Network Working Group
Internet-Draft
Intended status: Experimental
Expires: May 4, 2017

R. Hamilton
J. Iyengar
I. Swett
A. Wilk
Google

October 31, 2016

QUIC: A UDP-Based Multiplexed and Secure Transport draft-hamilton-quic-transport-protocol-01

Abstract

QUIC is a multiplexed and secure transport protocol that runs on top of UDP. QUIC builds on past transport experience, and implements mechanisms that make it useful as a modern general-purpose transport protocol. Using UDP as the basis of QUIC is intended to address compatibility issues with legacy clients and middleboxes. QUIC authenticates all of its headers, preventing third parties from from changing them. QUIC encrypts most of its headers, thereby limiting protocol evolution to QUIC endpoints only. Therefore, middleboxes, in large part, are not required to be updated as new protocol versions are deployed. This document describes the core QUIC protocol, including the conceptual design, wire format, and mechanisms of the QUIC protocol for connection establishment, stream multiplexing, stream and connection-level flow control, and data reliability. Accompanying documents describe QUIC's loss recovery and congestion control, and the use of TLS 1.3 for key negotiation.

Table of Contents

1. Introduction

QUIC is a multiplexed and secure transport protocol that runs on top of UDP. QUIC builds on past transport experience and implements mechanisms that make it useful as a modern general-purpose transport protocol. Using UDP as the substrate, QUIC seeks to be compatible with legacy clients and middleboxes. QUIC authenticates all of its headers, preventing middleboxes and other third parties from changing them, and encrypts most of its headers, limiting protocol evolution largely to QUIC endpoints only.

This document describes the core QUIC protocol, including the conceptual design, wire format, and mechanisms of the QUIC protocol

for connection establishment, stream multiplexing, stream and connection-level flow control, and data reliability. Accompanying documents describe QUIC's loss detection and congestion control [draft-iyengar-quic-loss-detection], and the use of TLS 1.3 for key negotiation [draft-thomson-quic-tls].

2. Conventions and Definitions

Definitions of terms that are used in this document:

- Client: The endpoint initiating a QUIC connection.
- Server: The endpoint accepting incoming QUIC connections.
- Endpoint: The client or server end of a connection.
- Stream: A logical, bi-directional channel of ordered bytes within a QUIC connection.
- Connection: A conversation between two QUIC endpoints with a single encryption context that multiplexes streams within it.
- Connection ID: The identifier for a QUIC connection.
- QUIC packet: A well-formed UDP payload that can be parsed by a QUIC receiver. QUIC packet size in this document refers to the UDP payload size.

3. A QUIC Overview

This section briefly describes QUIC's key mechanisms and benefits. Key strengths of QUIC include:

- Low-latency Version Negotiation
- Low-latency connection establishment
- Multiplexing without head-of-line blocking
- Authenticated and encrypted header and payload
- Rich signaling for congestion control and loss recovery
- Stream and connection flow control
- Connection Migration and Resilience to NAT rebinding

3.1. Low-Latency Version Negotiation

QUIC combines version negotiation with the rest of connection establishment to avoid unnecessary roundtrip delays. A QUIC client proposes a version to use for the connection, and encodes the rest of the handshake using the proposed version. If the server does not speak the client-chosen version, it forces version negotiation by sending back a Version Negotiation packet to the client, causing a roundtrip of delay before connection establishment.

This mechanism eliminates roundtrip latency when the client's optimistically-chosen version is spoken by the server, and incentivizes servers to not lag behind clients in deployment of newer versions. Additionally, an application may negotiate QUIC versions out-of-band to increase chances of success in the first roundtrip and to obviate the additional roundtrip in the case of version mismatch.

3.2. Low-Latency Connection Establishment

QUIC relies on a combined crypto and transport handshake for setting up a secure transport connection. QUIC connections are expected to commonly use 0-RTT handshakes, meaning that for most QUIC connections, data can be sent immediately following the client handshake packet, without waiting for a reply from the server. QUIC provides a dedicated stream (Stream ID 1) to be used for performing the crypto handshake and QUIC options negotiation. The format of the QUIC options and parameters used during negotiation are described in this document, but the handshake protocol that runs on Stream ID 1 is described in the accompanying crypto handshake draft [draft-thomson-quic-tls].

3.3. Stream Multiplexing

When application messages are transported over TCP, independent application messages can suffer from head-of-line blocking. When an application multiplexes many streams atop TCP's single-bytestream abstraction, a loss of a TCP segment results in blocking of all subsequent segments until a retransmission arrives, irrespective of the application streams that are encapsulated in subsequent segments. QUIC ensures that lost packets carrying data for an individual stream only impact that specific stream. Data received on other streams can continue to be reassembled and delivered to the application.

3.4. Rich Signaling for Congestion Control and Loss Recovery

QUIC's packet framing and acknowledgments carry rich information that help both congestion control and loss recovery in fundamental ways. Each QUIC packet carries a new packet number, including those

carrying retransmitted data. This obviates the need for a separate mechanism to distinguish acks for retransmissions from those for original transmissions, avoiding TCP's retransmission ambiguity problem. QUIC acknowledgments also explicitly encode the delay between the receipt of a packet and its acknowledgment being sent, and together with the monotonically-increasing packet numbers, this allows for precise network roundtrip-time (RTT) calculation. QUIC's ACK frames support up to 256 ack blocks, so QUIC is more resilient to reordering than TCP with SACK support, as well as able to keep more bytes on the wire when there is reordering or loss.

3.5. Stream and Connection Flow Control

QUIC implements stream- and connection-level flow control, closely following HTTP/2's flow control mechanisms. At a high level, a QUIC receiver advertises the absolute byte offset within each stream up to which the receiver is willing to receive data. As data is sent, received, and delivered on a particular stream, the receiver sends WINDOW_UPDATE frames that increase the advertised offset limit for that stream, allowing the peer to send more data on that stream. In addition to this stream-level flow control, QUIC implements connection-level flow control to limit the aggregate buffer that a QUIC receiver is willing to allocate to all streams on a connection. Connection-level flow control works in the same way as stream-level flow control, but the bytes delivered and highest received offset are all aggregates across all streams.

3.6. Authenticated and Encrypted Header and Payload

TCP headers appear in plaintext on the wire and are not authenticated, causing a plethora of injection and header manipulation issues for TCP, such as receive-window manipulation and sequence-number overwriting. While some of these are mechanisms used by middleboxes to improve TCP performance, others are active attacks. Even "performance-enhancing" middleboxes that routinely interpose on the transport state machine end up limiting the evolvability of the transport protocol, as has been observed in the design of MPTCP and in its subsequent deployability issues.

Generally, QUIC packets are always authenticated and the payload is typically fully encrypted. The parts of the packet header which are not encrypted are still authenticated by the receiver, so as to thwart any packet injection or manipulation by third parties. Some early handshake packets, such as the Version Negotiation packet, are not encrypted, but information sent in these unencrypted handshake packets is later verified under crypto cover.

PUBLIC_RESET packets that reset a connection are currently not authenticated.

3.7. Connection Migration and Resilience to NAT Rebinding

QUIC connections are identified by a 64-bit Connection ID, randomly generated by the client. QUIC's consistent connection ID allows connections to survive changes to the client's IP and port, such as those caused by NAT rebindings or by the client changing network connectivity to a new address. QUIC provides automatic cryptographic verification of a rebound client, since the client continues to use the same session key for encrypting and decrypting packets. The consistent connection ID can be used to allow migration of the connection to a new server IP address as well, since the Connection ID remains consistent across changes in the client's and the server's network addresses.

4. Packet Types and Formats

We first describe QUIC's packet types and formats, since some are referenced in subsequent mechanisms. Note that unless otherwise noted, all values specified in this document are in little-endian format and all field sizes are in bits.

4.1. Common Header

All QUIC packets begin with a QUIC Common header, as shown below:

The fields in the Common Header are the following:

- Flags:
  - 0x01 = VERSION. The semantics of this flag depends on whether the packet is sent by the server or the client. A client MAY set this flag and include exactly one proposed version. A server may set this flag when the client-proposed version was unsupported, and may then provide a list (0 or more) of acceptable versions as a part of version negotiation (described in Section XXX.)
  - 0x02 = PUBLIC_RESET. Set to indicate that the packet is a Public Reset packet.

  * 0x04 = DIVERSIFICATION_NONCE. Set to indicate the presence of a 32-byte diversification nonce in the header. (DISCUSS_AND_MODIFY: This flag should be removed along with the Diversification Nonce bits, as discussed further below.)

  * 0x08 = CONNECTION_ID. Indicates the Connection ID is present in the packet. This must be set in all packets until negotiated to a different value for a given direction. For instance, if a client indicates that the 5-tuple fully identifies the connection at the client, the connection ID is optional in the server-to-client direction.

  * 0x30 = PACKET_NUMBER_SIZE. These two bits indicate the number of low-order-bytes of the packet number that are present in each packet.

