated` and this is a separate event mainly because of all the additional data that should be logged.

Data:

```

{
  ip_version?: "v4" | "v6",
  src_ip?: IPAddress,
  dst_ip?: IPAddress,

  protocol?: string, // transport layer protocol (default "QUIC")
  src_port?: uint32,
  dst_port?: uint32,

  src_cid?: bytes,
  dst_cid?: bytes,
}

```

Note: some QUIC stacks do not handle sockets directly and are thus unable to log IP and/or port information.

5.1.3. connection_closed

Importance: Base

Used for logging when a connection was closed, typically when an error or timeout occurred. Note that this event has overlap with `connectivity:connection_state_updated`, as well as the `CONNECTION_CLOSE` frame. However, in practice, when analyzing large deployments, it can be useful to have a single event representing a `connection_closed` event, which also includes an additional reason field to provide additional information. Additionally, it is useful to log closures due to timeouts, which are difficult to reflect using the other options.

In QUIC there are two main connection-closing error categories: connection and application errors. They have well-defined error codes and semantics. Next to these however, there can be internal errors that occur that may or may not get mapped to the official error codes in implementation-specific ways. As such, multiple error codes can be set on the same event to reflect this.

```

{
  owner?: "local" | "remote", // which side closed the connection

  connection_code?: TransportError | CryptoError | uint32,
  application_code?: ApplicationError | uint32,
  internal_code?: uint32,

  reason?: string
}

```

Triggers: * clean * handshake_timeout * idle_timeout * error // this is called the "immediate close" in the QUIC specification * stateless_reset * version_mismatch * application // for example HTTP/3's GOAWAY frame

5.1.4. connection_id_updated

Importance: Base

This event is emitted when either party updates their current Connection ID. As this typically happens only sparingly over the course of a connection, this event allows loggers to be more efficient than logging the observed CID with each packet in the .header field of the "packet_sent" or "packet_received" events.

This is viewed from the perspective of the one applying the new id. As such, if we receive a new connection id from our peer, we will see the dst_ fields are set. If we update our own connection id (e.g., NEW_CONNECTION_ID frame), we log the src_ fields.

Data:

```
{
  owner: "local" | "remote",

  old?:bytes,
  new?:bytes,
}
```

5.1.5. spin_bit_updated

Importance: Base

To be emitted when the spin bit changes value. It SHOULD NOT be emitted if the spin bit is set without changing its value.

Data:

```
{
  state: boolean
}
```

5.1.6. connection_retried

TODO

5.1.7. connection_state_updated

Importance: Base

This event is used to track progress through QUIC's complex handshake and connection close procedures. It is intended to provide exhaustive options to log each state individually, but also provides a more basic, simpler set for implementations less interested in tracking each smaller state transition. As such, users should not expect to see -all- these states reflected in all qlogs and implementers should focus on support for the SimpleConnectionState set.

Data: ~~~ { old?: ConnectionState | SimpleConnectionState, new: ConnectionState | SimpleConnectionState }

```
enum ConnectionState { attempted, // initial sent/received
peer_validated, // peer address validated by: client sent Handshake
packet OR client used CONNID chosen by the server. transport-draft-
32, section-8.1 handshake_started, early_write, // 1 RTT can be sent,
but handshake isn't done yet handshake_complete, // TLS handshake
complete: Finished received and sent. tls-draft-32, section-4.1.1
handshake_confirmed, // HANDSHAKE_DONE sent/received (connection is
now "active", 1RTT can be sent). tls-draft-32, section-4.1.2 closing,
draining, // connection_close sent/received closed // draining period
done, connection state discarded }
```

```
enum SimpleConnectionState { attempted, handshake_started,
handshake_confirmed, closed } ~~~
```

These states correspond to the following transitions for both client and server:

Client:

* send initial

- state = attempted

* get initial

- state = validated _(not really "needed" at the client, but somewhat useful to indicate progress nonetheless)_

* get first Handshake packet

- state = handshake_started

- * get Handshake packet containing ServerFinished
 - state = handshake_complete
- * send ClientFinished
 - state = early_write (1RTT can now be sent)
- * get HANDSHAKE_DONE
 - state = handshake_confirmed
- *Server:*
- * get initial
 - state = attempted
- * send initial _(don't think this needs a separate state, since some handshake will always be sent in the same flight as this?)_
- * send handshake EE, CERT, CV, ...
 - state = handshake_started
- * send ServerFinished
 - state = early_write (1RTT can now be sent)
- * get first handshake packet / something using a server-issued CID of min length
 - state = validated
- * get handshake packet containing ClientFinished
 - state = handshake_complete
- * send HANDSHAKE_DONE
 - state = handshake_confirmed

Note: connection_state_changed with a new state of "attempted" is the same conceptual event as the connection_started event above from the client's perspective. Similarly, a state of "closing" or "draining" corresponds to the connection_closed event.

5.1.8. MIGRATION-related events

e.g., path_updated

TODO: read up on the draft how migration works and whether to best fit this here or in TRANSPORT TODO: integrate
<https://tools.ietf.org/html/draft-deconinck-quic-multipath-02>

For now, infer from other connectivity events and path_challenge/
path_response frames

5.2. security

5.2.1. key_updated

Importance: Base

Note: secret_updated would be more correct, but in the draft it's called KEY_UPDATE, so stick with that for consistency

Data:

```
{
  key_type:KeyType,
  old?:bytes,
  new:bytes,
  generation?:uint32 // needed for 1RTT key updates
}
```

Triggers:

- * "tls" // (e.g., initial, handshake and 0-RTT keys are generated by TLS)
- * "remote_update"
- * "local_update"

5.2.2. key_retired

Importance: Base

Data:

```
{
    key_type:KeyType,
    key?:bytes,
    generation?:uint32 // needed for 1RTT key updates
}
```

Triggers:

- * "tls" // (e.g., initial, handshake and 0-RTT keys are dropped implicitly)
- * "remote_update"
- * "local_update"

5.3. transport

5.3.1. version_information

Importance: Core

QUIC endpoints each have their own list of of QUIC versions they support. The client uses the most likely version in their first initial. If the server does support that version, it replies with a version_negotiation packet, containing supported versions. From this, the client selects a version. This event aggregates all this information in a single event type. It also allows logging of supported versions at an endpoint without actual version negotiation needing to happen.

Data:

```
{
    server_versions?:Array<bytes>,
    client_versions?:Array<bytes>,
    chosen_version?:bytes
}
```

Intended use:

- * When sending an initial, the client logs this event with client_versions and chosen_version set
- * Upon receiving a client initial with a supported version, the server logs this event with server_versions and chosen_version set

- * Upon receiving a client initial with an unsupported version, the server logs this event with `server_versions` set and `client_versions` to the single-element array containing the client's attempted version. The absence of `chosen_version` implies no overlap was found.
- * Upon receiving a version negotiation packet from the server, the client logs this event with `client_versions` set and `server_versions` to the versions in the version negotiation packet and `chosen_version` to the version it will use for the next initial packet

5.3.2. `alpn_information`

Importance: Core

QUIC implementations each have their own list of application level protocols and versions thereof they support. The client includes a list of their supported options in its first initial as part of the TLS Application Layer Protocol Negotiation (alpn) extension. If there are common option(s), the server chooses the most optimal one and communicates this back to the client. If not, the connection is closed.

Data:

```
{
  server_alpns?:Array<string>,
  client_alpns?:Array<string>,
  chosen_alpn?:string
}
```

Intended use:

- * When sending an initial, the client logs this event with `client_alpns` set
- * When receiving an initial with a supported alpn, the server logs this event with `server_alpns` set, `client_alpns` equalling the client-provided list, and `chosen_alpn` to the value it will send back to the client.
- * When receiving an initial with an alpn, the client logs this event with `chosen_alpn` to the received value.
- * Alternatively, a client can choose to not log the first event, but wait for the receipt of the server initial to log this event with both `client_alpns` and `chosen_alpn` set.

5.3.3. parameters_set

Importance: Core

This event groups settings from several different sources (transport parameters, TLS ciphers, etc.) into a single event. This is done to minimize the amount of events and to decouple conceptual setting impacts from their underlying mechanism for easier high-level reasoning.

All these settings are typically set once and never change. However, they are typically set at different times during the connection, so there will typically be several instances of this event with different fields set.

Note that some settings have two variations (one set locally, one requested by the remote peer). This is reflected in the "owner" field. As such, this field **MUST** be correct for all settings included a single event instance. If you need to log settings from two sides, you **MUST** emit two separate event instances.

In the case of connection resumption and 0-RTT, some of the server's parameters are stored up-front at the client and used for the initial connection startup. They are later updated with the server's reply. In these cases, utilize the separate "parameters_restored" event to indicate the initial values, and this event to indicate the updated values, as normal.

Data:

```

{
    owner?: "local" | "remote",

    resumption_allowed?: boolean, // valid session ticket was received
    early_data_enabled?: boolean, // early data extension was enabled on the TLS layer
    tls_cipher?: string, // (e.g., "AES_128_GCM_SHA256")
    aead_tag_length?: uint8, // depends on the TLS cipher, but it's easier to be explicit. Default value is 16

    // transport parameters from the TLS layer:
    original_destination_connection_id?: bytes,
    initial_source_connection_id?: bytes,
    retry_source_connection_id?: bytes,
    stateless_reset_token?: Token,
    disable_active_migration?: boolean,

    max_idle_timeout?: uint64,
    max_udp_payload_size?: uint32,
    ack_delay_exponent?: uint16,
    max_ack_delay?: uint16,
    active_connection_id_limit?: uint32,

    initial_max_data?: uint64,
    initial_max_stream_data_bidi_local?: uint64,
    initial_max_stream_data_bidi_remote?: uint64,
    initial_max_stream_data_uni?: uint64,
    initial_max_streams_bidi?: uint64,
    initial_max_streams_uni?: uint64,

    preferred_address?: PreferredAddress
}

interface PreferredAddress {
    ip_v4: IPAddress,
    ip_v6: IPAddress,

    port_v4: uint16,
    port_v6: uint16,

    connection_id: bytes,
    stateless_reset_token: Token
}

```

Additionally, this event can contain any number of unspecified fields. This is to reflect setting of for example unknown (greased) transport parameters or employed (proprietary) extensions.

5.3.4. parameters_restored

Importance: Base

When using QUIC 0-RTT, clients are expected to remember and restore the server's transport parameters from the previous connection. This event is used to indicate which parameters were restored and to which values when utilizing 0-RTT. Note that not all transport parameters should be restored (many are even prohibited from being re-utilized). The ones listed here are the ones expected to be useful for correct 0-RTT usage.

Data:

```
{
  disable_active_migration?:boolean,

  max_idle_timeout?:uint64,
  max_udp_payload_size?:uint32,
  active_connection_id_limit?:uint32,

  initial_max_data?:uint64,
  initial_max_stream_data_bidi_local?:uint64,
  initial_max_stream_data_bidi_remote?:uint64,
  initial_max_stream_data_uni?:uint64,
  initial_max_streams_bidi?:uint64,
  initial_max_streams_uni?:uint64,
}
```

Note that, like parameters_set above, this event can contain any number of unspecified fields to allow for additional/custom parameters.

5.3.5. packet_sent

Importance: Core

Data:

```
{
  header:PacketHeader,

  frames?:Array<QuicFrame>, // see appendix for the definitions

  is_coalesced?:boolean, // default value is false

  retry_token?:Token, // only if header.packet_type === retry

  stateless_reset_token?:bytes, // only if header.packet_type === stateless_reset. Is always 128 bits in length.

  supported_versions:Array<bytes>, // only if header.packet_type === version_negotiation

  raw?:RawInfo,
  datagram_id?:uint32
}
```

Note: We do not explicitly log the `encryption_level` or `packet_number_space`: the `header.packet_type` specifies this by inference (assuming correct implementation)

Triggers:

- * "retransmit_reordered" // draft-23 5.1.1
- * "retransmit_timeout" // draft-23 5.1.2
- * "pto_probe" // draft-23 5.3.1
- * "retransmit_crypto" // draft-19 6.2
- * "cc_bandwidth_probe" // needed for some CCs to figure out bandwidth allocations when there are no normal sends

Note: for more details on "datagram_id", see Section 5.3.10. It is only needed when keeping track of packet coalescing.

5.3.6. packet_received

Importance: Core

Data:

```
{
  header:PacketHeader,

  frames?:Array<QuicFrame>, // see appendix for the definitions

  is_coalesced?:boolean,

  retry_token?:Token, // only if header.packet_type === retry

  stateless_reset_token?:bytes, // only if header.packet_type === stateless_reset. Is always 128 bits in length.

  supported_versions:Array<bytes>, // only if header.packet_type === version_negotiation

  raw?:RawInfo,
  datagram_id?:uint32
}
```

Note: We do not explicitly log the encryption_level or packet_number_space: the header.packet_type specifies this by inference (assuming correct implementation)

Triggers:

* "keys_available" // if packet was buffered because it couldn't be decrypted before

Note: for more details on "datagram_id", see Section 5.3.10. It is only needed when keeping track of packet coalescing.

5.3.7. packet_dropped

Importance: Base

This event indicates a QUIC-level packet was dropped after partial or no parsing.

Data:

```
{
  header?:PacketHeader, // primarily packet_type should be filled here, as other fields might not be parseable

  raw?:RawInfo,
  datagram_id?:uint32
}
```

For this event, the "trigger" field SHOULD be set (for example to one of the values below), as this helps tremendously in debugging.

Triggers:

- * "key_unavailable"
- * "unknown_connection_id"
- * "header_parse_error"
- * "payload_decrypt_error"
- * "protocol_violation"
- * "dos_prevention"
- * "unsupported_version"
- * "unexpected_packet"
- * "unexpected_source_connection_id"
- * "unexpected_version"
- * "duplicate"
- * "invalid_initial"

Note: sometimes packets are dropped before they can be associated with a particular connection (e.g., in case of "unsupported_version"). This situation is discussed more in Section 3.

Note: for more details on "datagram_id", see Section 5.3.10. It is only needed when keeping track of packet coalescing.

5.3.8. packet_buffered

Importance: Base

This event is emitted when a packet is buffered because it cannot be processed yet. Typically, this is because the packet cannot be parsed yet, and thus we only log the full packet contents when it was parsed in a packet_received event.

Data:

```
{
  header?:PacketHeader, // primarily packet_type and possible packet_number should
  // be filled here, as other elements might not be available yet

  raw?:RawInfo,
  datagram_id?:uint32
}
```

Note: for more details on "datagram_id", see Section 5.3.10. It is only needed when keeping track of packet coalescing.

Triggers:

- * "backpressure" // indicates the parser cannot keep up, temporarily buffers packet for later processing
- * "keys_unavailable" // if packet cannot be decrypted because the proper keys were not yet available

5.3.9. packets_acked

Importance: Extra

This event is emitted when a (group of) sent packet(s) is acknowledged by the remote peer `_for the first time_`. This information could also be deduced from the contents of received ACK frames. However, ACK frames require additional processing logic to determine when a given packet is acknowledged for the first time, as QUIC uses ACK ranges which can include repeated ACKs. Additionally, this event can be used by implementations that do not log frame contents.

Data: ~~~ { packet_number_space?:PacketNumberSpace,
packet_numbers?:Array<uint64> } ~~~

Note: if packet_number_space is omitted, it assumes the default value of PacketNumberSpace.application_data, as this is by far the most prevalent packet number space a typical QUIC connection will use.