    + 11 indicates that 6 bytes of the packet number are present

    + 10 indicates that 4 bytes of the packet number are present

    + 01 indicates that 2 bytes of the packet number are present

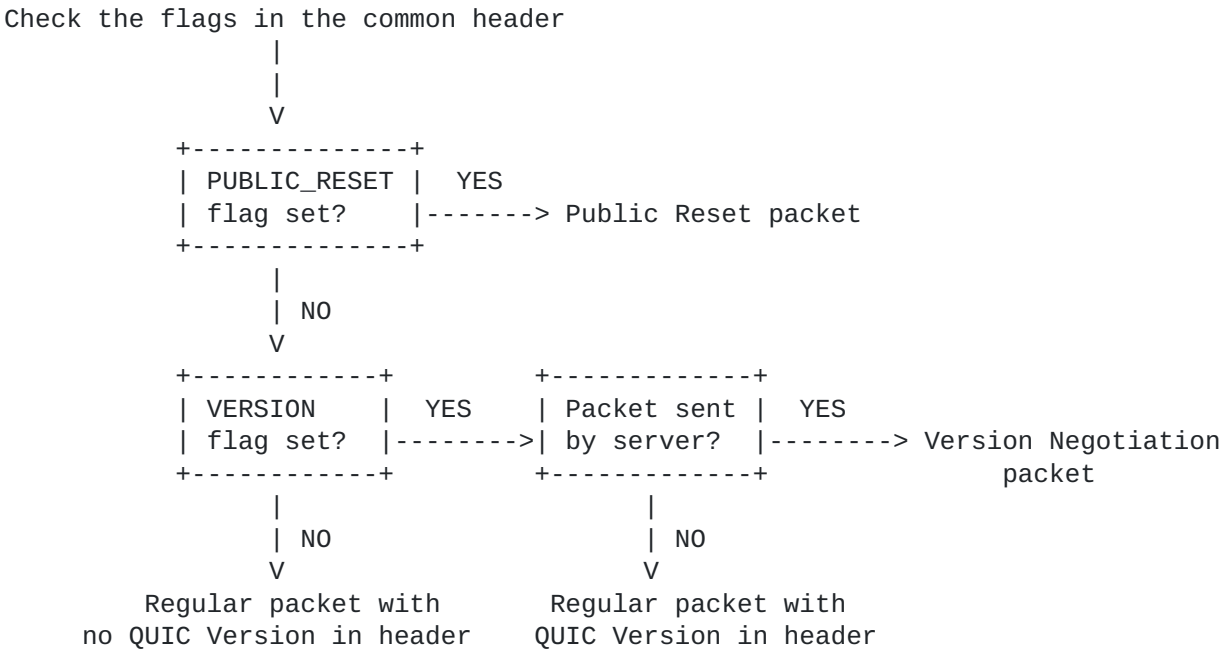
    + 00 indicates that 1 byte of the packet number is present

  * 0x40 = MULTIPATH. This bit is reserved for multipath use.

  * 0x80 is currently unused, and must be set to 0.

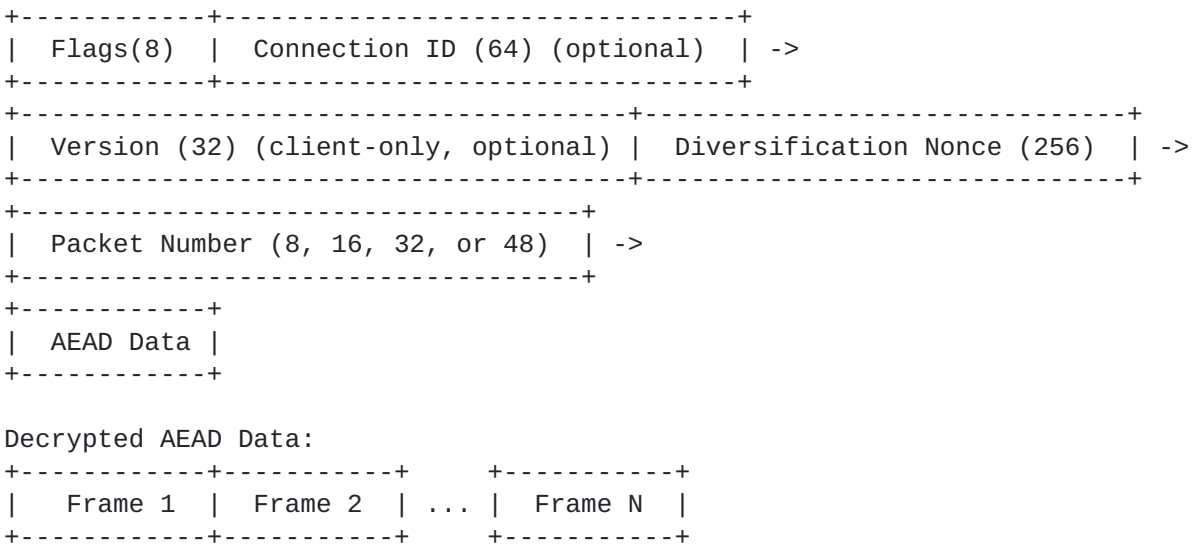
- Connection ID: An unsigned 64-bit random number chosen by the client, used as the identifier of the connection. Connection ID is tied to a QUIC connection, and remains consistent across client and/or server IP and port changes.

While all QUIC packets have the same common header, there are three types of packets: Regular packets, Version Negotiation packets, and Public Reset packets. The flowchart below shows how a packet is classified into one of these three packet types:



4.2. Regular Packets

Each Regular packet's header consists of a Common Header followed by fields specific to Regular packets, as shown below:



The fields in a Regular packet past the Common Header are the following:

- QUIC Version: A 32-bit opaque tag that represents the version of the QUIC protocol. Only present in the client-to-server direction, and if the VERSION flag is set. Version Negotiation is described in Section XXX.

- DISCUSS_AND_REPLACE: Diversification Nonce: A 32-byte nonce generated by the server and used only in the Server->Client direction to ensure that the server is able to generate unique keys per connection. Specifically, when using QUIC's 0-RTT crypto handshake, a repeated CHLO with the exact same connection ID and CHLO can lead to the same (intermediate) initial-encryption keys being derived for the connection. A server-generated nonce disallows a client from causing the same keys to be derived for two distinct connections. Once the connection is forward-secure, this nonce is no longer present in packets. This nonce can be removed from the packet header if a requirement can be added for the crypto handshake to ensure key uniqueness. The expectation is that TLS1.3 meets this requirement. Upon working group adoption of this document, this requirement should be added to the crypto handshake requirements, and the nonce should be removed from the packet format.

- Packet Number: The lower 8, 16, 32, or 48 bits of the packet number, based on the PACKET_NUMBER_SIZE flag. Each Regular packet is assigned a packet number by the sender. The first packet sent by an endpoint MUST have a packet number of 1.

- AEAD Data: A Regular packet's header, which includes the Common Header, and the Version, Diversification Nonce, and Packet Number fields, is authenticated but not encrypted. The rest of a Regular packet, starting with the first frame, is both authenticated and encrypted. Immediately following the header, Regular packets contain AEAD (Authenticated Encryption with Associated Data) data. This data must be decrypted in order for the contents to be interpreted. After decryption, the plaintext consists of a sequence of frames, as shown (frames are described in Section XXX).

4.2.1. Packet Number Compression and Reconstruction

The complete packet number is a 64-bit unsigned number and is used as part of a cryptographic nonce for packet encryption. To reduce the number of bits required to represent the packet number over the wire, at most 48 bits of the packet number are transmitted over the wire. A QUIC endpoint MUST NOT reuse a complete packet number within the same connection (that is, under the same cryptographic keys). If the total number of packets transmitted in this connection reaches 2^64 - 1, the sender MUST close the connection by sending a CONNECTION_CLOSE

frame with the error code QUIC_SEQUENCE_NUMBER_LIMIT_REACHED (connection termination is described in Section XXX.) For unambiguous reconstruction of the complete packet number by a receiver from the lower-order bits, a QUIC sender MUST NOT have more than 2^(packet_number_size - 2) in flight at any point in the connection. In other words,

- If a sender sets PACKET_NUMBER_SIZE bits to 11, it MUST NOT have more than (2^46) packets in flight.
- If a sender sets PACKET_NUMBER_SIZE bits to 10, it MUST NOT have more than (2^30) packets in flight.
- If a sender sets PACKET_NUMBER_SIZE bits to 01, it MUST NOT have more than (2^14) packets in flight.
- If a sender sets PACKET_NUMBER_SIZE bits to 00, it MUST NOT have more than (2^6) packets in flight.

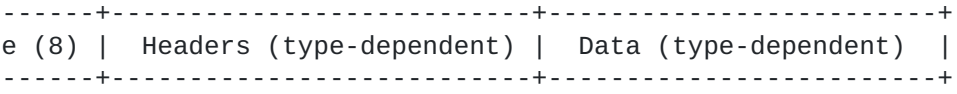
DISCUSS: Should the receiver be required to enforce this rule that the sender MUST NOT exceed the inflight limit? Specifically, should the receiver drop packets that are received outside this window?

Any truncated packet number received from a peer MUST be reconstructed as the value closest to the next expected packet number from that peer.

(TODO: Clarify how packet number size can change mid-connection.)

4.2.2. Frames and Frame Types

A Regular packet MUST contain at least one frame, and MAY contain multiple frames and multiple frame types. Frames MUST fit within a single QUIC packet and MUST NOT span a QUIC packet boundary. Each frame begins with a Frame Type byte, indicating its type, followed by type-dependent headers, and variable-length data, as follows:

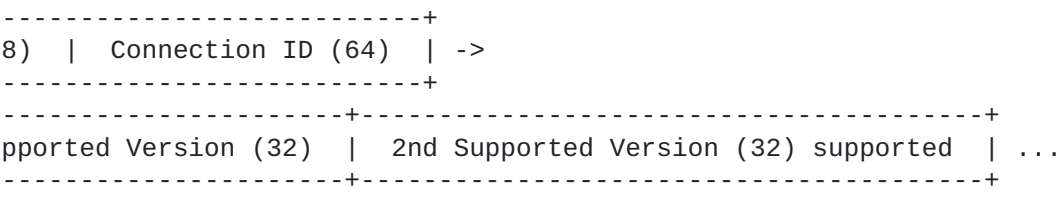


The following table lists currently defined frame types. Note that the Frame Type byte in STREAM and ACK frames is used to carry other frame-specific flags. For all other frames, the Frame Type byte simply identifies the frame. These frames are explained in more detail as they are referenced later in the document.

Type-field value	Frame type
1FDOOOSS	STREAM
01NTLLMM	ACK
00000000 (0x00)	PADDING
00000001 (0x01)	RST_STREAM
00000010 (0x02)	CONNECTION_CLOSE
00000011 (0x03)	GOAWAY
00000100 (0x04)	WINDOW_UPDATE
00000101 (0x05)	BLOCKED
00000110 (0x06)	STOP_WAITING
00000111 (0x07)	PING

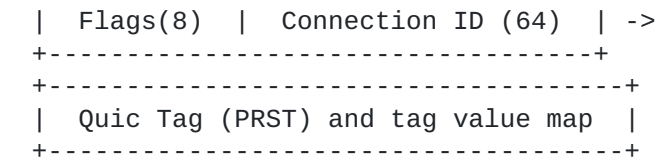
4.3. Version Negotiation Packet

A Version Negotiation packet is only sent by the server, MUST have the VERSION flag set, and MUST include the full 64-bit Connection ID. The rest of the Version Negotiation packet is a list of 4-byte versions which the server supports, as shown below.