5.3.10. datagrams_sent

Importance: Extra

When we pass one or more UDP-level datagrams to the socket. This is useful for determining how QUIC packet buffers are drained to the OS.

Data:

```

{
    count?:uint16, // to support passing multiple at once
    raw?:Array<RawInfo>, // RawInfo:length field indicates total length of the da
tagrams, including UDP header length

    datagram_ids?:Array<uint32>
}

```

Note: QUIC itself does not have a concept of a "datagram_id". This field is a purely qlong-specific construct to allow tracking how multiple QUIC packets are coalesced inside of a single UDP datagram, which is an important optimization during the QUIC handshake. For this, implementations assign a (per-endpoint) unique ID to each datagram and keep track of which packets were coalesced into the same datagram. As packet coalescing typically only happens during the handshake (as it requires at least one long header packet), this can be done without much overhead.

5.3.11. datagrams_received

Importance: Extra

When we receive one or more UDP-level datagrams from the socket. This is useful for determining how datagrams are passed to the user space stack from the OS.

Data:

```

{
    count?:uint16, // to support passing multiple at once
    raw?:Array<RawInfo>, // RawInfo:length field indicates total length of the da
tagrams, including UDP header length

    datagram_ids?:Array<uint32>
}

```

Note: for more details on "datagram_ids", see Section 5.3.10.

5.3.12. datagram_dropped

Importance: Extra

When we drop a UDP-level datagram. This is typically if it does not contain a valid QUIC packet (in that case, use packet_dropped instead).

Data:


```
{  
    raw?:RawInfo  
}
```

5.3.13. stream_state_updated

Importance: Base

This event is emitted whenever the internal state of a QUIC stream is updated, as described in QUIC transport draft-23 section 3. Most of this can be inferred from several types of frames going over the wire, but it's much easier to have explicit signals for these state changes.

Data:

```

{
    stream_id:uint64,
    stream_type?:"unidirectional"|"bidirectional", // mainly useful when opening
the stream

    old?:StreamState,
    new:StreamState,

    stream_side?:"sending"|"receiving"
}

enum StreamState {
    // bidirectional stream states, draft-23 3.4.
    idle,
    open,
    half_closed_local,
    half_closed_remote,
    closed,

    // sending-side stream states, draft-23 3.1.
    ready,
    send,
    data_sent,
    reset_sent,
    reset_received,

    // receive-side stream states, draft-23 3.2.
    receive,
    size_known,
    data_read,
    reset_read,

    // both-side states
    data_received,

    // qlog-defined
    destroyed // memory actually freed
}

```

Note: QUIC implementations SHOULD mainly log the simplified bidirectional (HTTP/2-alike) stream states (e.g., idle, open, closed) instead of the more finegrained stream states (e.g., data_sent, reset_received). These latter ones are mainly for more in-depth debugging. Tools SHOULD be able to deal with both types equally.

5.3.14. frames_processed

Importance: Extra

This event's main goal is to prevent a large proliferation of specific purpose events (e.g., `packets_acknowledged`, `flow_control_updated`, `stream_data_received`). We want to give implementations the opportunity to (selectively) log this type of signal without having to log packet-level details (e.g., in `packet_received`). Since for almost all cases, the effects of applying a frame to the internal state of an implementation can be inferred from that frame's contents, we aggregate these events in this single `"frames_processed"` event.

Note: This event can be used to signal internal state change not resulting directly from the actual "parsing" of a frame (e.g., the frame could have been parsed, data put into a buffer, then later processed, then logged with this event).

Note: Implementations logging `"packet_received"` and which include all of the packet's constituent frames therein, are not expected to emit this `"frames_processed"` event (contrary to the HTTP-level `"frames_parsed"` event). Rather, implementations not wishing to log full packets or that wish to explicitly convey extra information about when frames are processed (if not directly tied to their reception) can use this event.

Note: for some events, this approach will lose some information (e.g., for which encryption level are packets being acknowledged?). If this information is important, please use the `packet_received` event instead.

Note: in some implementations, it can be difficult to log frames directly, even when using `packet_sent` and `packet_received` events. For these cases, this event also contains the direct `packet_number` field, which can be used to more explicitly link this event to the `packet_sent/received` events.

Data:

```
{
  frames:Array<QuicFrame>, // see appendix for the definitions
  packet_number?:uint64
}
```

5.3.15. `data_moved`

Importance: Base

Used to indicate when data moves between the different layers (for example passing from HTTP/3 to QUIC stream buffers and vice versa) or between HTTP/3 and the actual user application on top (for example a browser engine). This helps make clear the flow of data, how long data remains in various buffers and the overheads introduced by individual layers.

For example, this helps make clear whether received data on a QUIC stream is moved to the HTTP layer immediately (for example per received packet) or in larger batches (for example, all QUIC packets are processed first and afterwards the HTTP layer reads from the streams with newly available data). This in turn can help identify bottlenecks or scheduling problems.

Data:

```
{
  stream_id?:uint64,
  offset?:uint64,
  length?:uint64, // byte length of the moved data

  from?:string, // typically: use either of "application","http","transport"
  to?:string, // typically: use either of "application","http","transport"

  data?:bytes // raw bytes that were transferred
}
```

Note: we do not for example use a "direction" field (with values "up" and "down") to specify the data flow. This is because in some optimized implementations, data might skip some individual layers. Additionally, using explicit "from" and "to" fields is more flexible and allows the definition of other conceptual "layers" (for example to indicate data from QUIC CRYPTO frames being passed to a TLS library ("security") or from HTTP/3 to QPACK ("qpack")).

Note: this event type is part of the "transport" category, but really spans all the different layers. This means we have a few leaky abstractions here (for example, the stream_id or stream offset might not be available at some logging points, or the raw data might not be in a byte-array form). In these situations, implementers can decide to define new, in-context fields to aid in manual debugging.

5.4. recovery

Note: most of the events in this category are kept generic to support different recovery approaches and various congestion control algorithms. Tool creators SHOULD make an effort to support and visualize even unknown data in these events (e.g., plot unknown congestion states by name on a timeline visualization).

5.4.1. parameters_set

Importance: Base

This event groups initial parameters from both loss detection and congestion control into a single event. All these settings are typically set once and never change. Implementation that do, for some reason, change these parameters during execution, MAY emit the parameters_set event twice.

Data:

```
{
  // Loss detection, see recovery draft-23, Appendix A.2
  reordering_threshold?:uint16, // in amount of packets
  time_threshold?:float, // as RTT multiplier
  timer_granularity?:uint16, // in ms
  initial_rtt?:float, // in ms

  // congestion control, Appendix B.1.
  max_datagram_size?:uint32, // in bytes // Note: this could be updated after p
mtud
  initial_congestion_window?:uint64, // in bytes
  minimum_congestion_window?:uint32, // in bytes // Note: this could change whe
n max_datagram_size changes
  loss_reduction_factor?:float,
  persistent_congestion_threshold?:uint16 // as PTO multiplier
}
```

Additionally, this event can contain any number of unspecified fields to support different recovery approaches.

5.4.2. metrics_updated

Importance: Core

This event is emitted when one or more of the observable recovery metrics changes value. This event SHOULD group all possible metric updates that happen at or around the same time in a single event (e.g., if `min_rtt` and `smoothed_rtt` change at the same time, they should be bundled in a single `metrics_updated` entry, rather than split out into two). Consequently, a `metrics_updated` event is only guaranteed to contain at least one of the listed metrics.

Data:

```
{
  // Loss detection, see recovery draft-23, Appendix A.3
  min_rtt?:float, // in ms or us, depending on the overarching qlog's configura
tion
  smoothed_rtt?:float, // in ms or us, depending on the overarching qlog's conf
iguration
  latest_rtt?:float, // in ms or us, depending on the overarching qlog's config
uration
  rtt_variance?:float, // in ms or us, depending on the overarching qlog's conf
iguration

  pto_count?:uint16,

  // Congestion control, Appendix B.2.
  congestion_window?:uint64, // in bytes
  bytes_in_flight?:uint64,

  ssthresh?:uint64, // in bytes

  // qlog defined
  packets_in_flight?:uint64, // sum of all packet number spaces

  pacing_rate?:uint64 // in bps
}
```

Note: to make logging easier, implementations MAY log values even if they are the same as previously reported values (e.g., two subsequent `METRIC_UPDATE` entries can both report the exact same value for `min_rtt`). However, applications SHOULD try to log only actual updates to values.

Additionally, this event can contain any number of unspecified fields to support different recovery approaches.

5.4.3. `congestion_state_updated`

Importance: Base

This event signifies when the congestion controller enters a significant new state and changes its behaviour. This event's definition is kept generic to support different Congestion Control algorithms. For example, for the algorithm defined in the Recovery draft ("enhanced" New Reno), the following states are defined:

- * slow_start
- * congestion_avoidance
- * application_limited
- * recovery

Data:

```
{
  old?:string,
  new:string
}
```

The "trigger" field SHOULD be logged if there are multiple ways in which a state change can occur but MAY be omitted if a given state can only be due to a single event occurring (e.g., slow start is exited only when ssthresh is exceeded).

Some triggers for ("enhanced" New Reno):

- * persistent_congestion
- * ECN

5.4.4. loss_timer_updated

Importance: Extra

This event is emitted when a recovery loss timer changes state. The three main event types are:

- * set: the timer is set with a delta timeout for when it will trigger next
- * expired: when the timer effectively expires after the delta timeout
- * cancelled: when a timer is cancelled (e.g., all outstanding packets are acknowledged, start idle period)

Note: to indicate an active timer's timeout update, a new "set" event is used.

Data:

```
{
  timer_type?: "ack" | "pto", // called "mode" in draft-23 A.9.
  packet_number_space?: PacketNumberSpace,

  event_type: "set" | "expired" | "cancelled",

  delta?: float // if event_type === "set": delta time in ms or us (see configur
ation) from this event's timestamp until when the timer will trigger
}
```

TODO: how about CC algo's that use multiple timers? How generic do these events need to be? Just support QUIC-style recovery from the spec or broader?

TODO: read up on the loss detection logic in draft-27 onward and see if this suffices

5.4.5. packet_lost

Importance: Core

This event is emitted when a packet is deemed lost by loss detection.

Data:

```
{
  header?: PacketHeader, // should include at least the packet_type and packet_n
umber

  // not all implementations will keep track of full packets, so these are opti
onal
  frames?: Array<QuicFrame> // see appendix for the definitions
}
```

For this event, the "trigger" field SHOULD be set (for example to one of the values below), as this helps tremendously in debugging.

Triggers:

- * "reordering_threshold",
- * "time_threshold"
- * "pto_expired" // draft-23 section 5.3.1, MAY

5.4.6. marked_for_retransmit

Importance: Extra

This event indicates which data was marked for retransmit upon detecting a packet loss (see `packet_lost`). Similar to our reasoning for the "frames_processed" event, in order to keep the amount of different events low, we group this signal for all types of retransmittable data in a single event based on existing QUIC frame definitions.

Implementations retransmitting full packets or frames directly can just log the constituent frames of the lost packet here (or do away with this event and use the contents of the `packet_lost` event instead). Conversely, implementations that have more complex logic (e.g., marking ranges in a stream's data buffer as in-flight), or that do not track sent frames in full (e.g., only stream offset + length), can translate their internal behaviour into the appropriate frame instance here even if that frame was never or will never be put on the wire.

Note: much of this data can be inferred if implementations log `packet_sent` events (e.g., looking at overlapping stream data offsets and length, one can determine when data was retransmitted).

Data:

```
{
  frames:Array<QuicFrame>, // see appendix for the definitions
}
```

6. HTTP/3 event definitions

6.1. http

Note: like all category values, the "http" category is written in lowercase.

6.1.1. parameters_set

Importance: Base

This event contains HTTP/3 and QPACK-level settings, mostly those received from the HTTP/3 SETTINGS frame. All these parameters are typically set once and never change. However, they are typically set at different times during the connection, so there can be several instances of this event with different fields set.

Note that some settings have two variations (one set locally, one requested by the remote peer). This is reflected in the "owner" field. As such, this field MUST be correct for all settings included a single event instance. If you need to log settings from two sides, you MUST emit two separate event instances.

Data:

```
{
  owner?: "local" | "remote",

  max_header_list_size?: uint64, // from SETTINGS_MAX_HEADER_LIST_SIZE
  max_table_capacity?: uint64, // from SETTINGS_QPACK_MAX_TABLE_CAPACITY
  blocked_streams_count?: uint64, // from SETTINGS_QPACK_BLOCKED_STREAMS

  // qlog-defined
  waits_for_settings?: boolean // indicates whether this implementation waits for
  a SETTINGS frame before processing requests
}
```

Note: enabling server push is not explicitly done in HTTP/3 by use of a setting or parameter. Instead, it is communicated by use of the MAX_PUSH_ID frame, which should be logged using the frame_created and frame_parsed events below.

Additionally, this event can contain any number of unspecified fields. This is to reflect setting of for example unknown (greased) settings or parameters of (proprietary) extensions.

6.1.2. parameters_restored

Importance: Base

When using QUIC 0-RTT, clients are expected to remember and reuse the server's SETTINGS from the previous connection. This event is used to indicate which settings were restored and to which values when utilizing 0-RTT.

Data:

```
{
  max_header_list_size?: uint64,
  max_table_capacity?: uint64,
  blocked_streams_count?: uint64
}
```

Note that, like for parameters_set above, this event can contain any number of unspecified fields to allow for additional and custom settings.

6.1.3. stream_type_set

Importance: Base

Emitted when a stream's type becomes known. This is typically when a stream is opened and the stream's type indicator is sent or received.

Note: most of this information can also be inferred by looking at a stream's id, since id's are strictly partitioned at the QUIC level. Even so, this event has a "Base" importance because it helps a lot in debugging to have this information clearly spelled out.

Data:

```
{
  stream_id:uint64,

  owner?:"local"|"remote"

  old?:StreamType,
  new:StreamType,

  associated_push_id?:uint64 // only when new == "push"
}

enum StreamType {
  data, // bidirectional request-response streams
  control,
  push,
  reserved,
  qpack_encode,
  qpack_decode
}
```

6.1.4. frame_created

Importance: Core

HTTP equivalent to the packet_sent event. This event is emitted when the HTTP/3 framing actually happens. Note: this is not necessarily the same as when the HTTP/3 data is passed on to the QUIC layer. For that, see the "data_moved" event.

Data:

```
{
  stream_id:uint64,
  length?:uint64, // payload byte length of the frame
  frame:HTTP3Frame, // see appendix for the definitions,

  raw?:RawInfo
}
```

Note: in HTTP/3, DATA frames can have arbitrarily large lengths to reduce frame header overhead. As such, DATA frames can span many QUIC packets and can be created in a streaming fashion. In this case, the `frame_created` event is emitted once for the frame header, and further streamed data is indicated using the `data_moved` event.

6.1.5. `frame_parsed`

Importance: Core

HTTP equivalent to the `packet_received` event. This event is emitted when we actually parse the HTTP/3 frame. Note: this is not necessarily the same as when the HTTP/3 data is actually received on the QUIC layer. For that, see the `"data_moved"` event.

Data:

```
{
  stream_id:uint64,
  length?:uint64, // payload byte length of the frame
  frame:HTTP3Frame, // see appendix for the definitions,

  raw?:RawInfo
}
```

Note: in HTTP/3, DATA frames can have arbitrarily large lengths to reduce frame header overhead. As such, DATA frames can span many QUIC packets and can be processed in a streaming fashion. In this case, the `frame_parsed` event is emitted once for the frame header, and further streamed data is indicated using the `data_moved` event.

6.1.6. `push_resolved`

Importance: Extra

This event is emitted when a pushed resource is successfully claimed (used) or, conversely, abandoned (rejected) by the application on top of HTTP/3 (e.g., the web browser). This event is added to help debug problems with unexpected PUSH behaviour, which is commonplace with HTTP/2.

```
{
  push_id?:uint64,
  stream_id?:uint64, // in case this is logged from a place that does not have
access to the push_id

  decision:"claimed"|"abandoned"
}
```

6.2. qpack

Note: like all category values, the "qpack" category is written in lowercase.

The QPACK events mainly serve as an aid to debug low-level QPACK issues. The higher-level, plaintext header values SHOULD (also) be logged in the http.frame_created and http.frame_parsed event data (instead).

Note: qpack does not have its own parameters_set event. This was merged with http.parameters_set for brevity, since qpack is a required extension for HTTP/3 anyway. Other HTTP/3 extensions MAY also log their SETTINGS fields in http.parameters_set or MAY define their own events.

6.2.1. state_updated

Importance: Base

This event is emitted when one or more of the internal QPACK variables changes value. Note that some variables have two variations (one set locally, one requested by the remote peer). This is reflected in the "owner" field. As such, this field MUST be correct for all variables included a single event instance. If you need to log settings from two sides, you MUST emit two separate event instances.

Data:

```
{
  owner:"local" | "remote",

  dynamic_table_capacity?:uint64,
  dynamic_table_size?:uint64, // effective current size, sum of all the entries

  known_received_count?:uint64,
  current_insert_count?:uint64
}
```

6.2.2. stream_state_updated

Importance: Core

This event is emitted when a stream becomes blocked or unblocked by header decoding requests or QPACK instructions.

Note: This event is of "Core" importance, as it might have a large impact on HTTP/3's observed performance.

Data:

```
{
  stream_id:uint64,

  state:"blocked"|"unblocked" // streams are assumed to start "unblocked" until
  they become "blocked"
}
```

6.2.3. dynamic_table_updated

Importance: Extra

This event is emitted when one or more entries are inserted or evicted from QPACK's dynamic table.

Data:

```
{
  owner:"local" | "remote", // local = the encoder's dynamic table. remote = th
  e decoder's dynamic table

  update_type:"inserted"|"evicted",

  entries:Array<DynamicTableEntry>
}

class DynamicTableEntry {
  index:uint64;
  name?:string | bytes;
  value?:string | bytes;
}
```

6.2.4. headers_encoded

Importance: Base

This event is emitted when an uncompressed header block is encoded successfully.

Note: this event has overlap with `http.frame_created` for the `HeadersFrame` type. When outputting both events, implementers MAY omit the "headers" field in this event.

Data:

```
{
  stream_id?:uint64,

  headers?:Array<HTTPHeader>,

  block_prefix:QPackHeaderBlockPrefix,
  header_block:Array<QPackHeaderBlockRepresentation>,

  length?:uint32,
  raw?:bytes
}
```

6.2.5. headers_decoded

Importance: Base

This event is emitted when a compressed header block is decoded successfully.

Note: this event has overlap with `http.frame_parsed` for the `HeadersFrame` type. When outputting both events, implementers MAY omit the "headers" field in this event.

Data:

```
{
  stream_id?:uint64,

  headers?:Array<HTTPHeader>,

  block_prefix:QPackHeaderBlockPrefix,
  header_block:Array<QPackHeaderBlockRepresentation>,

  length?:uint32,
  raw?:bytes
}
```

6.2.6. instruction_created

Importance: Base

This event is emitted when a QPACK instruction (both decoder and encoder) is created and added to the encoder/decoder stream.

Data:

```
{
  instruction:QPackInstruction // see appendix for the definitions,
  length?:uint32,
  raw?:bytes
}
```

Note: encoder/decoder semantics and stream_id's are implicit in either the instruction types or can be logged via other events (e.g., http.stream_type_set)

6.2.7. instruction_parsed

Importance: Base

This event is emitted when a QPACK instruction (both decoder and encoder) is read from the encoder/decoder stream.

Data:

```
{
  instruction:QPackInstruction // see appendix for the definitions,
  length?:uint32,
  raw?:bytes
}
```

Note: encoder/decoder semantics and stream_id's are implicit in either the instruction types or can be logged via other events (e.g., http.stream_type_set)

7. Generic events and Simulation indicators

7.1. generic

The main goal of the events in this category is to allow implementations to fully replace their existing text-based logging by qlog. This is done by providing events to log generic strings for typical well-known logging levels (error, warning, info, debug, verbose).

7.1.1. error

Importance: Core

Used to log details of an internal error. For errors that effectively lead to the closure of a QUIC connection, it is recommended to use `transport:connection_closed` instead.

Data:

```
{
  code?:uint32,
  message?:string
}
```

7.1.2. warning

Importance: Base

Used to log details of an internal warning that might not get reflected on the wire.

Data:

```
{
  code?:uint32,
  message?:string
}
```

7.1.3. info

Importance: Extra

Used mainly for implementations that want to use `qlog` as their one and only logging format but still want to support unstructured string messages.

Data:

```
{
  message:string
}
```

7.1.4. debug

Importance: Extra

Used mainly for implementations that want to use qlog as their one and only logging format but still want to support unstructured string messages.

Data:

```
{
  message:string
}
```

7.1.5. verbose

Importance: Extra

Used mainly for implementations that want to use qlog as their one and only logging format but still want to support unstructured string messages.

Data:

```
{
  message:string
}
```

7.2. simulation

When evaluating a protocol evaluation, one typically sets up a series of interoperability or benchmarking tests, in which the test situations can change over time. For example, the network bandwidth or latency can vary during the test, or the network can be fully disable for a short time. In these setups, it is useful to know when exactly these conditions are triggered, to allow for proper correlation with other events.

7.2.1. scenario

Importance: Extra

Used to specify which specific scenario is being tested at this particular instance. This could also be reflected in the top-level qlog's "summary" or "configuration" fields, but having a separate event allows easier aggregation of several simulations into one trace.

```
{
  name?:string,
  details?:any
}
```

7.2.2. marker

Importance: Extra

Used to indicate when specific emulation conditions are triggered at set times (e.g., at 3 seconds in 2% packet loss is introduced, at 10s a NAT rebind is triggered).

```
{  
    type?:string,  
    message?:string  
}
```

8. Security Considerations

TBD

9. IANA Considerations

TBD

10. References

10.1. Normative References

[QLOG-MAIN]

Marx, R., Ed., "Main logging schema for qlog", Work in Progress, Internet-Draft, draft-marx-qlog-main-schema-02, 2 November 2020, <<https://tools.ietf.org/html/draft-marx-qlog-main-schema-02>>.

[QUIC-HTTP]

Bishop, M., Ed., "Hypertext Transfer Protocol Version 3 (HTTP/3)", Work in Progress, Internet-Draft, draft-ietf-quic-http-32, 1 October 2020, <<https://tools.ietf.org/html/draft-ietf-quic-http-32>>.

[QUIC-QPACK]

Frindell, A., Ed., "QPACK: Header Compression for HTTP/3", Work in Progress, Internet-Draft, draft-ietf-quic-qpack-19, 20 October 2020, <<https://tools.ietf.org/html/draft-ietf-quic-qpack-19>>.

[QUIC-TRANSPORT]

Iyengar, J., Ed. and M. Thomson, Ed., "QUIC: A UDP-Based Multiplexed and Secure Transport", Work in Progress, Internet-Draft, draft-ietf-quic-transport-32, 1 October 2020, <<https://tools.ietf.org/html/draft-ietf-quic-transport-32>>.

10.2. Informative 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>>.

Appendix A. QUIC data field definitions

A.1. IPAddress

```
class IPAddress : string | bytes;
```

// an IPAddress can either be a "human readable" form (e.g., "127.0.0.1" for v4 or "2001:0db8:85a3:0000:0000:8a2e:0370:7334" for v6) or use a raw byte-form (as the string forms can be ambiguous)

A.2. PacketType

```
enum PacketType {  
    initial,  
    handshake,  
    zerortt = "0RTT",  
    onertt = "1RTT",  
    retry,  
    version_negotiation,  
    stateless_reset,  
    unknown  
}
```

A.3. PacketNumberSpace

```
enum PacketNumberSpace {  
    initial,  
    handshake,  
    application_data  
}
```

A.4. PacketHeader

```

class PacketHeader {
    // Note: short vs long header is implicit through PacketType

    packet_type: PacketType;
    packet_number: uint64;

    flags?: uint8; // the bit flags of the packet headers (spin bit, key update b
it, etc. up to and including the packet number length bits if present) interprete
d as a single 8-bit integer

    token?:Token; // only if packet_type == initial

    length?: uint16, // only if packet_type == initial || handshake || 0RTT. Sign
ifies length of the packet_number plus the payload.

    // only if present in the header
    // if correctly using transport:connection_id_updated events,
    // dcid can be skipped for 1RTT packets
    version?: bytes; // e.g., "ff00001d" for draft-29
    scil?: uint8;
    dcil?: uint8;
    scid?: bytes;
    dcid?: bytes;
}

```

A.5. Token

```

class Token {
    type?: "retry"|"resumption"|"stateless_reset";

    length?:uint32; // byte length of the token
    data?:bytes; // raw byte value of the token

    details?:any; // decoded fields included in the token (typically: peer's IP a
ddress, creation time)
}

```

The token carried in an Initial packet can either be a retry token from a Retry packet, a stateless reset token from a Stateless Reset packet or one originally provided by the server in a NEW_TOKEN frame used when resuming a connection (e.g., for address validation purposes). Retry and resumption tokens typically contain encoded metadata to check the token's validity when it is used, but this metadata and its format is implementation specific. For that, this field includes a general-purpose "details" field.

A.6. KeyType

```
enum KeyType {
    server_initial_secret,
    client_initial_secret,

    server_handshake_secret,
    client_handshake_secret,

    server_0rtt_secret,
    client_0rtt_secret,

    server_1rtt_secret,
    client_1rtt_secret
}
```

A.7. QUIC Frames

```
type QuicFrame = PaddingFrame | PingFrame | AckFrame | ResetStreamFrame | StopSendingFrame | CryptoFrame | NewTokenFrame | StreamFrame | MaxDataFrame | MaxStreamDataFrame | MaxStreamsFrame | DataBlockedFrame | StreamDataBlockedFrame | StreamsBlockedFrame | NewConnectionIDFrame | RetireConnectionIDFrame | PathChallengeFrame | PathResponseFrame | ConnectionCloseFrame | HandshakeDoneFrame | UnknownFrame;
```

A.7.1. PaddingFrame

In QUIC, PADDING frames are simply identified as a single byte of value 0. As such, each padding byte could be theoretically interpreted and logged as an individual `PaddingFrame`.

However, as this leads to heavy logging overhead, implementations SHOULD instead emit just a single `PaddingFrame` and set the `payload_length` property to the amount of PADDING bytes/frames included in the packet.

```
class PaddingFrame{
    frame_type:string = "padding";

    length?:uint32; // total frame length, including frame header
    payload_length?:uint32;
}
```

A.7.2. PingFrame

```
class PingFrame{
    frame_type:string = "ping";

    length?:uint32; // total frame length, including frame header
    payload_length?:uint32;
}
```

A.7.3. AckFrame

```

class AckFrame{
    frame_type:string = "ack";

    ack_delay?:float; // in ms

    // first number is "from": lowest packet number in interval
    // second number is "to": up to and including // highest packet number in interval
    // e.g., looks like [[1,2],[4,5]]
    acked_ranges?:Array<[uint64, uint64]|[uint64]>;

    // ECN (explicit congestion notification) related fields (not always present)
    ect1?:uint64;
    ect0?:uint64;
    ce?:uint64;

    length?:uint32; // total frame length, including frame header
    payload_length?:uint32;
}

```

Note: the packet ranges in `AckFrame.acked_ranges` do not necessarily have to be ordered (e.g., `[[5,9],[1,4]]` is a valid value).

Note: the two numbers in the packet range can be the same (e.g., `[120,120]` means that packet with number 120 was ACKed). However, in that case, implementers SHOULD log `[120]` instead and tools MUST be able to deal with both notations.

A.7.4. ResetStreamFrame

```

class ResetStreamFrame{
    frame_type:string = "reset_stream";

    stream_id:uint64;
    error_code:ApplicationError | uint32;
    final_size:uint64; // in bytes

    length?:uint32; // total frame length, including frame header
    payload_length?:uint32;
}

```

A.7.5. StopSendingFrame

```

class StopSendingFrame{
    frame_type:string = "stop_sending";

    stream_id:uint64;
    error_code:ApplicationError | uint32;

    length?:uint32; // total frame length, including frame header
    payload_length?:uint32;
}

```

A.7.6. CryptoFrame

```

class CryptoFrame{
    frame_type:string = "crypto";

    offset:uint64;
    length:uint64;

    payload_length?:uint32;
}

```

A.7.7. NewTokenFrame

```

class NewTokenFrame{
    frame_type:string = "new_token";

    token:Token
}

```

A.7.8. StreamFrame

```

class StreamFrame{
    frame_type:string = "stream";

    stream_id:uint64;

    // These two MUST always be set
    // If not present in the Frame type, log their default values
    offset:uint64;
    length:uint64;

    // this MAY be set any time, but MUST only be set if the value is "true"
    // if absent, the value MUST be assumed to be "false"
    fin?:boolean;

    raw?:bytes;
}

```


A.7.9. MaxDataFrame

```
class MaxDataFrame{
    frame_type:string = "max_data";

    maximum:uint64;
}
```

A.7.10. MaxStreamDataFrame

```
class MaxStreamDataFrame{
    frame_type:string = "max_stream_data";

    stream_id:uint64;
    maximum:uint64;
}
```

A.7.11. MaxStreamsFrame

```
class MaxStreamsFrame{
    frame_type:string = "max_streams";

    stream_type:string = "bidirectional" | "unidirectional";
    maximum:uint64;
}
```

A.7.12. DataBlockedFrame

```
class DataBlockedFrame{
    frame_type:string = "data_blocked";

    limit:uint64;
}
```

A.7.13. StreamDataBlockedFrame

```
class StreamDataBlockedFrame{
    frame_type:string = "stream_data_blocked";

    stream_id:uint64;
    limit:uint64;
}
```

A.7.14. StreamsBlockedFrame

```
class StreamsBlockedFrame{
  frame_type:string = "streams_blocked";

  stream_type:string = "bidirectional" | "unidirectional";
  limit:uint64;
}
```

A.7.15. NewConnectionIDFrame

```
class NewConnectionIDFrame{
  frame_type:string = "new_connection_id";

  sequence_number:uint32;
  retire_prior_to:uint32;

  connection_id_length?:uint8;
  connection_id:bytes;

  stateless_reset_token?:Token;
}
```

A.7.16. RetireConnectionIDFrame

```
class RetireConnectionIDFrame{
  frame_type:string = "retire_connection_id";

  sequence_number:uint32;
}
```

A.7.17. PathChallengeFrame

```
class PathChallengeFrame{
  frame_type:string = "path_challenge";

  data?:bytes; // always 64-bit
}
```

A.7.18. PathResponseFrame

```
class PathResponseFrame{
  frame_type:string = "path_response";

  data?:bytes; // always 64-bit
}
```

A.7.19. ConnectionCloseFrame

raw_error_code is the actual, numerical code. This is useful because some error types are spread out over a range of codes (e.g., QUIC's crypto_error).

```
type ErrorSpace = "transport" | "application";

class ConnectionCloseFrame{
  frame_type:string = "connection_close";

  error_space?:ErrorSpace;
  error_code?:TransportError | ApplicationError | uint32;
  raw_error_code?:uint32;
  reason?:string;

  trigger_frame_type?:uint64 | string; // For known frame types, the appropriate "frame_type" string. For unknown frame types, the hex encoded identifier value
}
```

A.7.20. HandshakeDoneFrame

```
class HandshakeDoneFrame{
  frame_type:string = "handshake_done";
}
```

A.7.21. UnknownFrame

```
class UnknownFrame{
  frame_type:string = "unknown";
  raw_frame_type:uint64;

  raw_length?:uint32;
  raw?:bytes;
}
```

A.7.22. TransportError

```
enum TransportError {
    no_error,
    internal_error,
    connection_refused,
    flow_control_error,
    stream_limit_error,
    stream_state_error,
    final_size_error,
    frame_encoding_error,
    transport_parameter_error,
    connection_id_limit_error,
    protocol_violation,
    invalid_token,
    application_error,
    crypto_buffer_exceeded
}
```

A.7.23. CryptoError

These errors are defined in the TLS document as "A TLS alert is turned into a QUIC connection error by converting the one-byte alert description into a QUIC error code. The alert description is added to 0x100 to produce a QUIC error code from the range reserved for CRYPTO_ERROR."