4.4. Public Reset Packet

A Public Reset packet MUST have the PUBLIC_RESET flag set, and MUST include the full 64-bit connection ID. The rest of the Public Reset packet is encoded as if it were a crypto handshake message of the tag PRST, as shown below.



The tag value map contains the following tag-values:

- RNON (public reset nonce proof) - a 64-bit unsigned integer.
- RSEQ (rejected packet number) - a 64-bit packet number.

- CADR (client address) - the observed client IP address and port number. This is currently for debugging purposes only and hence is optional.

DISCUSS_AND_REPLACE: The crypto handshake message format is described in the QUIC crypto document, and should be replaced with something simpler when this document is adopted. The purpose of the tag-value map following the PRST tag is to enable the receiver of the Public Reset packet to reasonably authenticate the packet. This map is an extensible map format that allows specification of various tags, which should again be replaced by something simpler.

5. Life of a Connection

A QUIC connection is a single conversation between two QUIC endpoints. QUIC's connection establishment intertwines version negotiation with the crypto and transport handshakes to reduce connection establishment latency, as described in Section XXX. Once established, a connection may migrate to a different IP or port at either endpoint, due to NAT rebinding or mobility, as described in Section XXX. Finally a connection may be terminated by either endpoint, as described in Section XXX.

5.1. Version Negotiation

QUIC's connection establishment begins with version negotiation, since all communication between the endpoints, including packet and frame formats, relies on the two endpoints agreeing on a version.

A QUIC connection begins with a client sending a handshake packet. The details of the handshake mechanisms are described in Section XX, but all of the initial packets sent from the client to the server MUST have the VERSION flag set, and MUST specify the version of the protocol being used.

When the server receives a packet from a client with the VERSION flag set for a connection that has not yet been established, it compares the client's version to the versions it supports.

- If the client's version is acceptable to the server, the server MUST use this protocol version for the lifetime of the connection. All subsequent packets sent by the server MUST have the version flag off.

- If the client's version is not acceptable to the server, the server MUST send a Version Negotiation packet to the client. This packet will have the VERSION flag set and will include the server's set of supported versions. On subsequently received

packets for the same connection ID with the unacceptable version, the server MUST continue responding with a Version Negotiation packet.

When the client receives a Version Negotiation packet from the server, it should select an acceptable protocol version. If such a version is found, the client MUST resend all packets using the new version, and the resent packets MUST use new packet numbers. These packets MUST continue to have the VERSION flag set and MUST include the new negotiated protocol version.

The client MUST send its version on all packets until it receives a packet from the server with the VERSION flag off. If version negotiation is successful, the client should receive a packet from the server with the VERSION flag off indicating the end of version negotiation. All subsequent packets the client sends MUST have the version flag off.

Once the server receives a packet from the client with the VERSION flag off, it MUST ignore the VERSION flag in subsequently received packets.

The Version Negotiation packet is unencrypted and exchanged without authentication. To avoid a downgrade attack, the client needs to verify its record of the server's version list in the Version Negotiation packet and the server needs to verify its record of the client's originally proposed version. Therefore, the client and server MUST include this information later in their corresponding crypto handshake data.

5.2. Crypto and Transport Handshake

QUIC relies on a combined crypto and transport handshake to minimize connection establishment latency. QUIC provides a dedicated stream (Stream ID 1) to be used for performing a combined connection and security handshake (streams are described in detail in Section XXX). The crypto handshake protocol encapsulates and delivers QUIC's transport handshake to the peer on the crypto stream. The first QUIC packet from the client to the server MUST carry handshake information as data on Stream ID 1.

5.2.1. Transport Parameters and Options

During connection establishment, the handshake must negotiate various transport parameters. The currently defined transport parameters are described later in the document.

The transport component of the handshake is responsible for exchanging and negotiating the following parameters for a QUIC connection. Not all parameters are negotiated, some are parameters sent in just one direction. These parameters and options are encoded and handed off to the crypto handshake protocol to be transmitted to the peer.

5.2.1.1. Encoding

(TODO: Describe format with example)

QUIC encodes the transport parameters and options as tag-value pairs, all as 7-bit ASCII strings. QUIC parameter tags are listed below.

5.2.1.2. Required Transport Parameters

- SFCW: Stream Flow Control Window. The stream level flow control byte offset advertised by the sender of this parameter.

- CFCW: Connection Flow Control Window. The connection level flow control byte offset advertised by the sender of this parameter.

- MSPC: Maximum number of incoming streams per connection.

5.2.1.3. Optional Transport Parameters

- TCID: Indicates support for truncated Connection IDs. If sent by a peer, indicates that connection IDs sent to the peer should be truncated to 0 bytes. This is expected to commonly be used by an endpoint where the 5-tuple is sufficient to identify a connection. For instance, if the 5-tuple is unique at the client, the client MAY send a TCID parameter to the server. When a TCID parameter is received, an endpoint MAY choose to not send the connection ID on subsequent packets.

- COPT: Connection Options are a repeated tag field. The field contains any connection options being requested by the client or server. These are typically used for experimentation and will evolve over time. Example use cases include changing congestion control algorithms and parameters such as initial window. (TODO: List connection options.)

5.2.2. Proof of Source Address Ownership

Transport protocols commonly use a roundtrip time to verify a client's address ownership for protection from malicious clients that spoof their source address. QUIC uses a cookie, called the Source Address Token (STK), to mostly eliminate this roundtrip of delay.

This technique is similar to TCP Fast Open's use of a cookie to avoid a roundtrip of delay in TCP connection establishment.

On a new connection, a QUIC server sends an STK, which is opaque to and stored by the client. On a subsequent connection, the client echoes it in the transport handshake as proof of IP ownership.

A QUIC server also uses the STK to store server-designated connection IDs for Stateless Rejects, to verify that an incoming connection contains the correct connection ID.

A QUIC server MAY additionally store other data in a the STK, such as measured bandwidth and measured minimum RTT to the client that may help the server better bootstrap a subsequent connection from the same client. A server MAY send an updated STK message mid-connection to update server state that is stored at the client in the STK.

(TODO: Describe server and client actions on STK, encoding, recommendations for what to put in an STK. Describe SCUP messages.)

5.2.3. Crypto Handshake Protocol Features

QUIC's current crypto handshake mechanism is documented in [QUIC-CRYPTO]. QUIC does not restrict itself to using a specific handshake protocol, so the details of a specific handshake protocol are out of this document's scope. If not explicitly specified in the application mapping, TLS is assumed to be the default crypto handshake protocol, as described in [draft-mthomson-quic-tls]. An application that maps to QUIC MAY however specify an alternative crypto handshake protocol to be used.

The following list of requirements and recommendations documents properties of the current prototype handshake which should be provided by any handshake protocol.

- The crypto handshake MUST ensure that the final negotiated key is distinct for every connection between two endpoints.

- Transport Negotiation: The crypto handshake MUST provide a mechanism for the transport component to exchange transport parameters and Source Address Tokens. To avoid downgrade attacks, the transport parameters sent and received MUST be verified before the handshake completes successfully.

- Connection Establishment in 0-RTT: Since low-latency connection establishment is a critical feature of QUIC, the QUIC handshake protocol SHOULD attempt to achieve 0-RTT connection establishment latency for repeated connections between the same endpoints.

- Source Address Spoofing Defense: Since QUIC handles source address verification, the crypto protocol SHOULD NOT impose a separate source address verification mechanism.

- Server Config Update: A QUIC server may refresh the source-address token (STK) mid-connection, to update the information stored in the STK at the client and to extend the period over which 0-RTT connections can be established by the client.

- Certificate Compression: Early QUIC experience demonstrated that compressing certificates exchanged during a handshake is valuable in reducing latency. This additionally helps to reduce the amplification attack footprint when a server sends a large set of certificates, which is not uncommon with TLS. The crypto protocol SHOULD compress certificates and any other information to minimize the number of packets sent during a handshake.

The following information used during the QUIC handshake MUST be cryptographically verified by the crypto handshake protocol:

- Client's originally proposed version in its first packet.

- Server's version list in it's Version Negotiation packet, if one was sent.

5.3. Connection Migration

QUIC connections are identified by their 64-bit Connection ID. QUIC's consistent connection ID allows connections to survive changes to the client's IP and/or port, such as those caused by client or server migrating to a new network. QUIC also provides automatic cryptographic verification of a rebound client, since the client continues to use the same session key for encrypting and decrypting packets.

DISCUSS: Simultaneous migration. Is this reasonable?

TODO: Perhaps move mitigation techniques from Security Considerations here.

5.4. Connection Termination

Connections should remain open until they become idle for a pre-negotiated period of time. A QUIC connection, once established, can be terminated in one of three ways:

1. Explicit Shutdown: An endpoint sends a CONNECTION_CLOSE frame to the peer initiating a connection termination. An endpoint may

send a GOAWAY frame to the peer prior to a CONNECTION_CLOSE to indicate that the connection will soon be terminated. A GOAWAY frame signals to the peer that any active streams will continue to be processed, but the sender of the GOAWAY will not initiate any additional streams and will not accept any new incoming streams. On termination of the active streams, a CONNECTION_CLOSE may be sent. If an endpoint sends a CONNECTION_CLOSE frame while unterminated streams are active (no FIN bit or RST_STREAM frames have been sent or received for one or more streams), then the peer must assume that the streams were incomplete and were abnormally terminated.

2. Implicit Shutdown: The default idle timeout for a QUIC connection is 30 seconds, and is a required parameter (ICSL) in connection negotiation. The maximum is 10 minutes. If there is no network activity for the duration of the idle timeout, the connection is closed. By default a CONNECTION_CLOSE frame will be sent. A silent close option can be enabled when it is expensive to send an explicit close, such as mobile networks that must wake up the radio.

3. Abrupt Shutdown: An endpoint may send a Public Reset packet at any time during the connection to abruptly terminate an active connection. A Public Reset packet SHOULD only be used as a final recourse. Commonly, a public reset is expected to be sent when a packet on an established connection is received by an endpoint that is unable decrypt the packet. For instance, if a server reboots mid-connection and loses any cryptographic state associated with open connections, and then receives a packet on an open connection, it should send a Public Reset packet in return. (TODO: articulate rules around when a public reset should be sent.)