This approach maps badly to a pre-defined enum. As such, we define the `crypto_error` string as having a dynamic component here, which should include the hex-encoded value of the TLS alert description.

```
enum CryptoError {
    crypto_error_{TLS_ALERT}
}
```

Appendix B. HTTP/3 data field definitions

B.1. HTTP/3 Frames

```
type HTTP3Frame = DataFrame | HeadersFrame | PriorityFrame | CancelPushFrame | SettingsFrame | PushPromiseFrame | GoAwayFrame | MaxPushIDFrame | DuplicatePushFrame | ReservedFrame | UnknownFrame;
```

B.1.1. DataFrame

```
class DataFrame{
    frame_type:string = "data";

    raw?:bytes;
}
```

B.1.2. HeadersFrame

This represents an `_uncompressed_`, plaintext HTTP Headers frame (e.g., no QPACK compression is applied).

For example:

```
headers: [{"name":":path","value":"/"}, {"name":":method","value":"GET"}, {"name":":authority","value":"127.0.0.1:4433"}, {"name":":scheme","value":"https"}]
```

```
class HeadersFrame{
    frame_type:string = "header";
    headers:Array<HTTPHeader>;
}
```

```
class HTTPHeader {
    name:string;
    value:string;
}
```

B.1.3. CancelPushFrame

```
class CancelPushFrame{
    frame_type:string = "cancel_push";
    push_id:uint64;
}
```

B.1.4. SettingsFrame

```
class SettingsFrame{
    frame_type:string = "settings";
    settings:Array<Setting>;
}
```

```
class Setting{
    name:string;
    value:string;
}
```

B.1.5. PushPromiseFrame

```
class PushPromiseFrame{
    frame_type:string = "push_promise";
    push_id:uint64;

    headers:Array<HTTPHeader>;
}
```

B.1.6. GoAwayFrame

```
class GoAwayFrame{
    frame_type:string = "goaway";
    stream_id:uint64;
}
```

B.1.7. MaxPushIDFrame

```
class MaxPushIDFrame{
    frame_type:string = "max_push_id";
    push_id:uint64;
}
```

B.1.8. DuplicatePushFrame

```
class DuplicatePushFrame{
    frame_type:string = "duplicate_push";
    push_id:uint64;
}
```

B.1.9. ReservedFrame

```
class ReservedFrame{
    frame_type:string = "reserved";
}
```

B.1.10. UnknownFrame

HTTP/3 re-uses QUIC's UnknownFrame definition, since their values and usage overlaps.

B.2. ApplicationError

```
enum ApplicationError{
    http_no_error,
    http_general_protocol_error,
    http_internal_error,
    http_stream_creation_error,
    http_closed_critical_stream,
    http_frame_unexpected,
    http_frame_error,
    http_excessive_load,
    http_id_error,
    http_settings_error,
    http_missing_settings,
    http_request_rejected,
    http_request_cancelled,
    http_request_incomplete,
    http_early_response,
    http_connect_error,
    http_version_fallback
}
```

Appendix C. QPACK DATA type definitions

C.1. QPACK Instructions

Note: the instructions do not have explicit encoder/decoder types, since there is no overlap between the instructions of both types in neither name nor function.

```
type QPackInstruction = SetDynamicTableCapacityInstruction | InsertWithNameReferenceInstruction | InsertWithoutNameReferenceInstruction | DuplicateInstruction | HeaderAcknowledgementInstruction | StreamCancellationInstruction | InsertCountIncrementInstruction;
```

C.1.1. SetDynamicTableCapacityInstruction

```
class SetDynamicTableCapacityInstruction {
    instruction_type:string = "set_dynamic_table_capacity";

    capacity:uint32;
}
```

C.1.2. InsertWithNameReferenceInstruction

```
class InsertWithNameReferenceInstruction {
    instruction_type:string = "insert_with_name_reference";

    table_type:"static"|"dynamic";

    name_index:uint32;

    huffman_encoded_value:boolean;

    value_length?:uint32;
    value?:string;
}
```

C.1.3. InsertWithoutNameReferenceInstruction

```
class InsertWithoutNameReferenceInstruction {
    instruction_type:string = "insert_without_name_reference";

    huffman_encoded_name:boolean;

    name_length?:uint32;
    name?:string;

    huffman_encoded_value:boolean;

    value_length?:uint32;
    value?:string;
}
```

C.1.4. DuplicateInstruction

```
class DuplicateInstruction {
    instruction_type:string = "duplicate";

    index:uint32;
}
```

C.1.5. HeaderAcknowledgementInstruction

```
class HeaderAcknowledgementInstruction {
    instruction_type:string = "header_acknowledgement";

    stream_id:uint64;
}
```

C.1.6. StreamCancellationInstruction


```
class StreamCancellationInstruction {
    instruction_type:string = "stream_cancellation";

    stream_id:uint64;
}
```

C.1.7. InsertCountIncrementInstruction

```
class InsertCountIncrementInstruction {
    instruction_type:string = "insert_count_increment";

    increment:uint32;
}
```

C.2. QPACK Header compression

```
type QPackHeaderBlockRepresentation = IndexedHeaderField | LiteralHeaderFieldWith
Name | LiteralHeaderFieldWithoutName;
```

C.2.1. IndexedHeaderField

Note: also used for "indexed header field with post-base index"

```
class IndexedHeaderField {
    header_field_type:string = "indexed_header";

    table_type:"static"|"dynamic"; // MUST be "dynamic" if is_post_base is true
    index:uint32;

    is_post_base:boolean = false; // to represent the "indexed header field with
post-base index" header field type
}
```

C.2.2. LiteralHeaderFieldWithName

Note: also used for "Literal header field with post-base name reference"

```
class LiteralHeaderFieldWithName {
    header_field_type:string = "literal_with_name";

    preserve_literal:boolean; // the 3rd "N" bit
    table_type:"static"|"dynamic"; // MUST be "dynamic" if is_post_base is true
    name_index:uint32;

    huffman_encoded_value:boolean;
    value_length?:uint32;
    value?:string;

    is_post_base:boolean = false; // to represent the "Literal header field with
    post-base name reference" header field type
}
```

C.2.3. LiteralHeaderFieldWithoutName

```
class LiteralHeaderFieldWithoutName {
    header_field_type:string = "literal_without_name";

    preserve_literal:boolean; // the 3rd "N" bit

    huffman_encoded_name:boolean;
    name_length?:uint32;
    name?:string;

    huffman_encoded_value:boolean;
    value_length?:uint32;
    value?:string;
}
```

C.2.4. QPackHeaderBlockPrefix

```
class QPackHeaderBlockPrefix {
    required_insert_count:uint32;
    sign_bit:boolean;
    delta_base:uint32;
}
```

Appendix D. Change Log

D.1. Since draft-01:

Major changes:

- * Moved data_moved from http to transport. Also made the "from" and "to" fields flexible strings instead of an enum (#111,#65)

- * Moved packet_type fields to PacketHeader. Moved packet_size field out of PacketHeader to RawInfo:length (#40)
- * Made events that need to log packet_type and packet_number use a header field instead of logging these fields individually
- * Added support for logging retry, stateless reset and initial tokens (#94,#86,#117)
- * Moved separate general event categories into a single category "generic" (#47)
- * Added "transport:connection_closed" event (#43,#85,#78,#49)
- * Added version_information and alpn_information events (#85,#75,#28)
- * Added parameters_restored events to help clarify 0-RTT behaviour (#88)

Smaller changes:

- * Merged loss_timer events into one loss_timer_updated event
- * Field data types are now strongly defined (#10,#39,#36,#115)
- * Renamed qpack instruction_received and instruction_sent to instruction_created and instruction_parsed (#114)
- * Updated qpack:dynamic_table_updated.update_type. It now has the value "inserted" instead of "added" (#113)
- * Updated qpack:dynamic_table_updated. It now has an "owner" field to differentiate encoder vs decoder state (#112)
- * Removed push_allowed from http:parameters_set (#110)
- * Removed explicit trigger field indications from events, since this was moved to be a generic property of the "data" field (#80)
- * Updated transport:connection_id_updated to be more in line with other similar events. Also dropped importance from Core to Base (#45)
- * Added length property to PaddingFrame (#34)
- * Added packet_number field to transport:frames_processed (#74)

- * Added a way to generically log packet header flags (first 8 bits) to PacketHeader
- * Added additional guidance on which events to log in which situations (#53)
- * Added "simulation:scenario" event to help indicate simulation details
- * Added "packets_acked" event (#107)
- * Added "datagram_ids" to the datagram_X and packet_X events to allow tracking of coalesced QUIC packets (#91)
- * Extended connection_state_updated with more fine-grained states (#49)

D.2. Since draft-00:

- * Event and category names are now all lowercase
- * Added many new events and their definitions
- * "type" fields have been made more specific (especially important for PacketType fields, which are now called packet_type instead of type)
- * Events are given an importance indicator (issue #22)
- * Event names are more consistent and use past tense (issue #21)
- * Triggers have been redefined as properties of the "data" field and updated for most events (issue #23)

Appendix E. Design Variations

TBD

Appendix F. Acknowledgements

Thanks to Marten Seemann, Jana Iyengar, Brian Trammell, Dmitri Tikhonov, Stephen Petrides, Jari Arkko, Marcus Ihlar, Victor Vasiliev, Mirja Kuehlewind, Jeremy Laine, Kazu Yamamoto, Christian Huitema, and Lucas Pardue for their feedback and suggestions.

Author's Address

Robin Marx
Hasselt University

Email: robin.marx@uhasselt.be

QUIC
Internet-Draft
Intended status: Standards Track
Expires: November 16, 2021

R. Marx, Ed.
KU Leuven
L. Niccolini, Ed.
Facebook
M. Seemann, Ed.
Protocol Labs
May 15, 2021

Main logging schema for qlog
draft-marx-qlog-main-schema-03

Abstract

This document describes a high-level schema for a standardized logging format called qlog. This format allows easy sharing of data and the creation of reusable visualization and debugging tools. The high-level schema in this document is intended to be protocol-agnostic. Separate documents specify how the format should be used for specific protocol data. The schema is also format-agnostic, and can be represented in for example JSON, csv or protobuf.

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 16, 2021.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Notational Conventions	4
2. Design goals	5
3. The high level qlog schema	6
3.1. summary	7
3.2. traces	8
3.3. Individual Trace containers	9
3.3.1. configuration	10
3.3.2. vantage_point	12
3.4. Field name semantics	14
3.4.1. timestamps	15
3.4.2. category and event	17
3.4.3. data	18
3.4.4. protocol_type	18
3.4.5. triggers	19
3.4.6. group_id	19
3.4.7. common_fields	21
4. Guidelines for event definition documents	23
4.1. Event design guidelines	23
4.2. Event importance indicators	24
4.3. Custom fields	25
5. Generic events and data classes	25
5.1. Raw packet and frame information	26
5.2. Generic events	27
5.2.1. error	27
5.2.2. warning	27
5.2.3. info	27
5.2.4. debug	28
5.2.5. verbose	28
5.3. Simulation events	28
5.3.1. scenario	29
5.3.2. marker	29
6. Serializing qlog	29
6.1. qlog to JSON mapping	30
6.1.1. numbers	30
6.1.2. bytes	31
6.1.3. Summarizing table	32
6.1.4. Other JSON specifics	33
6.2. qlog to NDJSON mapping	33
6.2.1. Supporting NDJSON in tooling	35

6.3.	Other optimized formatting options	35
6.3.1.	Data structure optimizations	36
6.3.2.	Compression	37
6.3.3.	Binary formats	37
6.3.4.	Overview and summary	38
6.4.	Conversion between formats	39
7.	Methods of access and generation	40
7.1.	Set file output destination via an environment variable .	40
7.2.	Access logs via a well-known endpoint	41
8.	Tooling requirements	42
9.	Security and privacy considerations	42
10.	IANA Considerations	43
11.	References	43
11.1.	Normative References	43
11.2.	Informative References	43
11.3.	URIs	44
Appendix A.	Change Log	45
A.1.	Since draft-marx-qlog-main-schema-draft-02:	45
A.2.	Since draft-marx-qlog-main-schema-01:	45
A.3.	Since draft-marx-qlog-main-schema-00:	46
Appendix B.	Design Variations	46
Appendix C.	Acknowledgements	46
Authors' Addresses	46

1. Introduction

There is currently a lack of an easily usable, standardized endpoint logging format. Especially for the use case of debugging and evaluating modern Web protocols and their performance, it is often difficult to obtain structured logs that provide adequate information for tasks like problem root cause analysis.

This document aims to provide a high-level schema and harness that describes the general layout of an easily usable, shareable, aggregatable and structured logging format. This high-level schema is protocol agnostic, with logging entries for specific protocols and use cases being defined in other documents (see for example [QLOG-QUIC] for QUIC and [QLOG-H3] for HTTP/3 and QPACK-related event definitions).

The goal of this high-level schema is to provide amenities and default characteristics that each logging file should contain (or should be able to contain), such that generic and reusable toolsets can be created that can deal with logs from a variety of different protocols and use cases.

As such, this document contains concepts such as versioning, metadata inclusion, log aggregation, event grouping and log file size reduction techniques.

Feedback and discussion are welcome at <https://github.com/quiclog/internet-drafts> [1]. Readers are advised to refer to the "editor's draft" at that URL for an up-to-date version of this document.

1.1. Notational 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].

While the qlog schema's are format-agnostic, for readability the qlog documents will use a JSON-inspired format ([RFC8259]) for examples and definitions.

As qlog can be serialized both textually but also in binary, we employ a custom datatype definition language, inspired loosely by the "TypeScript" language [2].

This document describes how to employ JSON and NDJSON as textual serializations for qlog in Section 6. Other documents will describe how to utilize other concrete serialization options, though tips and requirements for these are also listed in this document (Section 6).

The main general conventions in this document a reader should be aware of are:

- o `obj?` : this object is optional
- o `type1 | type2` : a union of these two types (object can be either `type1` OR `type2`)
- o `obj:type` : this object has this concrete type
- o `obj:array<type>` : this object is an array of this type
- o `class` : defines a new type
- o `//` : single-line comment

The main data types are:

- o `int8` : signed 8-bit integer
- o `int16` : signed 16-bit integer

- o int32 : signed 32-bit integer
- o int64 : signed 64-bit integer
- o uint8 : unsigned 8-bit integer
- o uint16 : unsigned 16-bit integer
- o uint32 : unsigned 32-bit integer
- o uint64 : unsigned 64-bit integer
- o float : 32-bit floating point value
- o double : 64-bit floating point value
- o byte : an individual raw byte (8-bit) value (use array<byte> or the shorthand "bytes" to specify a binary blob)
- o string : list of Unicode (typically UTF-8) encoded characters
- o boolean : boolean
- o enum: fixed list of values (Unless explicitly defined, the value of an enum entry is the string version of its name (e.g., initial = "initial"))
- o any : represents any object type. Mainly used here as a placeholder for more concrete types defined in related documents (e.g., specific event types)

All timestamps and time-related values (e.g., offsets) in qlog are logged as doubles in the millisecond resolution.

Other qlog documents can define their own data types (e.g., separately for each Packet type that a protocol supports).

2. Design goals

The main tenets for the qlog schema design are:

- o Streamable, event-based logging
- o Flexibility in the format, complexity in the tooling (e.g., few components are a MUST, tools need to deal with this)
- o Extensible and pragmatic (e.g., no complex fixed schema with extension points)

- o Aggregation and transformation friendly (e.g., the top-level element is a container for individual traces, `group_id` can be used to tag events to a particular context)
- o Metadata is stored together with event data

3. The high level qlog schema

A qlog file should be able to contain several individual traces and logs from multiple vantage points that are in some way related. To that end, the top-level element in the qlog schema defines only a small set of "header" fields and an array of component traces. For this document, the required "qlog_version" field MUST have a value of "qlog-03-WIP".

Note: there have been several previously broadly deployed qlog versions based on older drafts of this document (see draft-marx-qlog-main-schema). The old values for the "qlog_version" field were "draft-00", "draft-01" and "draft-02". When qlog was moved to the QUIC working group, we decided to increment the existing counter, rather than reverting back to -00. As such, any numbering indicating in the "qlog_version" field is explicitly not tied to a particular version of the draft documents.

As qlog can be serialized in a variety of ways, the "qlog_format" field is used to indicate which serialization option was chosen. Its value MUST either be one of the options defined in this document (e.g., Section 6) or the field must be omitted entirely, in which case it assumes the default value of "JSON".

In order to make it easier to parse and identify qlog files and their serialization format, the "qlog_version" and "qlog_format" fields and their values SHOULD be in the first 256 characters/bytes of the resulting log file.

An example of the qlog file's top-level structure is shown in Figure 1.

Definition:

```
class QlogFile {
  qlog_version:string,
  qlog_format?:string,
  title?:string,
  description?:string,
  summary?: Summary,
  traces: array<Trace|TraceError>
}
```

JSON serialization:

```
{
  "qlog_version": "draft-03-WIP",
  "qlog_format": "JSON",
  "title": "Name of this particular qlog file (short)",
  "description": "Description for this group of traces (long)",
  "summary": {
    ...
  },
  "traces": [...]
}
```

Figure 1: Top-level element

3.1. summary

In a real-life deployment with a large amount of generated logs, it can be useful to sort and filter logs based on some basic summarized or aggregated data (e.g., log length, packet loss rate, log location, presence of error events, ...). The summary field (if present) SHOULD be on top of the qlog file, as this allows for the file to be processed in a streaming fashion (i.e., the implementation could just read up to and including the summary field and then only load the full logs that are deemed interesting by the user).

As the summary field is highly deployment-specific, this document does not specify any default fields or their semantics. Some examples of potential entries are shown in Figure 2.

Definition (purely illustrative example):

```
class Summary {  
    "trace_count":uint32, // amount of traces in this file  
    "max_duration":uint64, // time duration of the longest trace in ms  
    "max_outgoing_loss_rate":float, // highest loss rate for outgoing packets over  
    // all traces  
    "total_event_count":uint64, // total number of events across all traces,  
    "error_count":uint64 // total number of error events in this trace  
}
```

JSON serialization:

```
{  
    "trace_count": 1,  
    "max_duration": 5006,  
    "max_outgoing_loss_rate": 0.013,  
    "total_event_count": 568,  
    "error_count": 2  
}
```

Figure 2: Summary example definition

3.2. traces

It is often advantageous to group several related qlog traces together in a single file. For example, we can simultaneously perform logging on the client, on the server and on a single point on their common network path. For analysis, it is useful to aggregate these three individual traces together into a single file, so it can be uniquely stored, transferred and annotated.

As such, the "traces" array contains a list of individual qlog traces. Typical qlogs will only contain a single trace in this array. These can later be combined into a single qlog file by taking the "traces" entry/entries for each qlog file individually and copying them to the "traces" array of a new, aggregated qlog file. This is typically done in a post-processing step.

The "traces" array can thus contain both normal traces (for the definition of the Trace type, see Section 3.3), but also "error" entries. These indicate that we tried to find/convert a file for inclusion in the aggregated qlog, but there was an error during the process. Rather than silently dropping the erroneous file, we can opt to explicitly include it in the qlog file as an entry in the "traces" array, as shown in Figure 3.

Definition:

```
class TraceError {  
    error_description: string, // A description of the error  
    uri?: string, // the original URI at which we attempted to find the file  
    vantage_point?: VantagePoint // see {{vantage_point}}: the vantage point we w  
ere expecting to include here  
}
```

JSON serialization:

```
{  
  "error_description": "File could not be found",  
  "uri": "/srv/traces/today/latest.qlog",  
  "vantage_point": { type: "server" }  
}
```

Figure 3: TraceError definition

Note that another way to combine events of different traces in a single qlog file is through the use of the "group_id" field, discussed in Section 3.4.6.

3.3. Individual Trace containers

The exact conceptual definition of a Trace can be fluid. For example, a trace could contain all events for a single connection, for a single endpoint, for a single measurement interval, for a single protocol, etc. As such, a Trace container contains some metadata in addition to the logged events, see Figure 4.

In the normal use case however, a trace is a log of a single data flow collected at a single location or vantage point. For example, for QUIC, a single trace only contains events for a single logical QUIC connection for either the client or the server.

The semantics and context of the trace can mainly be deduced from the entries in the "common_fields" list and "vantage_point" field.

Definition:

```
class Trace {  
    title?: string,  
    description?: string,  
    configuration?: Configuration,  
    common_fields?: CommonFields,  
    vantage_point: VantagePoint,  
    events: array<Event>  
}
```

JSON serialization:

```
{  
  "title": "Name of this particular trace (short)",  
  "description": "Description for this trace (long)",  
  "configuration": {  
    "time_offset": 150  
  },  
  "common_fields": {  
    "ODCID": "abcde1234",  
    "time_format": "absolute"  
  },  
  "vantage_point": {  
    "name": "backend-67",  
    "type": "server"  
  },  
  "events": [...]  
}
```

Figure 4: Trace container definition

3.3.1. configuration

We take into account that a qlog file is usually not used in isolation, but by means of various tools. Especially when aggregating various traces together or preparing traces for a demonstration, one might wish to persist certain tool-based settings inside the qlog file itself. For this, the configuration field is used.

The configuration field can be viewed as a generic metadata field that tools can fill with their own fields, based on per-tool logic. It is best practice for tools to prefix each added field with their tool name to prevent collisions across tools. This document only defines two optional, standard, tool-independent configuration settings: "time_offset" and "original_uris".

Definition:

```
class Configuration {  
    time_offset:double, // in ms,  
    original_uris: array<string>,  
  
    // list of fields with any type  
}
```

JSON serialization:

```
{  
  "time_offset": 150, // starts 150ms after the first timestamp indicates  
  "original_uris": [  
    "https://example.org/trace1.qlog",  
    "https://example.org/trace2.qlog"  
  ]  
}
```

Figure 5: Configuration definition

3.3.1.1. time_offset

The `time_offset` field indicates by how many milliseconds the starting time of the current trace should be offset. This is useful when comparing logs taken from various systems, where clocks might not be perfectly synchronous. Users could use manual tools or automated logic to align traces in time and the found optimal offsets can be stored in this field for future usage. The default value is 0.

3.3.1.2. original_uris

The `original_uris` field is used when merging multiple individual qlog files or other source files (e.g., when converting .pcaps to qlog). It allows to keep better track where certain data came from. It is a simple array of strings. It is an array instead of a single string, since a single qlog trace can be made up out of an aggregation of multiple component qlog traces as well. The default value is an empty array.

3.3.1.3. custom fields

Tools can add optional custom metadata to the "configuration" field to store state and make it easier to share specific data viewpoints and view configurations.

Two examples from the qvis toolset [3] are shown in Figure 6.


```

{
  "configuration" : {
    "qvis" : {
      // when loaded into the qvis toolsuite's congestion graph tool
      // zoom in on the period between 1s and 2s and select the 124th event
      defined in this trace
      "congestion_graph": {
        "startX": 1000,
        "endX": 2000,
        "focusOnEventIndex": 124
      }

      // when loaded into the qvis toolsuite's sequence diagram tool
      // automatically scroll down the timeline to the 555th event defined
      in this trace
      "sequence_diagram" : {
        "focusOnEventIndex": 555
      }
    }
  }
}

```

Figure 6: Custom configuration fields example

3.3.2. vantage_point

The `vantage_point` field describes the vantage point from which the trace originates, see Figure 7. Each trace can have only a single `vantage_point` and thus all events in a trace MUST BE from the perspective of this `vantage_point`. To include events from multiple `vantage_points`, implementers can for example include multiple traces, split by `vantage_point`, in a single qlog file.

Definition:

```
class VantagePoint {
  name?: string,
  type: VantagePointType,
  flow?: VantagePointType
}

class VantagePointType {
  server, // endpoint which initiates the connection.
  client, // endpoint which accepts the connection.
  network, // observer in between client and server.
  unknown
}
```

JSON serialization examples:

```
{
  "name": "aioquic client",
  "type": "client",
}

{
  "name": "wireshark trace",
  "type": "network",
  "flow": "client"
}
```

Figure 7: VantagePoint definition

The flow field is only required if the type is "network" (for example, the trace is generated from a packet capture). It is used to disambiguate events like "packet sent" and "packet received". This is indicated explicitly because for multiple reasons (e.g., privacy) data from which the flow direction can be otherwise inferred (e.g., IP addresses) might not be present in the logs.

Meaning of the different values for the flow field: * "client" indicates that this vantage point follows client data flow semantics (a "packet sent" event goes in the direction of the server). * "server" indicates that this vantage point follow server data flow semantics (a "packet sent" event goes in the direction of the client). * "unknown" indicates that the flow's direction is unknown.

Depending on the context, tools confronted with "unknown" values in the vantage_point can either try to heuristically infer the semantics from protocol-level domain knowledge (e.g., in QUIC, the client

always sends the first packet) or give the user the option to switch between client and server perspectives manually.

3.4. Field name semantics

Inside of the "events" field of a qlog trace is a list of events logged by the endpoint. Each event is specified as a generic object with a number of member fields and their associated data. Depending on the protocol and use case, the exact member field names and their formats can differ across implementations. This section lists the main, pre-defined and reserved field names with specific semantics and expected corresponding value formats.

Each qlog event at minimum requires the "time" (Section 3.4.1), "name" (Section 3.4.2) and "data" (Section 3.4.3) fields. Other typical fields are "time_format" (Section 3.4.1), "protocol_type" (Section 3.4.4), "trigger" (Section 3.4.5), and "group_id" (Section 3.4.6). As especially these later fields typically have identical values across individual event instances, they are normally logged separately in the "common_fields" (Section 3.4.7).

The specific values for each of these fields and their semantics are defined in separate documents, specific per protocol or use case. For example: event definitions for QUIC, HTTP/3 and QPACK can be found in [QLOG-QUIC] and [QLOG-H3].

Other fields are explicitly allowed by the qlog approach, and tools SHOULD allow for the presence of unknown event fields, but their semantics depend on the context of the log usage (e.g., for QUIC, the ODCID field is used), see [QLOG-QUIC].

An example of a qlog event with its component fields is shown in Figure 8.

Definition:

```
class Event {
  time: double,
  name: string,
  data: any,

  protocol_type?: Array<string>,
  group_id?: string|uint32,

  time_format?: "absolute"|"delta"|"relative",

  // list of fields with any type
}
```

JSON serialization:

```
{
  time: 1553986553572,

  name: "transport:packet_sent",
  data: { ... }

  protocol_type: ["QUIC","HTTP3"],
  group_id: "127ecc830d98f9d54a42c4f0842aa87e181a",

  time_format: "absolute",

  ODCID: "127ecc830d98f9d54a42c4f0842aa87e181a", // QUIC specific
}
```

Figure 8: Event fields definition

3.4.1. timestamps

The "time" field indicates the timestamp at which the event occurred. Its value is typically the Unix timestamp since the 1970 epoch (number of milliseconds since midnight UTC, January 1, 1970, ignoring leap seconds). However, qlog supports two more succinct timestamps formats to allow reducing file size. The employed format is indicated in the "time_format" field, which allows one of three values: "absolute", "delta" or "relative":

- o Absolute: Include the full absolute timestamp with each event. This approach uses the largest amount of characters. This is also the default value of the "time_format" field.

- o **Delta:** Delta-encode each time value on the previously logged value. The first event in a trace typically logs the full absolute timestamp. This approach uses the least amount of characters.
- o **Relative:** Specify a full "reference_time" timestamp (typically this is done up-front in "common_fields", see Section 3.4.7) and include only relatively-encoded values based on this reference_time with each event. The "reference_time" value is typically the first absolute timestamp. This approach uses a medium amount of characters.

The first option is good for stateless loggers, the second and third for stateful loggers. The third option is generally preferred, since it produces smaller files while being easier to reason about. An example for each option can be seen in Figure 9.

The absolute approach will use:
1500, 1505, 1522, 1588

The delta approach will use:
1500, 5, 17, 66

The relative approach will:
- set the reference_time to 1500 in "common_fields"
- use: 0, 5, 22, 88

Figure 9: Three different approaches for logging timestamps

One of these options is typically chosen for the entire trace (put differently: each event has the same value for the "time_format" field). Each event MUST include a timestamp in the "time" field.

Events in each individual trace SHOULD be logged in strictly ascending timestamp order (though not necessarily absolute value, for the "delta" format). Tools CAN sort all events on the timestamp before processing them, though are not required to (as this could impose a significant processing overhead). This can be a problem especially for multi-threaded and/or streaming loggers, who could consider using a separate postprocessor to order qlog events in time if a tool do not provide this feature.

Timestamps do not have to use the UNIX epoch timestamp as their reference. For example for privacy considerations, any initial reference timestamps (for example "endpoint uptime in ms" or "time since connection start in ms") can be chosen. Tools SHOULD NOT assume the ability to derive the absolute Unix timestamp from qlog

traces, nor allow on them to relatively order events across two or more separate traces (in this case, clock drift should also be taken into account).