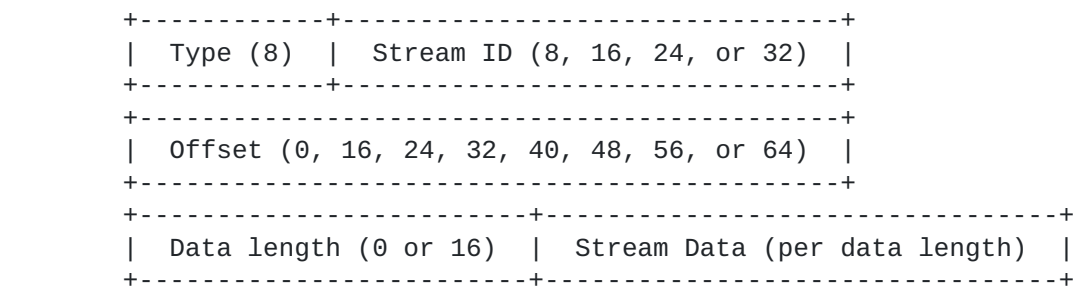
TODO: Connections that are terminated are added to a TIME_WAIT list at the server, so as to absorb any straggler packets in the network. Discuss TIME_WAIT list.

6. Frame Types and Formats

As described in Section XXX, Regular packets contain one or more frames. We now describe the various QUIC frame types that can be present in a Regular packet. The use of these frames and various frame header bits are described in subsequent sections.

6.1. STREAM Frame

STREAM frames implicitly create a stream and carry stream data. A STREAM frame is shown below.



The STREAM frame header fields are as follows:

- Frame Type: The Frame Type byte is an 8-bit value containing various flags, and is formatted as the following 8 bits: 1FDOOOSS.
  - The leftmost bit must be set to 1 indicating that this is a STREAM frame.
  - 'F' is the FIN bit, which is used for stream termination.
  - The 'D' bit indicates whether a Data Length field is present in the STREAM header. When set to 0, this field indicates that the Stream Data field extends to the end of the packet. When set to 1, this field indicates that Data Length field contains the length (in bytes) of the Stream Data field. The option to omit the length should only be used when the packet is a "full-sized" packet, to avoid the risk of corruption via padding.
  - The 'OOO' bits encode the length of the Offset header field as 0, 16, 24, 32, 40, 48, 56, or 64 bits long.
  - The 'SS' bits encode the length of the Stream ID header field as 8, 16, 24, or 32 bits. (DISCUSS: Consider making this 8, 16, 32, 64.)
- Stream ID: A variable-sized unsigned ID unique to this stream.
- Offset: A variable-sized unsigned number specifying the byte offset in the stream for the data in this STREAM frame. The first byte in the stream has an offset of 0.
- Data Length: An optional 16-bit unsigned number specifying the length of the Stream Data field in this STREAM frame.

A STREAM frame MUST have either non-zero data length or the FIN bit set.

Stream multiplexing is achieved by interleaving STREAM frames from multiple streams into one or more QUIC packets. A single QUIC packet MAY bundle STREAM frames from multiple streams.

Implementation note: One of the benefits of QUIC is avoidance of head-of-line blocking across multiple streams. When a packet loss occurs, only streams with data in that packet are blocked waiting for a retransmission to be received, while other streams can continue making progress. Note that when data from multiple streams is bundled into a single QUIC packet, loss of that packet blocks all those streams from making progress. An implementation is therefore advised to bundle as few streams as necessary in outgoing packets without losing transmission efficiency to underfilled packets.

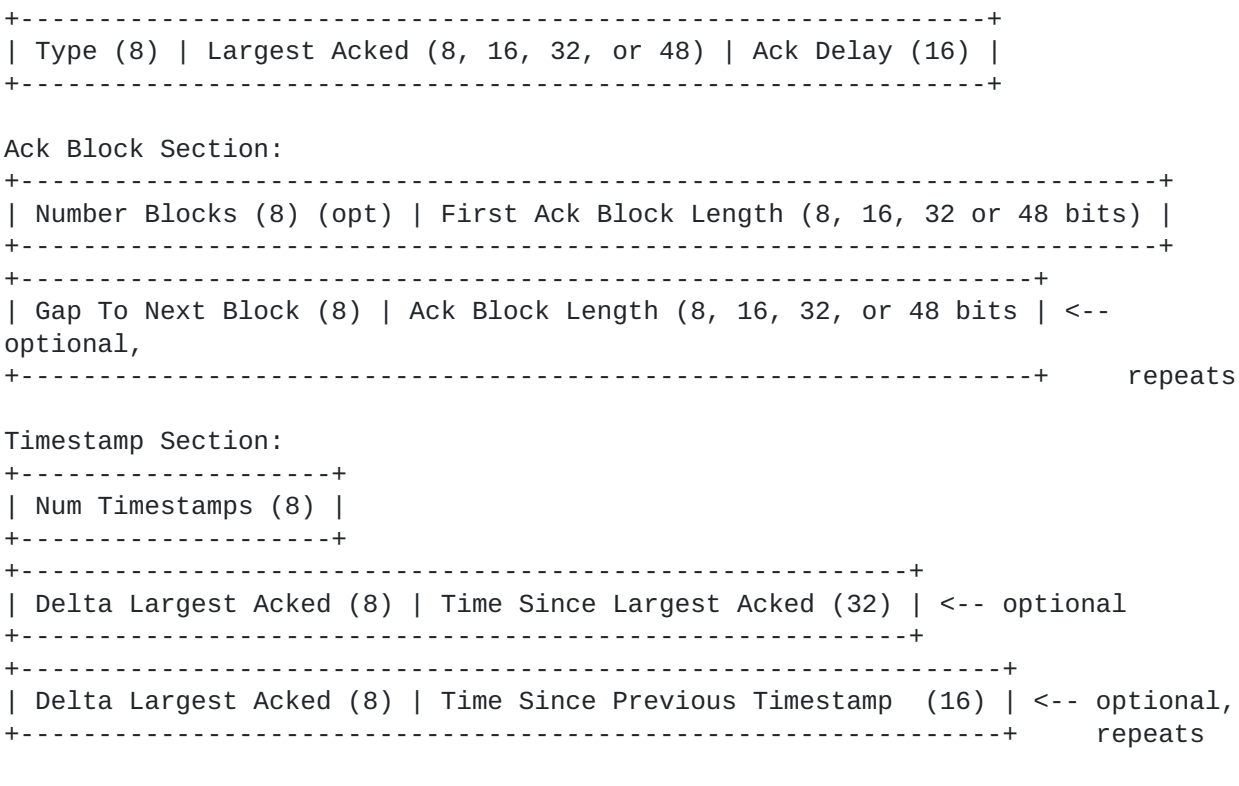
6.2. ACK Frame

Receivers send ACK frames to inform senders which packets they have received, as well as which packets are considered missing. The ACK frame contains between 1 and 256 ack blocks. Ack blocks are ranges of acknowledged packets.

To limit the ACK blocks to the ones that haven't yet been received by the sender, the sender periodically sends STOP_WAITING frames that signal the receiver to stop acking packets below a specified sequence number, raising the "least unacked" packet number at the receiver. A sender of an ACK frame thus reports only those ACK blocks between the received least unacked and the reported largest observed packet numbers. It is recommended for the sender to send the most recent largest acked packet it has received in an ack as the STOP_WAITING frame's least unacked value.

Unlike TCP SACKs, QUIC ACK blocks are irrevocable. Once a packet is acked, even if it does not appear in a future ack frame, it is assumed to be acked.

A sender MAY intentionally skip packet numbers to introduce entropy into the connection, to avoid opportunistic ack attacks. The sender MUST close the connection if an unsent packet number is acked. The format of the ACK frame is efficient at expressing blocks of missing packets; skipping packet numbers between 1 and 255 effectively provides up to 8 bits of efficient entropy on demand, which should be adequate protection against most opportunistic ack attacks.



The fields in the ACK frame are as follows:

- Frame Type: The Frame Type byte is an 8-bit value containing various flags. This byte is formatted as the following 8 bits: 01NULLMM.
  - The first two bits must be set to 01 indicating that this is an ACK frame.
  - The 'N' bit indicates whether the frame has more than 1 ack range.
  - The 'U' bit is unused.
  - The two 'LL' bits encode the length of the Largest Acked field as 1, 2, 4, or 6 bytes long.
  - The two 'MM' bits encode the length of the Ack Block Length fields as 1, 2, 4, or 6 bytes long.
- Largest Acked: A variable-sized unsigned value representing the largest packet number the peer is acking in this packet (typically the largest that the peer has seen thus far.)

- Ack Delay: Time from when the largest acked, as indicated in the Largest Acked field, was received by this peer to when this ack was sent.

- Ack Block Section:

  * Num Blocks (opt): An optional 8-bit unsigned value specifying the number of additional ack blocks (besides the required First Ack Block) in this ACK frame. Only present if the 'N' flag bit is 1.

  * First Ack Block Length: An unsigned packet number delta that indicates the number of contiguous additional packets being acked starting at the Largest Acked.

  * Gap To Next Block (opt, repeated): An unsigned number specifying the number of contiguous missing packets from the end of the previous ack block to the start of the next.

  * Ack Block Length (opt, repeated): An unsigned packet number delta that indicates the number of contiguous packets being acked starting after the end of the previous gap. Along with the previous field, this field is repeated "Num Blocks" times.

- Timestamp Section:

  * Num Timestamps: An unsigned 8-bit number specifying the total number of <packet number, timestamp> pairs following, including the First Timestamp.

  * Delta Largest Acked (opt): An optional 8-bit unsigned packet number delta specifying the delta between the largest acked and the first packet whose timestamp is being reported. In other words, this first packet number may be computed as (Largest Acked - Delta Largest Acked.)

  * First Timestamp (opt): An optional 32-bit unsigned value specifying the time delta in microseconds, from the beginning of the connection to the arrival of this packet.

  * Delta Largest Observed (opt, repeated): (Same as above.)

  * Time Since Previous Timestamp (opt, repeated): An optional 16-bit unsigned value specifying time delta from the previous reported timestamp. It is encoded in the same format as the Ack Delay. Along with the previous field, this field is repeated "Num Timestamps" times.

6.2.1. Time Format

DISCUSS_AND_REPLACE: Perhaps make this format simpler.