3.4.2. category and event

Events differ mainly in the type of metadata associated with them. To help identify a given event and how to interpret its metadata in the "data" field (see Section 3.4.3), each event has an associated "name" field. This can be considered as a concatenation of two other fields, namely event "category" and event "type".

Category allows a higher-level grouping of events per specific event type. For example for QUIC and HTTP/3, the different categories could be "transport", "http", "qpack", and "recovery". Within these categories, the event Type provides additional granularity. For example for QUIC and HTTP/3, within the "transport" Category, there would be "packet_sent" and "packet_received" events.

Logging category and type separately conceptually allows for fast and high-level filtering based on category and the re-use of event types across categories. However, it also considerably inflates the log size and this flexibility is not used extensively in practice at the time of writing.

As such, the default approach in qlog is to concatenate both field values using the ":" character in the "name" field, as can be seen in Figure 10. As such, qlog category and type names MUST NOT include this character.

JSON serialization using separate fields:

```
{
  category: "transport",
  type: "packet_sent"
}
```

JSON serialization using ":" concatenated field:

```
{
  name: "transport:packet_sent"
}
```

Figure 10: Ways of logging category, type and name of an event.

Certain serializations CAN emit category and type as separate fields, and qlog tools SHOULD be able to deal with both the concatenated "name" field, and the separate "category" and "type" fields. Text-based serializations however are encouraged to employ the concatenated "name" field for efficiency.

3.4.3. data

The data field is a generic object. It contains the per-event metadata and its form and semantics are defined per specific sort of event. For example, data field value definitions for QUIC and HTTP/3 can be found in [QLOG-QUIC] and [QLOG-H3].

One purely illustrative example for a QUIC "packet_sent" event is shown in Figure 11.

Definition:

```
class TransportPacketSentEvent {  
    packet_size?:uint32,  
    header:PacketHeader,  
    frames?:Array<QuicFrame>  
}
```

JSON serialization:

```
{  
  packet_size: 1280,  
  header: {  
    packet_type: "1RTT",  
    packet_number: 123  
  },  
  frames: [  
    {  
      frame_type: "stream",  
      length: 1000,  
      offset: 456  
    },  
    {  
      frame_type: "padding"  
    }  
  ]  
}
```

Figure 11: Example of the 'data' field for a QUIC packet_sent event

3.4.4. protocol_type

The "protocol_type" array field indicates to which protocols (or protocol "stacks") this event belongs. This allows a single qlog file to aggregate traces of different protocols (e.g., a web server offering both TCP+HTTP/2 and QUIC+HTTP/3 connections).

For example, QUIC and HTTP/3 events have the "QUIC" and "HTTP3" protocol_type entry values, see [QLOG-QUIC] and [QLOG-H3].

Typically however, all events in a single trace are of the same few protocols, and this array field is logged once in "common_fields", see Section 3.4.7.

3.4.5. triggers

Sometimes, additional information is needed in the case where a single event can be caused by a variety of other events. In the normal case, the context of the surrounding log messages gives a hint as to which of these other events was the cause. However, in highly-parallel and optimized implementations, corresponding log messages might be separated in time. Another option is to explicitly indicate these "triggers" in a high-level way per-event to get more fine-grained information without much additional overhead.

In qlog, the optional "trigger" field contains a string value describing the reason (if any) for this event instance occurring. While this "trigger" field could be a property of the qlog Event itself, it is instead a property of the "data" field instead. This choice was made because many event types do not include a trigger value, and having the field at the Event-level would cause overhead in some serializations. Additional information on the trigger can be added in the form of additional member fields of the "data" field value, yet this is highly implementation-specific, as are the trigger field's string values.

One purely illustrative example of some potential triggers for QUIC's "packet_dropped" event is shown in Figure 12.

Definition:

```
class QuicPacketDroppedEvent {
    packet_type?:PacketType,
    raw_length?:uint32,

    trigger?: "key_unavailable" | "unknown_connection_id" | "decrypt_error" | "un
supported_version"
}
```

Figure 12: Trigger example

3.4.6. group_id

As discussed in Section 3.3, a single qlog file can contain several traces taken from different vantage points. However, a single trace from one endpoint can also contain events from a variety of sources.

For example, a server implementation might choose to log events for all incoming connections in a single large (streamed) qlog file. As such, we need a method for splitting up events belonging to separate logical entities.

The simplest way to perform this splitting is by associating a "group identifier" to each event that indicates to which conceptual "group" each event belongs. A post-processing step can then extract events per group. However, this group identifier can be highly protocol and context-specific. In the example above, we might use QUIC's "Original Destination Connection ID" to uniquely identify a connection. As such, they might add a "ODCID" field to each event. However, a middlebox logging IP or TCP traffic might rather use four-tuples to identify connections, and add a "four_tuple" field.

As such, to provide consistency and ease of tooling in cross-protocol and cross-context setups, qlog instead defines the common "group_id" field, which contains a string value. Implementations are free to use their preferred string serialization for this field, so long as it contains a unique value per logical group. Some examples can be seen in Figure 13.

JSON serialization for events grouped by four tuples and QUIC connection IDs:

```
events: [
  {
    time: 1553986553579,
    protocol_type: ["TCP", "TLS", "HTTP2"],
    group_id: "ip1=2001:67c:1232:144:9498:6df6:f450:110b,ip2=2001:67c:2b0:1c1
::198,port1=59105,port2=80",
    name: "transport:packet_received",
    data: { ... },
  },
  {
    time: 1553986553581,
    protocol_type: ["QUIC", "HTTP3"],
    group_id: "127ecc830d98f9d54a42c4f0842aa87e181a",
    name: "transport:packet_sent",
    data: { ... },
  }
]
```

Figure 13: Example of group_id usage

Note that in some contexts (for example a Multipath transport protocol) it might make sense to add additional contextual per-event fields (for example "path_id"), rather than use the group_id field for that purpose.

Note also that, typically, a single trace only contains events belonging to a single logical group (for example, an individual QUIC connection). As such, instead of logging the "group_id" field with an identical value for each event instance, this field is typically logged once in "common_fields", see Section 3.4.7.

3.4.7. common_fields

As discussed in the previous sections, information for a typical qlog event varies in three main fields: "time", "name" and associated data. Additionally, there are also several more advanced fields that allow mixing events from different protocols and contexts inside of the same trace (for example "protocol_type" and "group_id"). In most "normal" use cases however, the values of these advanced fields are consistent for each event instance (for example, a single trace contains events for a single QUIC connection).

To reduce file size and making logging easier, qlog uses the "common_fields" list to indicate those fields and their values that are shared by all events in this component trace. This prevents these fields from being logged for each individual event. An example of this is shown in Figure 14.

JSON serialization with repeated field values per-event instance:

```
{
  events: [{
    group_id: "127ecc830d98f9d54a42c4f0842aa87e181a",
    protocol_type: ["QUIC", "HTTP3"],
    time_format: "relative",
    reference_time: "1553986553572",

    time: 2,
    name: "transport:packet_received",
    data: { ... }
  }, {
    group_id: "127ecc830d98f9d54a42c4f0842aa87e181a",
    protocol_type: ["QUIC", "HTTP3"],
    time_format: "relative",
    reference_time: "1553986553572",

    time: 7,
    name: "http:frame_parsed",
    data: { ... }
  }
]
```

JSON serialization with repeated field values extracted to common_fields:

```
{
  common_fields: {
    group_id: "127ecc830d98f9d54a42c4f0842aa87e181a",
    protocol_type: ["QUIC", "HTTP3"],
    time_format: "relative",
    reference_time: "1553986553572"
  },
  events: [
    {
      time: 2,
      name: "transport:packet_received",
      data: { ... }
    }, {
      7,
      name: "http:frame_parsed",
      data: { ... }
    }
  ]
}
```

Figure 14: Example of common_fields usage

The "common_fields" field is a generic dictionary of key-value pairs, where the key is always a string and the value can be of any type, but is typically also a string or number. As such, unknown entries in this dictionary MUST be disregarded by the user and tools (i.e., the presence of an unknown field is explicitly NOT an error).

The list of default qlog fields that are typically logged in common_fields (as opposed to as individual fields per event instance) are:

- o time_format
- o reference_time
- o protocol_type
- o group_id

Tools MUST be able to deal with these fields being defined either on each event individually or combined in common_fields. Note that if at least one event in a trace has a different value for a given field, this field MUST NOT be added to common_fields but instead defined on each event individually. Good example of such fields are "time" and "data", who are divergent by nature.

4. Guidelines for event definition documents

This document only defines the main schema for the qlog format. This is intended to be used together with specific, per-protocol event definitions that specify the name (category + type) and data needed for each individual event. This is with the intent to allow the qlog main schema to be easily re-used for several protocols. Examples include the QUIC event definitions [QLOG-QUIC] and HTTP/3 and QPACK event definitions [QLOG-H3].

This section defines some basic annotations and concepts the creators of event definition documents SHOULD follow to ensure a measure of consistency, making it easier for qlog implementers to extrapolate from one protocol to another.

4.1. Event design guidelines

TODO: pending QUIC working group discussion. This text reflects the initial (qlog draft 01 and 02) setup.

There are several ways of defining qlog events. In practice, we have seen two main types used so far: a) those that map directly to concepts seen in the protocols (e.g., "packet_sent") and b) those

that act as aggregating events that combine data from several possible protocol behaviours or code paths into one (e.g., "parameters_set"). The latter are typically used as a means to reduce the amount of unique event definitions, as reflecting each possible protocol event as a separate qlog entity would cause an explosion of event types.

Additionally, logging duplicate data is typically prevented as much as possible. For example, packet header values that remain consistent across many packets are split into separate events (for example "spin_bit_updated" or "connection_id_updated" for QUIC).

Finally, we have typically refrained from adding additional state change events if those state changes can be directly inferred from data on the wire (for example flow control limit changes) if the implementation is bug-free and spec-compliant. Exceptions have been made for common events that benefit from being easily identifiable or individually logged (for example "packets_acked").

4.2. Event importance indicators

Depending on how events are designed, it may be that several events allow the logging of similar or overlapping data. For example the separate QUIC "connection_started" event overlaps with the more generic "connection_state_updated". In these cases, it is not always clear which event should be logged or used, and which event should take precedence if e.g., both are present and provide conflicting information.

To aid in this decision making, we recommend that each event SHOULD have an "importance indicator" with one of three values, in decreasing order of importance and expected usage:

- o Core
- o Base
- o Extra

The "Core" events are the events that SHOULD be present in all qlog files for a given protocol. These are typically tied to basic packet and frame parsing and creation, as well as listing basic internal metrics. Tool implementers SHOULD expect and add support for these events, though SHOULD NOT expect all Core events to be present in each qlog trace.

The "Base" events add additional debugging options and CAN be present in qlog files. Most of these can be implicitly inferred from data in

Core events (if those contain all their properties), but for many it is better to log the events explicitly as well, making it clearer how the implementation behaves. These events are for example tied to passing data around in buffers, to how internal state machines change and help show when decisions are actually made based on received data. Tool implementers SHOULD at least add support for showing the contents of these events, if they do not handle them explicitly.

The "Extra" events are considered mostly useful for low-level debugging of the implementation, rather than the protocol. They allow more fine-grained tracking of internal behaviour. As such, they CAN be present in qlog files and tool implementers CAN add support for these, but they are not required to.

Note that in some cases, implementers might not want to log for example data content details in the "Core" events due to performance or privacy considerations. In this case, they SHOULD use (a subset of) relevant "Base" events instead to ensure usability of the qlog output. As an example, implementations that do not log QUIC "packet_received" events and thus also not which (if any) ACK frames the packet contains, SHOULD log "packets_acked" events instead.

Finally, for event types whose data (partially) overlap with other event types' definitions, where necessary the event definition document should include explicit guidance on which to use in specific situations.

4.3. Custom fields

Event definition documents are free to define new category and event types, top-level fields (e.g., a per-event field indicating its privacy properties or path_id in multipath protocols), as well as values for the "trigger" property within the "data" field, or other member fields of the "data" field, as they see fit.

They however SHOULD NOT expect non-specialized tools to recognize or visualize this custom data. However, tools SHOULD make an effort to visualize even unknown data if possible in the specific tool's context. If they do not, they MUST ignore these unknown fields.

5. Generic events and data classes

There are some event types and data classes that are common across protocols, applications and use cases that benefit from being defined in a single location. This section specifies such common definitions.

5.1. Raw packet and frame information

While qlog is a more high-level logging format, it also allows the inclusion of most raw wire image information, such as byte lengths and even raw byte values. This can be useful when for example investigating or tuning packetization behaviour or determining encoding/framing overheads. However, these fields are not always necessary and can take up considerable space if logged for each packet or frame. They can also have a considerable privacy and security impact. As such, they are grouped in a separate optional field called "raw" of type RawInfo (where applicable).

```
class RawInfo {  
    length?:uint64; // the full byte length of the entity (e.g., packet or frame)  
    including headers and trailers  
    payload_length?:uint64; // the byte length of the entity's payload, without h  
    eaders or trailers  
  
    data?:bytes; // the contents of the full entity, including headers and traile  
rs  
}
```

Note: The RawInfo:data field can be truncated for privacy or security purposes (for example excluding payload data). In this case, the length properties should still indicate the non-truncated lengths.

Note: We do not specify explicit header_length or trailer_length fields. In most protocols, header_length can be calculated by subtracing the payload_length from the length (e.g., if trailer_length is always 0). In protocols with trailers (e.g., QUIC's AEAD tag), event definitions documents SHOULD define other ways of logging the trailer_length to make the header_length calculation possible.

The exact definitions entities, headers, trailers and payloads depend on the protocol used. If this is non-trivial, event definitions documents SHOULD include a clear explanation of how entities are mapped into the RawInfo structure.

Note: Relatedly, many modern protocols use Variable-Length Integer Encoded (VLIE) values in their headers, which are of a dynamic length. Because of this, we cannot deterministically reconstruct the header encoding/length from non-RawInfo qlog data, as implementations might not necessarily employ the most efficient VLIE scheme for all values. As such, to make exact size-analysis possible, implementers should use explicit lengths in RawInfo rather than reconstructing them from other qlog data. Similarly, tool developers should only utilize RawInfo (and related information) in such tools to prevent errors.

5.2. Generic events

In typical logging setups, users utilize a discrete number of well-defined logging categories, levels or severities to log freeform (string) data. This generic events category replicates this approach to allow implementations to fully replace their existing text-based logging by qlog. This is done by providing events to log generic strings for the typical well-known logging levels (error, warning, info, debug, verbose).

For the events defined below, the "category" is "generic" and their "type" is the name of the heading in lowercase (e.g., the "name" of the error event is "generic:error").

5.2.1. error

Importance: Core

Used to log details of an internal error that might not get reflected on the wire.

Data:

```
{
  code?:uint32,
  message?:string
}
```

5.2.2. warning

Importance: Base

Used to log details of an internal warning that might not get reflected on the wire.

Data:

```
{
  code?:uint32,
  message?:string
}
```

5.2.3. info

Importance: Extra

Used mainly for implementations that want to use qlog as their one and only logging format but still want to support unstructured string messages.

Data:

```
{  
    message:string  
}
```

5.2.4. debug

Importance: Extra

Used mainly for implementations that want to use qlog as their one and only logging format but still want to support unstructured string messages.