The time format used in the ACK frame above is a 16-bit unsigned float with 11 explicit bits of mantissa and 5 bits of explicit exponent, specifying time in microseconds. The bit format is loosely modeled after IEEE 754. For example, 1 microsecond is represented as 0x1, which has an exponent of zero, presented in the 5 high order bits, and mantissa of 1, presented in the 11 low order bits. When the explicit exponent is greater than zero, an implicit high-order 12th bit of 1 is assumed in the mantissa. For example, a floating value of 0x800 has an explicit exponent of 1, as well as an explicit mantissa of 0, but then has an effective mantissa of 4096 (12th bit is assumed to be 1). Additionally, the actual exponent is one-less than the explicit exponent, and the value represents 4096 microseconds. Any values larger than the representable range are clamped to 0xFFFF.

6.3. STOP_WAITING Frame

The STOP_WAITING frame is sent to inform the peer that it should not continue to wait for packets with packet numbers lower than a specified value. The packet number is encoded in 1, 2, 4 or 6 bytes, using the same coding length as is specified for the packet number for the enclosing packet's header (specified in the QUIC Frame packet's Flags field.) The frame is as follows:

```
+-----------------------------------------------------+
| Type (8) | Least unacked delta (8, 16, 32, or 48) |
+-----------------------------------------------------+
```

The fields in the STOP_WAITING frame are as follows:

- Frame Type: The Frame Type byte is an 8-bit value that must be set to 0x06 indicating that this is a STOP_WAITING frame.

- Least Unacked Delta: A variable-length packet number delta with the same length as the packet header's packet number. Subtract it from the complete packet number of the enclosing packet to determine the least unacked packet number. The resulting least unacked packet number is the earliest packet for which the sender is still awaiting an ack. If the receiver is missing any packets earlier than this packet, the receiver SHOULD consider those packets to be irrecoverably lost and MUST NOT report those packets as missing in subsequent acks.

6.4. WINDOW_UPDATE Frame

The WINDOW_UPDATE frame informs the peer of an increase in an endpoint's flow control receive window. The StreamID can be zero, indicating this WINDOW_UPDATE applies to the connection level flow control window, or non-zero, indicating that the specified stream should increase its flow control window. The frame is as follows:

The fields in the WINDOW_UPDATE frame are as follows:

- Frame Type: The Frame Type byte is an 8-bit value that must be set to 0x04 indicating that this is a WINDOW_UPDATE frame.

- Stream ID: ID of the stream whose flow control windows is being updated, or 0 to specify the connection-level flow control window.

- Byte offset: A 64-bit unsigned integer indicating the absolute byte offset of data which can be sent on the given stream. In the case of connection level flow control, the cumulative number of bytes which can be sent on all currently open streams.

6.5. BLOCKED Frame

A sender sends a BLOCKED frame when it is ready to send data (and has data to send), but is currently flow control blocked. BLOCKED frames are purely informational frames, but extremely useful for debugging purposes. A receiver of a BLOCKED frame should simply discard it (after possibly printing a helpful log message). The frame is as follows:

The fields in the BLOCKED frame are as follows:

- Frame Type: The Frame Type byte is an 8-bit value that must be set to 0x05 indicating that this is a BLOCKED frame.

- Stream ID: A 32-bit unsigned number indicating the stream which is flow control blocked. A non-zero Stream ID field specifies the stream that is flow control blocked. When zero, the Stream ID field indicates that the connection is flow control blocked.

6.6. RST_STREAM Frame

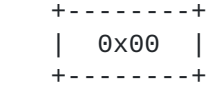
An endpoint may use a RST_STREAM frame to abruptly terminate a stream. The frame is as follows:

The fields are:

- Frame type: The Frame Type is an 8-bit value that must be set to 0x01 specifying that this is a RST_STREAM frame.
- Stream ID: The 32-bit Stream ID of the stream being terminated.
- Byte offset: A 64-bit unsigned integer indicating the absolute byte offset of the end of data written on this stream by the RST_STREAM sender.
- Error code: A 32-bit error code which indicates why the stream is being closed.

6.7. PADDING Frame

The PADDING frame pads a packet with 0x00 bytes. When this frame is encountered, the rest of the packet is expected to be padding bytes. The frame contains 0x00 bytes and extends to the end of the QUIC packet. A PADDING frame only has a Frame Type field, and must have the 8-bit Frame Type field set to 0x00.



6.8. PING frame

Endpoints can use PING frames to verify that their peers are still alive or to check reachability to the peer. The PING frame contains no payload. The receiver of a PING frame simply needs to ACK the packet containing this frame. The PING frame SHOULD be used to keep a connection alive when a stream is open. The default is to send a PING frame after 15 seconds of quiescence. A PING frame only has a Frame Type field, and must have the 8-bit Frame Type field set to 0x07.

```
   +--------+
   |  0x07  |
   +--------+
```

6.9. CONNECTION_CLOSE frame

An endpoint sends a CONNECTION_CLOSE frame to notify its peer that the connection is being closed. If there are open streams that haven't been explicitly closed, they are implicitly closed when the connection is closed. (Ideally, a GOAWAY frame would be sent with enough time that all streams are torn down.) The frame is as follows:

```
+-------------------------------------------------------------------------+
| Type(8) | Error code (32) | Reason phrase length (16) | Reason phrase |
+-------------------------------------------------------------------------+
```

The fields of a CONNECTION_CLOSE frame are as follows:

- Frame Type: An 8-bit value that must be set to 0x02 specifying that this is a CONNECTION_CLOSE frame.

- Error Code: A 32-bit error code which indicates the reason for closing this connection.

- Reason Phrase Length: A 16-bit unsigned number specifying the length of the reason phrase. This may be zero if the sender chooses to not give details beyond the QuicErrorCode.

- Reason Phrase: An optional human-readable explanation for why the connection was closed.

6.10. GOAWAY Frame

An endpoint may use a GOAWAY frame to notify its peer that the connection should stop being used, and will likely be aborted in the future. The endpoints will continue using any active streams, but the sender of the GOAWAY will not initiate any additional streams, and will not accept any new streams. The frame is as follows:

```
   +-----------------------------------------------------------+
   |  Type (8) |  Error code (32) |  Last Good Stream ID (32)  |
   +-----------------------------------------------------------+
   +-----------------------------------------------+
   | Reason phrase length (16)  |  Reason phrase  |
   +-----------------------------------------------+
```

The fields of a GOAWAY frame are as follows:

- Frame type: An 8-bit value that must be set to 0x03 specifying that this is a GOAWAY frame.

- Error Code: A 32-bit field error code which indicates the reason for closing this connection.

- Last Good Stream ID: The last Stream ID which was accepted by the sender of the GOAWAY message. If no streams were replied to, this value must be set to 0.

- Reason Phrase Length: A 16-bit unsigned number specifying the length of the reason phrase. This may be zero if the sender chooses to not give details beyond the error code.

- Reason Phrase: An optional human-readable explanation for why the connection was closed.

7. Packetization and Reliability

The maximum packet size for QUIC is the maximum size of the encrypted payload of the resulting UDP datagram. All QUIC packets SHOULD be sized to fit within the path's MTU to avoid IP fragmentation. The recommended default maximum packet size is 1350 bytes for IPv6 and 1370 bytes for IPv4. To optimize better, endpoints MAY use PLPMTUD [RFC4821] for detecting the path's MTU and setting the maximum packet size appropriately.

A sender bundles one or more frames in a Regular QUIC packet. A sender MAY bundle any set of frames in a packet. All QUIC packets MUST contain a packet number and MAY contain one or more frames (Section XX). Packet numbers MUST be unique within a connection and MUST NOT be reused within the same connection. Packet numbers MUST be assigned to packets in a strictly monotonically increasing order. The initial packet number used, at both the client and the server, MUST be 0. That is, the first packet in both directions of the connection MUST have a packet number of 0.

A sender SHOULD minimize per-packet bandwidth and computational costs by bundling as many frames as possible within a QUIC packet. A sender MAY wait for a short period of time to bundle multiple frames before sending a packet that is not maximally packed, to avoid sending out large numbers of small packets. An implementation may use heuristics about expected application sending behavior to determine whether and for how long to wait. This waiting period is an implementation decision, and an implementation should be careful to delay conservatively, since any delay is likely to increase application-visible latency.

Regular QUIC packets are "containers" of frames; a packet is never retransmitted whole, but frames in a lost packet may be rebundled and transmitted in a subsequent packet as necessary.

A packet may contain frames and/or application data, only some of which may require reliability. When a packet is detected as lost, the sender SHOULD only resend frames that require retransmission.

- All application data sent in STREAM frames MUST be retransmitted, with one exception. When an endpoint sends a RST_STREAM frame, data outstanding on that stream SHOULD NOT be retransmitted, since subsequent data on this stream is expected to not be delivered by the receiver.

- ACK, STOP_WAITING, and PADDING frames MUST NOT be retransmitted. New frames of these types may however be bundled with any outgoing packet.

- All other frames MUST be retransmitted.

Upon detecting losses, a sender MUST take appropriate congestion control action. The details of loss detection and congestion control are described in [draft-loss-recovery].

A receiver acknowledges receipt of a received packet by sending one or more ACK frames containing the packet number of the received packet. To avoid perpetual acking between endpoints, a receiver MUST NOT generate an ack in response to every packet containing only ACK frames. However, since it is possible that an endpoint sends only packets containing ACK frame (or other non-retransmittable frames), the receiving peer MAY send an ACK frame after a reasonable number (currently 20) of such packets have been received.

Strategies and implications of the frequency of generating acknowledgments are discussed in more detail in [draft-loss-recovery].