Data:

```
{  
    message:string  
}
```

5.2.5. verbose

Importance: Extra

Used mainly for implementations that want to use qlog as their one and only logging format but still want to support unstructured string messages.

Data:

```
{  
    message:string  
}
```

5.3. Simulation events

When evaluating a protocol implementation, one typically sets up a series of interoperability or benchmarking tests, in which the test situations can change over time. For example, the network bandwidth or latency can vary during the test, or the network can be fully disable for a short time. In these setups, it is useful to know when exactly these conditions are triggered, to allow for proper correlation with other events.

For the events defined below, the "category" is "simulation" and their "type" is the name of the heading in lowercase (e.g., the "name" of the scenario event is "simulation:scenario").

5.3.1. scenario

Importance: Extra

Used to specify which specific scenario is being tested at this particular instance. This could also be reflected in the top-level qlog's "summary" or "configuration" fields, but having a separate event allows easier aggregation of several simulations into one trace (e.g., split by "group_id").

```
{
  name?:string,
  details?:any
}
```

5.3.2. marker

Importance: Extra

Used to indicate when specific emulation conditions are triggered at set times (e.g., at 3 seconds in 2% packet loss is introduced, at 10s a NAT rebind is triggered).

```
{
  type?:string,
  message?:string
}
```

6. Serializing qlog

This document and other related qlog schema definitions are intentionally serialization-format agnostic. This means that implementers themselves can choose how to represent and serialize qlog data practically on disk or on the wire. Some examples of possible formats are JSON, CBOR, CSV, protocol buffers, flatbuffers, etc.

All these formats make certain tradeoffs between flexibility and efficiency, with textual formats like JSON typically being more flexible but also less efficient than binary formats like protocol buffers. The format choice will depend on the practical use case of the qlog user. For example, for use in day to day debugging, a plaintext readable (yet relatively large) format like JSON is probably preferred. However, for use in production, a more optimized

yet restricted format can be better. In this latter case, it will be more difficult to achieve interoperability between qlog implementations of various protocol stacks, as some custom or tweaked events from one might not be compatible with the format of the other. This will also reflect in tooling: not all tools will support all formats.

This being said, the authors prefer JSON as the basis for storing qlog, as it retains full flexibility and maximum interoperability. Storage overhead can be managed well in practice by employing compression. For this reason, this document details both how to practically transform qlog schema definitions to JSON and to the streamable NDJSON. We discuss concrete options to bring down JSON size and processing overheads in Section 6.3.

As depending on the employed format different deserializers/parsers should be used, the "qlog_format" field is used to indicate the chosen serialization approach. This field is always a string, but can be made hierarchical by the use of the "." separator between entries. For example, a value of "JSON.optimizationA" can indicate that a default JSON format is being used, but that a certain optimization of type A was applied to the file as well (see also Section 6.3).

6.1. qlog to JSON mapping

When mapping qlog to normal JSON, the "qlog_format" field MUST have the value "JSON". This is also the default qlog serialization and default value of this field.

To facilitate this mapping, the qlog documents employ a format that is close to pure JSON for its examples and data definitions. Still, as JSON is not a typed format, there are some practical peculiarities to observe.

6.1.1. numbers

While JSON has built-in support for integers up to 64 bits in size, not all JSON parsers do. For example, none of the major Web browsers support full 64-bit integers at this time, as all numerical values (both floating-point numbers and integers) are internally represented as floating point IEEE 754 [4] values. In practice, this limits their integers to a maximum value of $2^{53}-1$. Integers larger than that are either truncated or produce a JSON parsing error. While this is expected to improve in the future (as "BigInt" support [5] has been introduced in most Browsers, though not yet integrated into JSON parsers), we still need to deal with it here.

When transforming an `int64`, `uint64` or `double` from qlog to JSON, the implementer can thus choose to either log them as JSON numbers (taking the risk of truncation or un-parseability) or to log them as strings instead. Logging as strings should however only be practically needed if the value is likely to exceed $2^{53}-1$. In practice, even though protocols such as QUIC allow 64-bit values for for example stream identifiers, these high numbers are unlikely to be reached for the overwhelming majority of cases. As such, it is probably a valid trade-off to take the risk and log 64-bit values as JSON numbers instead of strings.

Tools processing JSON-based qlog SHOULD however be able to deal with 64-bit fields being serialized as either strings or numbers.

6.1.2. bytes

Unlike most binary formats, JSON does not allow the logging of raw binary blobs directly. As such, when serializing a byte or `array<byte>`, a scheme needs to be chosen.

To represent qlog bytes in JSON, they MUST be serialized to their lowercase hexadecimal equivalents (with 0 prefix for values lower than 10). All values are directly appended to each other, without delimiters. The full value is not prefixed with 0x (as is sometimes common). An example is given in Figure 15.

For the five raw unsigned byte input values of: 5 20 40 171 255, the JSON serialization is:

```
{
  raw: "051428abff"
}
```

Figure 15: Example for serializing bytes

As such, the resulting string will always have an even amount of characters and the original byte-size can be retrieved by dividing the string length by 2.

6.1.2.1. Truncated values

In some cases, it can be interesting not to log a full raw blob but instead a truncated value (for example, only the first 100 bytes of an HTTP response body to be able to discern which file it actually contained). In these cases, the original byte-size length cannot be obtained from the serialized value directly. As such, all qlog schema definitions SHOULD include a separate, length-indicating field for all fields of type `array<byte>` they specify. This allows always retrieving the original length, but also allows the omission of any

raw value bytes of the field completely (e.g., out of privacy or security considerations).

To reduce overhead however and in the case the full raw value is logged, the extra length-indicating field can be left out. As such, tools MUST be able to deal with this situation and derive the length of the field from the raw value if no separate length-indicating field is present. All possible permutations are shown by example in Figure 16.

```
// both the full raw value and its length are present (length is redundant)
{
  "raw_length": 5,
  "raw": "051428abff"
}

// only the raw value is present, indicating it represents the fields full value
// the byte length is obtained by calculating raw.length / 2
{
  "raw": "051428abff"
}

// only the length field is present, meaning the value was omitted
{
  "raw_length": 5,
}

// both fields are present and the lengths do not match: the value was truncated
// to the first three bytes.
{
  "raw_length": 5,
  "raw": "051428"
}
```

Figure 16: Example for serializing truncated bytes

6.1.3. Summarizing table

By definition, JSON strings are serialized surrounded by quotes. Numbers without.

qlog type	JSON type
int8	number
int16	number
int32	number
uint8	number
uint16	number
uint32	number
float	number
int64	number or string
uint64	number or string
double	number or string
bytes	string (lowercase hex value)
string	string
boolean	string ("true" or "false")
enum	string (full value/name, not index)
any	object ({ ... })
array	array ([...])

6.1.4. Other JSON specifics

JSON files by definition ([RFC8259]) MUST utilize the UTF-8 encoding, both for the file itself and the string values.

Most JSON parsers strictly follow the JSON specification. This includes the rule that trailing comma's are not allowed. As it is frequently annoying to remove these trailing comma's when logging events in a streaming fashion, tool implementers SHOULD allow the last event entry of a qlog trace to be an empty object. This allows loggers to simply close the qlog file by appending "{}]]]]" after their last added event.

Finally, while not specifically required by the JSON specification, all qlog field names in a JSON serialization MUST be lowercase.

6.2. qlog to NDJSON mapping

One of the downsides of using pure JSON is that it is inherently a non-streamable format. Put differently, it is not possible to simply append new qlog events to a log file without "closing" this file at the end by appending "]]]]". Without these closing tags, most JSON parsers will be unable to parse the file entirely. As most platforms do not provide a standard streaming JSON parser (which would be able to deal with this problem), this document also provides a qlog mapping to a streamable JSON format called Newline-Delimited JSON (NDJSON) [6].

When mapping qlog to NDJSON, the "qlog_format" field MUST have the value "NDJSON".

NDJSON is very similar to JSON, except that it interprets each line in a file as a fully separate JSON object. Put differently, unlike default JSON, it does not require a file to be wrapped as a full object with "{ ... }" or "[...]". Using this setup, qlog events can simply be appended as individually serialized lines at the back of a streamed logging file.

For this to work, some qlog definitions have to be adjusted however. Mainly, events are no longer part of the "events" array in the Trace object, but are instead logged separately from the qlog "file header" (QlogFile class in Section 3). Additionally, qlog's NDJSON mapping does not allow logging multiple individual traces in a single qlog file. As such, the QlogFile:traces field is replaced by the singular "trace" field, which simply contains the Trace data directly. An example can be seen in Figure 17. Note that the "group_id" field can still be used on a per-event basis to include events from conceptually different sources in a single NDJSON qlog file.

Note as well from Figure 17 that the file's header (QlogFileNDJSON) also needs to be fully serialized on a single line to be NDJSON compatible.

Definition:

```
class QlogFileNDJSON {
    qlog_format: "NDJSON",

    qlog_version:string,
    title?:string,
    description?:string,
    summary?: Summary,
    trace: Trace
}
// list of qlog events, separated by newlines
```

NDJSON serialization:

```
{"qlog_format":"NDJSON","qlog_version":"draft-03-WIP","title":"Name of this parti
cular NDJSON qlog file (short)","description":"Description for this NDJSON qlog f
ile (long)","trace":{"common_fields":{"protocol_type": ["QUIC","HTTP3"],"group_id
":"127ecc830d98f9d54a42c4f0842aa87e181a","time_format":"relative","reference_time
":"1553986553572"},"vantage_point":{"name":"backend-67","type":"server"}}}
{"time": 2, "name": "transport:packet_received", "data": { ... } }
{"time": 7, "name": "http:frame_parsed", "data": { ... } }
```

Figure 17: Top-level element

Finally, while not specifically required by the NDJSON specification, all qlog field names in a NDJSON serialization MUST be lowercase.

6.2.1. Supporting NDJSON in tooling

Note that NDJSON is not supported in most default programming environments (unlike normal JSON). However, several custom NDJSON parsing libraries exist [7] that can be used and the format is easy enough to parse with existing implementations (i.e., by splitting the file into its component lines and feeding them to a normal JSON parser individually, as each line by itself is a valid JSON object).

6.3. Other optimized formatting options

Both the JSON and NDJSON formatting options described above are serviceable in general small to medium scale (debugging) setups. However, these approaches tend to be relatively verbose, leading to larger file sizes. Additionally, generalized (ND)JSON (de)serialization performance is typically (slightly) lower than that of more optimized and predictable formats. Both aspects make these formats more challenging (though still practical [8]) to use in large scale setups.

During the development of qlog, we compared a multitude of alternative formatting and optimization options. The results of this study are summarized on the qlog github repository [9]. The rest of this section discusses some of these approaches implementations could choose and the expected gains and tradeoffs inherent therein. Tools SHOULD support mainly the compression options listed in Section 6.3.2, as they provide the largest wins for the least cost overall.

Over time, specific qlog formats and encodings can be created that more formally define and combine some of the discussed optimizations or add new ones. We choose to define these schemes in separate documents to keep the main qlog definition clean and generalizable, as not all contexts require the same performance or flexibility as others and qlog is intended to be a broadly usable and extensible format (for example more flexibility is needed in earlier stages of protocol development, while more performance is typically needed in later stages). This is also the main reason why the general qlog format is the less optimized JSON instead of a more performant option.

To be able to easily distinguish between these options in qlog compatible tooling (without the need to have the user provide out-of-band information or to (heuristically) parse and process files in a multitude of ways, see also Section 8), we recommend using explicit file extensions to indicate specific formats. As there are no standards in place for this type of extension to format mapping, we employ a commonly used scheme here. Our approach is to list the

applied optimizations in the extension in ascending order of application (e.g., if a qlog file is first optimized with technique A and then compressed with technique B, the resulting file would have the extension ".qlog.A.B"). This allows tooling to start at the back of the extension to "undo" applied optimizations to finally arrive at the expected qlog representation.

6.3.1. Data structure optimizations

The first general category of optimizations is to alter the representation of data within an (ND)JSON qlog file to reduce file size.

The first option is to employ a scheme similar to the CSV (comma separated value [rfc4180]) format, which utilizes the concept of column "headers" to prevent repeating field names for each datapoint instance. Concretely for JSON qlog, several field names are repeated with each event (i.e., time, name, data). These names could be extracted into a separate list, after which qlog events could be serialized as an array of values, as opposed to a full object. This approach was a key part of the original qlog format (prior to draft 02) using the "event_fields" field. However, tests showed that this optimization only provided a mean file size reduction of 5% (100MB to 95MB) while significantly increasing the implementation complexity, and this approach was abandoned in favor of the default JSON setup. Implementations using this format should not employ a separate file extension (as it still uses JSON), but rather employ a new value of "JSON.namedheaders" (or "NDJSON.namedheaders") for the "qlog_format" field (see Section 3).

The second option is to replace field values and/or names with indices into a (dynamic) lookup table. This is a common compression technique and can provide significant file size reductions (up to 50% in our tests, 100MB to 50MB). However, this approach is even more difficult to implement efficiently and requires either including the (dynamic) table in the resulting file (an approach taken by for example Chromium's NetLog format [10]) or defining a (static) table up-front and sharing this between implementations. Implementations using this approach should not employ a separate file extension (as it still uses JSON), but rather employ a new value of "JSON.dictionary" (or "NDJSON.dictionary") for the "qlog_format" field (see Section 3).

As both options either proved difficult to implement, reduced qlog file readability, and provided too little improvement compared to other more straightforward options (for example Section 6.3.2), these schemes are not inherently part of qlog.

6.3.2. Compression

The second general category of optimizations is to utilize a (generic) compression scheme for textual data. As qlog in the (ND)JSON format typically contains a large amount of repetition, off-the-shelf (text) compression techniques typically succeed very well in bringing down file sizes (regularly with up to two orders of magnitude in our tests, even for "fast" compression levels). As such, utilizing compression is recommended before attempting other optimization options, even though this might (somewhat) increase processing costs due to the additional compression step.

The first option is to use GZIP compression ([RFC1952]). This generic compression scheme provides multiple compression levels (providing a trade-off between compression speed and size reduction). Utilized at level 6 (a medium setting thought to be applicable for streaming compression of a qlog stream in commodity devices), gzip compresses qlog JSON files to 7% of their initial size on average (100MB to 7MB). For this option, the file extension .qlog.gz SHOULD BE used. The "qlog_format" field should still reflect the original JSON formatting of the qlog data (e.g., "JSON" or "NDJSON").

The second option is to use Brotli compression ([RFC7932]). While similar to gzip, this more recent compression scheme provides a better efficiency. It also allows multiple compression levels. Utilized at level 4 (a medium setting thought to be applicable for streaming compression of a qlog stream in commodity devices), brotli compresses qlog JSON files to 7% of their initial size on average (100MB to 7MB). For this option, the file extension .qlog.br SHOULD BE used. The "qlog_format" field should still reflect the original JSON formatting of the qlog data (e.g., "JSON" or "NDJSON").

Other compression algorithms of course exist (for example xz, zstd, and lz4). We mainly recommend gzip and brotli because of their tweakable behaviour and wide support in web-based environments, which we envision as the main tooling ecosystem (see also Section 8).

6.3.3. Binary formats

The third general category of optimizations is to use a more optimized (often binary) format instead of the textual JSON format. This approach inherently produces smaller files and often has better (de)serialization performance. However, the resultant files are no longer human readable and some formats require hard tradeoffs between flexibility for performance.