8. Streams: QUIC's Data Structuring Abstraction

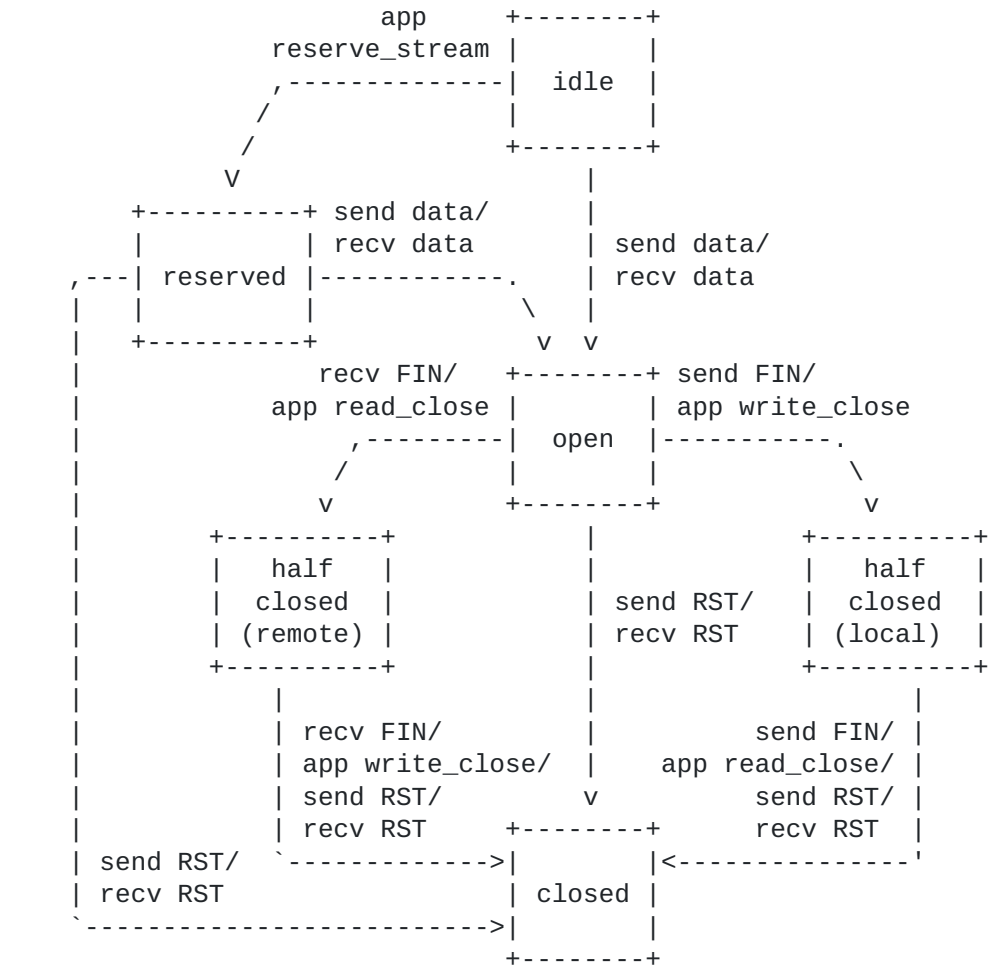
Streams in QUIC provide a lightweight, ordered, and bidirectional byte-stream abstraction. Streams can be created either by the client or the server, can concurrently send data interleaved with other streams, and can be cancelled. QUIC's stream lifetime is modeled closely after HTTP/2's [RFC7540]. Streams are independent of each other in delivery order. That is, data that is received on a stream is delivered in order within that stream, but there is no particular delivery order across streams. Transmit ordering among streams is left to the implementation. QUIC streams are considered lightweight

in that the creation and destruction of streams are expected to have minimal bandwidth and computational cost. A single STREAM frame may create, carry data for, and terminate a stream, or a stream may last the entire duration of a connection. Implementations are therefore advised to keep these extremes in mind and to implement stream creation and destruction to be as lightweight as possible.

An alternative view of QUIC streams is as an elastic "message" abstraction, similar to the way ephemeral streams are used in SST [cite SST], which may be a more appealing description for some applications.

8.1. Life of a Stream

The semantics of QUIC streams is based on HTTP/2 streams, and the lifecycle of a QUIC stream therefore closely follows that of an HTTP/2 stream [RFC7540], with some differences to accommodate the possibility of out-of-order delivery due to the use of multiple streams in QUIC. The lifecycle of a QUIC stream is shown in the following figure and described below.



send: endpoint sends this frame
recv: endpoint receives this frame

data: application data in a STREAM frame
FIN: FIN flag in a STREAM frame
RST: RST_STREAM frame

app: application API signals to QUIC
reserve_stream: causes a StreamID to be reserved for later use
read_close: causes stream to be half-closed without receiving a FIN
write_close: causes stream to be half-closed without sending a FIN

Note that this diagram shows stream state transitions and the frames and flags that affect those transitions only. For the purpose of state transitions, the FIN flag is processed as a separate event to the frame that bears it; a STREAM frame with the FIN flag set can cause two state transitions. When the FIN bit is sent on an empty

STREAM frame, the offset in the STREAM frame MUST be one greater than the last data byte sent on this stream.

Both endpoints have a subjective view of the state of a stream that could be different when frames are in transit. Endpoints do not coordinate the creation of streams; they are created unilaterally by either endpoint. The negative consequences of a mismatch in states are limited to the "closed" state after sending RST_STREAM, where frames might be received for some time after closing.

Streams have the following states:

8.1.1. idle

All streams start in the "idle" state.

The following transitions are valid from this state:

Sending or receiving a STREAM frame causes the stream to become "open". The stream identifier is selected as described in Section XX. The same STREAM frame can also cause a stream to immediately become "half-closed".

An application can reserve an idle stream for later use. The stream state for the reserved stream transitions to "reserved".

Receiving any frame other than STREAM or RST_STREAM on a stream in this state MUST be treated as a connection error (Section XX) of type YYYY.

8.1.2. reserved

A stream in this state has been reserved for later use by the application. In this state only the following transitions are possible:

- Sending or receiving a STREAM frame causes the stream to become "open".

- Sending or receiving a RST_STREAM frame causes the stream to become "closed".

8.1.3. open

A stream in the "open" state may be used by both peers to send frames of any type. In this state, a sending peer must observe the flow-control limit advertised by its receiving peer (Section XX).

From this state, either endpoint can send a frame with the FIN flag set, which causes the stream to transition into one of the "half-closed" states. An endpoint sending an FIN flag causes the stream state to become "half-closed (local)". An endpoint receiving a FIN flag causes the stream state to become "half-closed (remote)"; the receiving endpoint MUST NOT process the FIN flag until all preceding data on the stream has been received.

Either endpoint can send a RST_STREAM frame from this state, causing it to transition immediately to "closed".

8.1.4. half-closed (local)

A stream that is in the "half-closed (local)" state MUST NOT be used for sending STREAM frames; WINDOW_UPDATE and RST_STREAM MAY be sent in this state.

A stream transitions from this state to "closed" when a frame that contains an FIN flag is received or when either peer sends a RST_STREAM frame.

An endpoint can receive any type of frame in this state. Providing flow-control credit using WINDOW_UPDATE frames is necessary to continue receiving flow-controlled frames. In this state, a receiver MAY ignore WINDOW_UPDATE frames for this stream, which might arrive for a short period after a frame bearing the FIN flag is sent.

8.1.5. half-closed (remote)

A stream that is "half-closed (remote)" is no longer being used by the peer to send any data. In this state, a sender is no longer obligated to maintain a receiver stream-level flow-control window.

If an endpoint receives any STREAM frames for a stream that is in this state, it MUST close the connection with a QUIC_STREAM_DATA_AFTER_TERMINATION error (Section XX).

A stream in this state can be used by the endpoint to send frames of any type. In this state, the endpoint continues to observe advertised stream-level and connection-level flow-control limits (Section XX).

A stream can transition from this state to "closed" by sending a frame that contains a FIN flag or when either peer sends a RST_STREAM frame.

8.1.6. closed

The "closed" state is the terminal state.

A final offset is present in both a frame bearing a FIN flag and in a RST_STREAM frame. Upon sending either of these frames for a stream, the endpoint MUST NOT send a STREAM frame carrying data beyond the final offset.

An endpoint that receives any frame for this stream after receiving either a FIN flag and all stream data preceding it, or a RST_STREAM frame, MUST quietly discard the frame, with one exception. If a STREAM frame carrying data beyond the received final offset is received, the endpoint MUST close the connection with a QUIC_STREAM_DATA_AFTER_TERMINATION error (Section XX).

An endpoint that receives a RST_STREAM frame (and which has not sent a FIN or a RST_STREAM) MUST immediately respond with a RST_STREAM frame, and MUST NOT send any more data on the stream. This endpoint may continue receiving frames for the stream on which a RST_STREAM is received.

If this state is reached as a result of sending a RST_STREAM frame, the peer that receives the RST_STREAM might have already sent -- or enqueued for sending -- frames on the stream that cannot be withdrawn. An endpoint MUST ignore frames that it receives on closed streams after it has sent a RST_STREAM frame. An endpoint MAY choose to limit the period over which it ignores frames and treat frames that arrive after this time as being in error.

STREAM frames received after sending RST_STREAM are counted toward the connection and stream flow-control windows. Even though these frames might be ignored, because they are sent before their sender receives the RST_STREAM, the sender will consider the frames to count against its flow-control windows.

In the absence of more specific guidance elsewhere in this document, implementations SHOULD treat the receipt of a frame that is not expressly permitted in the description of a state as a connection error (Section XX). Frames of unknown types are ignored.

(TODO: QUIC_STREAM_NO_ERROR is a special case. Write it up.)

8.2. Stream Identifiers

Streams are identified by an unsigned 32-bit integer, referred to as the StreamID. To avoid StreamID collision, clients MUST initiate

streams usinge odd-numbered StreamIDs; streams initiated by the server MUST use even-numbered StreamIDs.

A StreamID of zero (0x0) is reserved and used for connection-level flow control frames (Section XX); the StreamID of zero cannot be used to establish a new stream.

StreamID 1 (0x1) is reserved for the crypto handshake. StreamID 1 MUST NOT be used for application data, and MUST be the first client-initiated stream.

Streams MUST be created or reserved in sequential order, but MAY be used in arbitrary order. A QUIC endpoint MUST NOT reuse a StreamID on a given connection.

8.3. Stream Concurrency

An endpoint can limit the number of concurrently active incoming streams by setting the MSPC parameter (see Section XX) in the transport parameters. The maximum concurrent streams setting is specific to each endpoint and applies only to the peer that receives the setting. That is, clients specify the maximum number of concurrent streams the server can initiate, and servers specify the maximum number of concurrent streams the client can initiate.

Streams that are in the "open" state or in either of the "half-closed" states count toward the maximum number of streams that an endpoint is permitted to open. Streams in any of these three states count toward the limit advertised in the MSPC setting.

Endpoints MUST NOT exceed the limit set by their peer. An endpoint that receives a STREAM frame that causes its advertised concurrent stream limit to be exceeded MUST treat this as a stream error of type QUIC_TOO_MANY_OPEN_STREAMS (Section XX).