The first option is to use the CBOR (Concise Binary Object Representation [rfc7049]) format. For our purposes, CBOR can be

viewed as a straightforward binary variant of JSON. As such, existing JSON qlog files can be trivially converted to and from CBOR (though slightly more work is needed for NDJSON qlogs). While CBOR thus does retain the full qlog flexibility, it only provides a 25% file size reduction (100MB to 75MB) compared to textual (ND)JSON. As CBOR support in programming environments is not as widespread as that of textual JSON and the format lacks human readability, CBOR was not chosen as the default qlog format. For this option, the file extension `.qlog.cbor` SHOULD BE used. The `"qlog_format"` field should still reflect the original JSON formatting of the qlog data (e.g., `"JSON"` or `"NDJSON"`).

A second option is to use a more specialized binary format, such as Protocol Buffers [11] (protobuf). This format is battle-tested, has support for optional fields and has libraries in most programming languages. Still, it is significantly less flexible than textual JSON or CBOR, as it relies on a separate, pre-defined schema (a `.proto` file). As such, it is not possible to (easily) log new event types in protobuf files without adjusting this schema as well, which has its own practical challenges. As qlog is intended to be a flexible, general purpose format, this type of format was not chosen as its basic serialization. The lower flexibility does lead to significantly reduced file sizes. Our straightforward mapping of the qlog main schema and QUIC/HTTP3 event types to protobuf created qlog files 24% as large as the raw JSON equivalents (100MB to 24MB). For this option, the file extension `.qlog.protobuf` SHOULD BE used. The `"qlog_format"` field should reflect the different internal format, for example: `"qlog_format": "protobuf"`.

Note that binary formats can (and should) also be used in conjunction with compression (see Section 6.3.2). For example, CBOR compresses well (to about 6% of the original textual JSON size (100MB to 6MB) for both gzip and brotli) and so does protobuf (5% (gzip) to 3% (brotli)). However, these gains are similar to the ones achieved by simply compressing the textual JSON equivalents directly (7%, see Section 6.3.2). As such, since compression is still needed to achieve optimal file size reductions even with binary formats, we feel the more flexible compressed textual JSON options are a better default for the qlog format in general.

6.3.4. Overview and summary

In summary, textual JSON was chosen as the main qlog format due to its high flexibility and because its inefficiencies can be largely solved by the utilization of compression techniques (which are needed to achieve optimal results with other formats as well).

Still, qlog implementers are free to define other qlog formats depending on their needs and context of use. These formats should be described in their own documents, the discussion in this document mainly acting as inspiration and high-level guidance. Implementers are encouraged to add concrete qlog formats and definitions to the designated public repository [12].

The following table provides an overview of all the discussed qlog formatting options with examples:

format	qlog_format	extension
JSON Section 6.1	JSON	.qlog
NDJSON Section 6.2	NDJSON	.qlog
named headers Section 6.3.1	(ND)JSON.namedheaders	.qlog
dictionary Section 6.3.1	(ND)JSON.dictionary	.qlog
CBOR Section 6.3.3	(ND)JSON	.qlog.cbor
protobuf Section 6.3.3	protobuf	.qlog.protobuf
gzip Section 6.3.2	no change	.gz suffix
brrotli Section 6.3.2	no change	.br suffix

6.4. Conversion between formats

As discussed in the previous sections, a qlog file can be serialized in a multitude of formats, each of which can conceivably be transformed into or from one another without loss of information. For example, a number of NDJSON streamed qlogs could be combined into a JSON formatted qlog for later processing. Similarly, a captured binary qlog could be transformed to JSON for easier interpretation and sharing.

Secondly, we can also consider other structured logging approaches that contain similar (though typically not identical) data to qlog, like raw packet capture files (for example .pcap files from tcpdump) or endpoint-specific logging formats (for example the NetLog format in Google Chrome). These are sometimes the only options, if an implementation cannot or will not support direct qlog output for any reason, but does provide other internal or external (e.g., SSLKEYLOGFILE export to allow decryption of packet captures) logging options. For this second category, a (partial) transformation from/to qlog can also be defined.

As such, when defining a new qlog serialization format or wanting to utilize qlog-compatible tools with existing codebases lacking qlog

support, it is recommended to define and provide a concrete mapping from one format to default JSON-serialized qlog. Several of such mappings exist. Firstly, [pcap2qlog] (<https://github.com/quiclog/pcap2qlog>) transforms QUIC and HTTP/3 packet capture files to qlog. Secondly, netlog2qlog [13] converts chromium's internal dictionary-encoded JSON format to qlog. Finally, quictrace2qlog [14] converts the older quictrace format to JSON qlog. Tools can then easily integrate with these converters (either by incorporating them directly or for example using them as a (web-based) API) so users can provide different file types with ease. For example, the qvis [15] toolsuite supports a multitude of formats and qlog serializations.

7. Methods of access and generation

Different implementations will have different ways of generating and storing qlogs. However, there is still value in defining a few default ways in which to steer this generation and access of the results.

7.1. Set file output destination via an environment variable

To provide users control over where and how qlog files are created, we define two environment variables. The first, QLOGFILE, indicates a full path to where an individual qlog file should be stored. This path **MUST** include the full file extension. The second, QLOGDIR, sets a general directory path in which qlog files should be placed. This path **MUST** include the directory separator character at the end.

In general, QLOGDIR should be preferred over QLOGFILE if an endpoint is prone to generate multiple qlog files. This can for example be the case for a QUIC server implementation that logs each QUIC connection in a separate qlog file. An alternative that uses QLOGFILE would be a QUIC server that logs all connections in a single file and uses the "group_id" field (Section 3.4.6) to allow post-hoc separation of events.

Implementations **SHOULD** provide support for QLOGDIR and **MAY** provide support for QLOGFILE.

When using QLOGDIR, it is up to the implementation to choose an appropriate naming scheme for the qlog files themselves. The chosen scheme will typically depend on the context or protocols used. For example, for QUIC, it is recommended to use the Original Destination Connection ID (ODCID), followed by the vantage point type of the logging endpoint. Examples of all options for QUIC are shown in Figure 18.

Command: `QLOGFILE=/srv/qlogs/client.qlog quicclientbinary`

Should result in the the `quicclientbinary` executable logging a single qlog file named `client.qlog` in the `/srv/qlogs` directory.

This is for example useful in tests when the client sets up just a single connection and then exits.

Command: `QLOGDIR=/srv/qlogs/ quicserverbinary`

Should result in the `quicserverbinary` executable generating several logs files, one for each QUIC connection.

Given two QUIC connections, with ODCID values `"abcde"` and `"12345"` respectively, this would result in two files:

`/srv/qlogs/abcde_server.qlog`

`/srv/qlogs/12345_server.qlog`

Command: `QLOGFILE=/srv/qlogs/server.qlog quicserverbinary`

Should result in the the `quicserverbinary` executable logging a single qlog file named `server.qlog` in the `/srv/qlogs` directory.

Given that the server handled two QUIC connections before it was shut down, with ODCID values `"abcde"` and `"12345"` respectively, this would result in event instances in the qlog file being tagged with the `"group_id"` field with values `"abcde"` and `"12345"`.

Figure 18: Environment variable examples for a QUIC implementation

7.2. Access logs via a well-known endpoint

After generation, qlog implementers MAY make available generated logs and traces on an endpoint (typically the server) via the following .well-known URI:

`.well-known/qlog/IDENTIFIER.extension`

The `IDENTIFIER` variable depends on the context and the protocol. For example for QUIC, the lowercase Original Destination Connection ID (ODCID) is recommended, as it can uniquely identify a connection. Additionally, the extension depends on the chosen format (see Section 6.3.4). For example, for a QUIC connection with ODCID `"abcde"`, the endpoint for fetching its default JSON-formatted .qlog file would be:

`.well-known/qlog/abcde.qlog`

Implementers SHOULD allow users to fetch logs for a given connection on a 2nd, separate connection. This helps prevent pollution of the logs by fetching them over the same connection that one wishes to observe through the log. Ideally, for the QUIC use case, the logs should also be approachable via an HTTP/2 or HTTP/1.1 endpoint (i.e., on TCP port 443), to for example aid debugging in the case where QUIC/UDP is blocked on the network.

qlog implementers SHOULD NOT enable this .well-known endpoint in typical production settings to prevent (malicious) users from

downloading logs from other connections. Implementers are advised to disable this endpoint by default and require specific actions from the end users to enable it (and potentially qlog itself). Implementers MUST also take into account the general privacy and security guidelines discussed in Section 9 before exposing qlogs to outside actors.

8. Tooling requirements

Tools ingestion qlog MUST indicate which qlog version(s), qlog format(s), compression methods and potentially other input file formats (for example .pcap) they support. Tools SHOULD at least support .qlog files in the default JSON format (Section 6.1). Additionally, they SHOULD indicate exactly which values for and properties of the name (category and type) and data fields they look for to execute their logic. Tools SHOULD perform a (high-level) check if an input qlog file adheres to the expected qlog schema. If a tool determines a qlog file does not contain enough supported information to correctly execute the tool's logic, it SHOULD generate a clear error message to this effect.

Tools MUST NOT produce breaking errors for any field names and/or values in the qlog format that they do not recognize. Tools SHOULD indicate even unknown event occurrences within their context (e.g., marking unknown events on a timeline for manual interpretation by the user).

Tool authors should be aware that, depending on the logging implementation, some events will not always be present in all traces. For example, using a circular logging buffer of a fixed size, it could be that the earliest events (e.g., connection setup events) are later overwritten by "newer" events. Alternatively, some events can be intentionally omitted out of privacy or file size considerations. Tool authors are encouraged to make their tools robust enough to still provide adequate output for incomplete logs.

9. Security and privacy considerations

TODO : discuss privacy and security considerations (e.g., what NOT to log, what to strip out of a log before sharing, ...)

TODO: strip out/don't log IPs, ports, specific CIDs, raw user data, exact times, HTTP HEADERS (or at least :path), SNI values

TODO: see if there is merit in encrypting the logs and having the server choose an encryption key (e.g., sent in transport parameters)

Good initial reference: Christian Huitema's blogpost [16]

10. IANA Considerations

TODO: primarily the .well-known URI

11. References

11.1. Normative References

[QLOG-H3] Marx, R., Ed., Niccolini, L., Ed., and M. Seemann, Ed., "HTTP/3 and QPACK event definitions for qlog", draft-marx-qlog-h3-events-00 (work in progress).

[QLOG-QUIC] Marx, R., Ed., Niccolini, L., Ed., and M. Seemann, Ed., "QUIC event definitions for qlog", draft-marx-qlog-quic-events-00 (work in progress).

11.2. Informative References

[RFC1952] Deutsch, P., "GZIP file format specification version 4.3", RFC 1952, DOI 10.17487/RFC1952, May 1996, <<https://www.rfc-editor.org/info/rfc1952>>.

[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>>.

[rfc4180] Shafranovich, Y., "Common Format and MIME Type for Comma-Separated Values (CSV) Files", RFC 4180, DOI 10.17487/RFC4180, October 2005, <<https://www.rfc-editor.org/info/rfc4180>>.

[rfc7049] Bormann, C. and P. Hoffman, "Concise Binary Object Representation (CBOR)", RFC 7049, DOI 10.17487/RFC7049, October 2013, <<https://www.rfc-editor.org/info/rfc7049>>.

[RFC7932] Alakuijala, J. and Z. Szabadka, "Brotli Compressed Data Format", RFC 7932, DOI 10.17487/RFC7932, July 2016, <<https://www.rfc-editor.org/info/rfc7932>>.

[RFC8259] Bray, T., Ed., "The JavaScript Object Notation (JSON) Data Interchange Format", STD 90, RFC 8259, DOI 10.17487/RFC8259, December 2017, <<https://www.rfc-editor.org/info/rfc8259>>.

11.3. URIs

- [1] <https://github.com/quiclog/internet-drafts>
- [2] <https://www.typescriptlang.org/>
- [3] <https://qvis.edm.uhasselt.be>
- [4] https://en.wikipedia.org/wiki/Floating-point_arithmetic
- [5] https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference/Global_Objects/BigInt
- [6] <http://ndjson.org/>
- [7] <http://ndjson.org/libraries.html>
- [8] <https://qlog.edm.uhasselt.be/anrw/>
- [9] <https://github.com/quiclog/internet-drafts/issues/30#issuecomment-617675097>
- [10] <https://www.chromium.org/developers/design-documents/network-stack/netlog>
- [11] <https://developers.google.com/protocol-buffers>
- [12] <https://github.com/quiclog/qlog>
- [13] <https://github.com/quiclog/qvis/tree/master/visualizations/src/components/filemanager/netlogconverter>
- [14] <https://github.com/quiclog/quictrace2qlog>
- [15] <https://qvis.edm.uhasselt.be>
- [16] <https://huitema.wordpress.com/2020/07/21/scrubbing-quic-logs-for-privacy/>
- [17] <https://github.com/google/quic-trace>
- [18] <https://github.com/EricssonResearch/spindump>
- [19] <https://www.wireshark.org/>

Appendix A. Change Log

A.1. Since draft-marx-qlog-main-schema-draft-02:

- o These changes were done in preparation of the adoption of the drafts by the QUIC working group (#137)
- o Moved RawInfo, Importance, Generic events and Simulation events to this document.
- o Added basic event definition guidelines
- o Made protocol_type an array instead of a string (#146)

A.2. Since draft-marx-qlog-main-schema-01:

- o Decoupled qlog from the JSON format and described a mapping instead (#89)
 - * Data types are now specified in this document and proper definitions for fields were added in this format
 - * 64-bit numbers can now be either strings or numbers, with a preference for numbers (#10)
 - * binary blobs are now logged as lowercase hex strings (#39, #36)
 - * added guidance to add length-specifiers for binary blobs (#102)
- o Removed "time_units" from Configuration. All times are now in ms instead (#95)
- o Removed the "event_fields" setup for a more straightforward JSON format (#101, #89)
- o Added a streaming option using the NDJSON format (#109, #2, #106)
- o Described optional optimization options for implementers (#30)
- o Added QLOGDIR and QLOGFILE environment variables, clarified the .well-known URL usage (#26, #33, #51)
- o Overall tightened up the text and added more examples

A.3. Since draft-marx-qlog-main-schema-00:

- o All field names are now lowercase (e.g., category instead of CATEGORY)
- o Triggers are now properties on the "data" field value, instead of separate field types (#23)
- o group_ids in common_fields is now just also group_id

Appendix B. Design Variations

- o Quic-trace [17] takes a slightly different approach based on protocolbuffers.
- o Spindump [18] also defines a custom text-based format for in-network measurements
- o Wireshark [19] also has a QUIC dissector and its results can be transformed into a json output format using tshark.

The idea is that qlog is able to encompass the use cases for both of these alternate designs and that all tooling converges on the qlog standard.

Appendix C. Acknowledgements

Much of the initial work by Robin Marx was done at Hasselt University.

Thanks to Jana Iyengar, Brian Trammell, Dmitri Tikhonov, Stephen Petrides, Jari Arkko, Marcus Ihlar, Victor Vasiliev, Mirja Kuehlewind, Jeremy Laine and Lucas Pardue for their feedback and suggestions.

Authors' Addresses

Robin Marx (editor)
KU Leuven

Email: robin.marx@kuleuven.be

Luca Niccolini (editor)
Facebook

Email: lniccolini@fb.com

Marten Seemann (editor)
Protocol Labs

Email: marten@protocol.ai