8.4. Sending and Receiving Data

Once a stream is created, endpoints may use the stream to send and receive data. Each endpoint may send a series of STREAM frames encapsulating data on a stream until the stream is terminated in that direction. Streams are an ordered byte-stream abstraction, and they have no other structure within them. STREAM frame boundaries are not expected to be preserved in retransmissions from the sender or during delivery to the application at the receiver.

When new data is to be sent on a stream, a sender MUST set the encapsulating STREAM frame's offset field to the stream offset of the first byte of this new data. The first byte of data that is sent on

a stream has the stream offset 0. A receiver MUST ensure that received stream data is delivered to the application as an ordered byte-stream. Data received out of order MUST be buffered for later delivery, as long as it is not in violation of the receiver's flow control limits.

An endpoint MUST NOT send any stream data without consulting the congestion controller and the flow controller, with the following two exceptions.

- The crypto handshake stream, Stream 1, MUST NOT be subject to congestion control or connection-level flow control, but MUST be subject to stream-level flow control.

- An application MAY exclude specific stream IDs from connection-level flow control. If so, these streams MUST NOT be subject to connection-level flow control.

Flow control is described in detail in Section XX, and congestion control is described in the companion document [draft-iyengar-quic-loss-recovery].

9. Flow Control

It is necessary to limit the amount of data that a sender may have outstanding at any time, so as to prevent a fast sender from overwhelming a slow receiver, or to prevent a malicious sender from consuming significant resources at a receiver. This section describes QUIC's flow-control mechanisms.

QUIC employs a credit-based flow-control scheme similar to HTTP/2's flow control [RFC7540]. A receiver advertises the number of octets it is prepared to receive on a given stream and for the entire connection. This leads to two levels of flow control in QUIC: (i) Connection flow control, which prevents senders from exceeding a receiver's buffer capacity for the connection, and (ii) Stream flow control, which prevents a single stream from consuming the entire receive buffer for a connection.

A receiver sends WINDOW_UPDATE frames to the sender to advertise additional credit, for both connection and stream flow control. A receiver advertises the maximum absolute byte offset in the stream or in the connection which the receiver is willing to receive.

The initial flow control credit is 65536 bytes for both the stream and connection flow controllers.

A receiver MAY advertise a larger offset at any point in the connection by sending a WINDOW_UPDATE frame. A receiver MUST NOT renege on an advertisement; that is, once a receiver advertises an offset via a WINDOW_UPDATE frame, it MUST NOT subsequently advertise a smaller offset. A sender may receive WINDOW_UPDATE frames out of order; a sender MUST therefore ignore any reductions in flow control credit.

A sender MUST send BLOCKED frames to indicate it has data to write but is blocked by lack of connection or stream flow control credit. BLOCKED frames are expected to be sent infrequently in common cases, but they are considered useful for debugging and monitoring purposes.

A receiver advertises credit for a stream by sending a WINDOW_UPDATE frame with the StreamID set appropriately. A receiver may simply use the current received offset to determine the flow control offset to be advertised.

Connection flow control is a limit to the total bytes of stream data sent in STREAM frames. A receiver advertises credit for a connection by sending a WINDOW_UPDATE frame with the StreamID set to zero (0x00). A receiver may maintain a cumulative sum of bytes received cumulatively on all streams to determine the value of the connection flow control offset to be advertised in WINDOW_UPDATE frames. A sender may maintain a cumulative sum of stream data bytes sent to impose the connection flow control limit.

9.1. Edge Cases and Other Considerations

There are some edge cases which must be considered when dealing with stream and connection level flow control. Given enough time, both endpoints must agree on flow control state. If one end believes it can send more than the other end is willing to receive, the connection will be torn down when too much data arrives. Conversely if a sender believes it is blocked, while endpoint B expects more data can be received, then the connection can be in a deadlock, with the sender waiting for a WINDOW_UPDATE which will never come.

9.1.1. Mid-stream RST_STREAM

On receipt of an RST_STREAM frame, an endpoint will tear down state for the matching stream and ignore further data arriving on that stream. This could result in the endpoints getting out of sync, since the RST_STREAM frame may have arrived out of order and there may be further bytes in flight. The data sender would have counted the data against its connection level flow control budget, but a receiver that has not received these bytes would not know to include them as well. The receiver must learn of the number of bytes that

were sent on the stream to make the same adjustment in its connection flow controller.

To avoid this de-synchronization, a RST_STREAM sender MUST include the final byte offset sent on the stream in the RST_STREAM frame. On receiving a RST_STREAM frame, a receiver definitively knows how many bytes were sent on that stream before the RST_STREAM frame, and the receiver MUST use the final offset to account for all bytes sent on the stream in its connection level flow controller.

9.1.2. Response to a RST_STREAM

Since streams are bidirectional, a sender of a RST_STREAM needs to know how many bytes the peer has sent on the stream. If an endpoint receives a RST_STREAM frame and has sent neither a FIN nor a RST_STREAM, it MUST send a RST_STREAM in response, bearing the offset of the last byte sent on this stream as the final offset.

9.1.3. Offset Increment

This document leaves when and how many bytes to advertise in a WINDOW_UPDATE to the implementation, but offers a few considerations. WINDOW_UPDATE frames constitute overhead, and therefore, sending a WINDOW_UPDATE with small offset increments is undesirable. At the same time, sending WINDOW_UPDATES with large offset increments requires the sender to commit to that amount of buffer. Implementations must find the correct tradeoff between these sides to determine how large an offset increment to send in a WINDOW_UPDATE.

A receiver MAY use an autotuning mechanism to tune the size of the offset increment to advertise based on a roundtrip time estimate and the rate at which the receiving application consumes data, similar to common TCP implementations.

9.1.4. BLOCKED frames

If a sender does not receive a WINDOW_UPDATE frame when it has run out of flow control credit, the sender will be blocked and MUST send a BLOCKED frame. A BLOCKED frame is expected to be useful for debugging at the receiver. A receiver SHOULD NOT wait for a BLOCKED frame before sending with a WINDOW_UPDATE, since doing so will cause at least one roundtrip of quiescence. For smooth operation of the congestion controller, it is generally considered best to not let the sender go into quiescence if avoidable. To avoid blocking a sender, and to reasonably account for the possibiity of loss, a receiver should send a WINDOW_UPDATE frame at least two roundtrips before it expects the sender to get blocked.

10. Error Codes

This section lists all the QUIC error codes that may be used in a CONNECTION_CLOSE frame. TODO: Trim list and group errors for readabiity.

- 0x01: QUIC_INTERNAL_ERROR. (Connection has reached an invalid state.)
- 0x02: QUIC_STREAM_DATA_AFTER_TERMINATION. (There were data frames after the a fin or reset.)
- 0x03: QUIC_INVALID_PACKET_HEADER. (Control frame is malformed.)
- 0x04: QUIC_INVALID_FRAME_DATA. (Frame data is malformed.)
- 0x30: QUIC_MISSING_PAYLOAD. (The packet contained no payload.)
- 0x2e: QUIC_INVALID_STREAM_DATA. (STREAM frame data is malformed.)
- 0x57: QUIC_OVERLAPPING_STREAM_DATA. (STREAM frame data overlaps with buffered data.)
- 0x3d: QUIC_UNENCRYPTED_STREAM_DATA. (Received STREAM frame data is not encrypted.)
- 0x58: QUIC_ATTEMPT_TO_SEND_UNENCRYPTED_STREAM_DATA. (Attempt to send unencrypted STREAM frame. Not sent on the wire, used for local logging.)
- 0x59: QUIC_MAYBE_CORRUPTED_MEMORY. (Received a frame which is likely the result of memory corruption.)
- 0x06: QUIC_INVALID_RST_STREAM_DATA. (RST_STREAM frame data is malformed.)
- 0x07: QUIC_INVALID_CONNECTION_CLOSE_DATA. (CONNECTION_CLOSE frame data is malformed.)
- 0x08: QUIC_INVALID_GOAWAY_DATA. (GOAWAY frame data is malformed.)
- 0x39: QUIC_INVALID_WINDOW_UPDATE_DATA. (WINDOW_UPDATE frame data is malformed.)
- 0x3a: QUIC_INVALID_BLOCKED_DATA. (BLOCKED frame data is malformed.)

- 0x3c: QUIC_INVALID_STOP_WAITING_DATA. (STOP_WAITING frame data is malformed.)
- 0x4e: QUIC_INVALID_PATH_CLOSE_DATA. (PATH_CLOSE frame data is malformed.)
- 0x09: QUIC_INVALID_ACK_DATA. (ACK frame data is malformed.)
- 0x0a: QUIC_INVALID_VERSION_NEGOTIATION_PACKET. (Version negotiation packet is malformed.)
- 0x0b: QUIC_INVALID_PUBLIC_RST_PACKET. (Public RST packet is malformed.)
- 0x0c: QUIC_DECRYPTION_FAILURE. (There was an error decrypting.)
- 0x0d: QUIC_ENCRYPTION_FAILURE. (There was an error encrypting.)
- 0x0e: QUIC_PACKET_TOO_LARGE. (The packet exceeded kMaxPacketSize.)
- 0x10: QUIC_PEER_GOING_AWAY. (The peer is going away. May be a client or server.)
- 0x11: QUIC_INVALID_STREAM_ID. (A stream ID was invalid.)
- 0x31: QUIC_INVALID_PRIORITY. (A priority was invalid.)
- 0x12: QUIC_TOO_MANY_OPEN_STREAMS. (Too many streams already open.)
- 0x4c: QUIC_TOO_MANY_AVAILABLE_STREAMS. (The peer created too many available streams.)
- 0x13: QUIC_PUBLIC_RESET. (Received public reset for this connection.)
- 0x14: QUIC_INVALID_VERSION. (Invalid protocol version.)
- 0x16: QUIC_INVALID_HEADER_ID. (The Header ID for a stream was too far from the previous.)
- 0x17: QUIC_INVALID_NEGOTIATED_VALUE. (Negotiable parameter received during handshake had invalid value.)
- 0x18: QUIC_DECOMPRESSION_FAILURE. (There was an error decompressing data.)

- 0x19: QUIC_NETWORK_IDLE_TIMEOUT. (The connection timed out due to no network activity.)

- 0x43: QUIC_HANDSHAKE_TIMEOUT. (The connection timed out waiting for the handshake to complete.)

- 0x1a: QUIC_ERROR_MIGRATING_ADDRESS. (There was an error encountered migrating addresses.)

- 0x56: QUIC_ERROR_MIGRATING_PORT. (There was an error encountered migrating port only.)

- 0x1b: QUIC_PACKET_WRITE_ERROR. (There was an error while writing to the socket.)

- 0x33: QUIC_PACKET_READ_ERROR. (There was an error while reading from the socket.)

- 0x32: QUIC_EMPTY_STREAM_FRAME_NO_FIN. (We received a STREAM_FRAME with no data and no fin flag set.)

- 0x38: QUIC_INVALID_HEADERS_STREAM_DATA. (We received invalid data on the headers stream.)

- 0x3b: QUIC_FLOW_CONTROL_RECEIVED_TOO_MUCH_DATA. (The peer received too much data, violating flow control.)

- 0x3f: QUIC_FLOW_CONTROL_SENT_TOO_MUCH_DATA. (The peer sent too much data, violating flow control.)

- 0x40: QUIC_FLOW_CONTROL_INVALID_WINDOW. (The peer received an invalid flow control window.)

- 0x3e: QUIC_CONNECTION_IP_POOLED. (The connection has been IP pooled into an existing connection.)

- 0x44: QUIC_TOO_MANY_OUTSTANDING_SENT_PACKETS. (The connection has too many outstanding sent packets.)

- 0x45: QUIC_TOO_MANY_OUTSTANDING_RECEIVED_PACKETS. (The connection has too many outstanding received packets.)

- 0x46: QUIC_CONNECTION_CANCELLED. (The quic connection has been cancelled.)

- 0x47: QUIC_BAD_PACKET_LOSS_RATE. (Disabled QUIC because of high packet loss rate.)

- 0x49: QUIC_PUBLIC_RESETS_POST_HANDSHAKE. (Disabled QUIC because of too many PUBLIC_RESETs post handshake.)

- 0x4a: QUIC_TIMEOUTS_WITH_OPEN_STREAMS. (Disabled QUIC because of too many timeouts with streams open.)

- 0x4b: QUIC_FAILED_TO_SERIALIZE_PACKET. (Closed because we failed to serialize a packet.)

- 0x55: QUIC_TOO_MANY_RTOS. (QUIC timed out after too many RTOs.)

- 0x1c: QUIC_HANDSHAKE_FAILED. (Crypto errors.Hanshake failed.)

- 0x1d: QUIC_CRYPTO_TAGS_OUT_OF_ORDER. (Handshake message contained out of order tags.)

- 0x1e: QUIC_CRYPTO_TOO_MANY_ENTRIES. (Handshake message contained too many entries.)

- 0x1f: QUIC_CRYPTO_INVALID_VALUE_LENGTH. (Handshake message contained an invalid value length.)

- 0x20: QUIC_CRYPTO_MESSAGE_AFTER_HANDSHAKE_COMPLETE. (A crypto message was received after the handshake was complete.)

- 0x21: QUIC_INVALID_CRYPTO_MESSAGE_TYPE. (A crypto message was received with an illegal message tag.)

- 0x22: QUIC_INVALID_CRYPTO_MESSAGE_PARAMETER. (A crypto message was received with an illegal parameter.)

- 0x34: QUIC_INVALID_CHANNEL_ID_SIGNATURE. (An invalid channel id signature was supplied.)

- 0x23: QUIC_CRYPTO_MESSAGE_PARAMETER_NOT_FOUND. (A crypto message was received with a mandatory parameter missing.)

- 0x24: QUIC_CRYPTO_MESSAGE_PARAMETER_NO_OVERLAP. (A crypto message was received with a parameter that has no overlapwith the local parameter.)

- 0x25: QUIC_CRYPTO_MESSAGE_INDEX_NOT_FOUND. (A crypto message was received that contained a parameter with too fewvalues.)

- 0x5e: QUIC_UNSUPPORTED_PROOF_DEMAND. (A demand for an unsupport proof type was received.)

- 0x26: QUIC_CRYPTO_INTERNAL_ERROR. (An internal error occured in crypto processing.)

- 0x27: QUIC_CRYPTO_VERSION_NOT_SUPPORTED. (A crypto handshake message specified an unsupported version.)

- 0x48: QUIC_CRYPTO_HANDSHAKE_STATELESS_REJECT. (A crypto handshake message resulted in a stateless reject.)

- 0x28: QUIC_CRYPTO_NO_SUPPORT. (There was no intersection between the crypto primitives supported by thepeer and ourselves.)

- 0x29: QUIC_CRYPTO_TOO_MANY_REJECTS. (The server rejected our client hello messages too many times.)

- 0x2a: QUIC_PROOF_INVALID. (The client rejected the server's certificate chain or signature.)

- 0x2b: QUIC_CRYPTO_DUPLICATE_TAG. (A crypto message was received with a duplicate tag.)

- 0x2c: QUIC_CRYPTO_ENCRYPTION_LEVEL_INCORRECT. (A crypto message was received with the wrong encryption level (i.e. itshould have been encrypted but was not.))

- 0x2d: QUIC_CRYPTO_SERVER_CONFIG_EXPIRED. (The server config for a server has expired.)

- 0x35: QUIC_CRYPTO_SYMMETRIC_KEY_SETUP_FAILED. (We failed to setup the symmetric keys for a connection.)

- 0x36: QUIC_CRYPTO_MESSAGE_WHILE_VALIDATING_CLIENT_HELLO. (A handshake message arrived, but we are still validating theprevious handshake message.)

- 0x41: QUIC_CRYPTO_UPDATE_BEFORE_HANDSHAKE_COMPLETE. (A server config update arrived before the handshake is complete.)

- 0x5a: QUIC_CRYPTO_CHLO_TOO_LARGE. (CHLO cannot fit in one packet.)

- 0x37: QUIC_VERSION_NEGOTIATION_MISMATCH. (This connection involved a version negotiation which appears to have beentampered with.)

- 0x50: QUIC_IP_ADDRESS_CHANGED. (IP address changed causing connection close.)

- 0x51: QUIC_CONNECTION_MIGRATION_NO_MIGRATABLE_STREAMS. (Connection migration errors.Network changed, but connection had no migratable streams.)

- 0x52: QUIC_CONNECTION_MIGRATION_TOO_MANY_CHANGES. (Connection changed networks too many times.)

- 0x53: QUIC_CONNECTION_MIGRATION_NO_NEW_NETWORK. (Connection migration was attempted, but there was no new network tomigrate to.)

- 0x54: QUIC_CONNECTION_MIGRATION_NON_MIGRATABLE_STREAM. (Network changed, but connection had one or more non-migratable streams.)

- 0x5d: QUIC_TOO_MANY_FRAME_GAPS. (Stream frames arrived too discontiguously so that stream sequencer buffermaintains too many gaps.)

- 0x5f: QUIC_STREAM_SEQUENCER_INVALID_STATE. (Sequencer buffer get into weird state where continuing read/write will leadto crash.)

- 0x60: QUIC_TOO_MANY_SESSIONS_ON_SERVER. (Connection closed because of server hits max number of sessions allowed.

11. Security and Privacy Considerations

11.1. Spoofed Ack Attack

An attacker receives an STK from the server and then releases the IP address on which it received the STK. The attacked may in the future, spoof this same address (which now presumably addresses a different endpoint), and initiates a 0-RTT connection with a server on the victim's behalf. The attacker then spoofs ack packets to the server which cause the server to potentially drown the victim in data.

There are two possible mitigations to this attack. The simplest one is that a server can unilaterally create a gap in packet-number space. In the non-attack scenario, the client will send an ack with a larger largest acked. In the attack scenario, the attacker may ack a packet in the gap. If the server sees an ack for a packet that was never sent, the connection can be aborted.

The second mitigation is that the server can require that acks for sent packets match the encryption level of the sent packet. This mitigation is useful if the connection has an ephemeral forward-secure key that is generated and used for every new connection. If a packet sent is encrypted with a forward-secure key, then any acks

that are received for them must also be forward-secure encrypted. Since the attacker will not have the forward secure key, the attacker will not be able to generate forward-secure encrypted ack packets.

12. Contributors

This protocol is the outcome of work by many engineers, not just the authors of this document. The design and rationale behind QUIC draw significantly from work by Jim Roskind [1]. In alphabetical order, the contributors to the project are: Britt Cyr, Jeremy Dorfman, Ryan Hamilton, Jana Iyengar, Fedor Kouranov, Charles Krasic, Jo Kulik, Adam Langley, Jim Roskind, Robbie Shade, Satyam Shekhar, Cherie Shi, Ian Swett, Raman Tenneti, Victor Vasiliev, Antonio Vicente, Patrik Westin, Alyssa Wilk, Dale Worley, Fan Yang, Dan Zhang, Daniel Ziegler.

13. Acknowledgments

Special thanks are due to the following for helping shape QUIC and its deployment: Chris Bentzel, Misha Efimov, Roberto Peon, Alistair Riddoch, Siddharth Vijayakrishnan, and Assar Westerlund. QUIC has also benefited immensely from discussions with folks in private conversations and public ones on the proto-quic@chromium.org mailing list.

.

14. References

14.1. Normative References

[RFC2119] Bradner, S., "Key Words for use in RFCs to Indicate Requirement Levels", March 1997.

[draft-thomson-quic-tls] Thomson, M. and R. Hamilton, "Porting QUIC to TLS", March 2016.

[draft-iyengar-quic-loss-recovery] Iyengar, J. and I. Swett, "QUIC Loss Recovery and Congestion Control", July 2016.

14.2. Informative References

[RFC7540] Belshe, M., Peon, R., and M. Thomson, "Hypertext Transfer Protocol Version 2 (HTTP/2)", May 2015.

[QUIC-CRYPTO]
Langley, A. and W. Chang, "QUIC Crypto", June 2015,
<http://goo.gl/OuVSxa>.

14.3. URIs

[1] https://goo.gl/dMVtFi

Authors' Addresses

Ryan Hamilton
Google

Email: rch@google.com

Janardhan Iyengar
Google

Email: jri@google.com

Ian Swett
Google

Email: ianswett@google.com

Alyssa Wilk
Google

Email: alyssar@google.com