Same Standards, Different Decisions: A Study of QUIC and HTTP/3 Implementation Diversity

Robin Marx
Hasselt University – tUL – EDM
Diepenbeek, Belgium
robin.marx@uhasselt.be

Joris Herbots
Wim Lamotte
Hasselt University – tUL – EDM
Diepenbeek, Belgium
{first.last}@uhasselt.be

Peter Quax
Hasselt University – tUL – Flanders
Make - EDM
Diepenbeek, Belgium
peter.quax@uhasselt.be

	aioquic	google	lsquic	mvfst	ngtcp2	picoquic	quic-go	quiche	quicly	quinn
Flow Control category (FC)	2	1	1	1	1	2	1	3	1	1
Multiplexing scheduler	SEQ	RR	RR	RR	SEQ	SEQ	RR	RR	RR	RR
Retransmission approach (RA)	2	1	2	3	2	2	2	1	4	2
0 RTT approach (ZR)	1	1	2	3	1	2	2	1	2	1
DATA frame size	large	medium	small	large	small	large	large	small	large	small
Worst case packetization goodput efficiency	90.34%	95.02%	92.54%	91.42%	90.88%	87.94%			91.52%	83.92%
Dynamic packet sizing (PMTUD)	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
Acknowledgment frequency (#packets)	2-8	2-10	2-8	10	2-4	2-6	2-9	1-38	2	1-17
Congestion Control (CC) New Reno | Cubic | BBRv1	✓|✗|✗	✗|✓|✓	✗|✓|✓	✓|✓|✓	✓|✗|✗	✓|✓|✓	✓|✗|✗	✓|✓|✗	✓|✗|✗	✓|✗|✗

Table 1: Selective behavioral comparison of 10 prevailing IETF QUIC implementations (May 2020). Empty slots indicate no results for this data point. SEQ = Sequential, RR = Round-Robin. Small = <100kB, medium = >100kB - <1MB, large = >1MB

ABSTRACT

The QUIC and HTTP/3 protocols are quickly maturing together with their implementations, though many of their low-level behaviours are not yet well-understood. To help improve this, we empirically compare 15 IETF QUIC+HTTP/3 implementations for advanced features like Flow and Congestion Control, 0-RTT, Multiplexing, and Packetization. We find a large heterogeneity between stacks, discuss uncovered bugs and conclude that most implementations are not fully optimized or validated yet. We argue that future work must prioritize rigorous root-cause analysis of observed behaviours, and show this is possible by employing our qlog and qvis tools.

EPIQ'20, August 10–14, 2020, Virtual Event, NY, USA

ACM ISBN 978-1-4503-8047-8/20/08…$15.00
https://doi.org/10.1145/3405796.3405828

CCS CONCEPTS

• **Networks** → **Transport protocols**; **Protocol testing and verification**; **Network protocol design**.

KEYWORDS

QUIC; HTTP/3; Transport Protocol; 0-RTT; Multiplexing

ACM Reference Format:
Robin Marx, Joris Herbots, Wim Lamotte, and Peter Quax. 2020. Same Standards, Different Decisions: A Study of QUIC and HTTP/3 Implementation Diversity. In *Workshop on Evolution, Performance, and Interoperability of QUIC (EPIQ'20), August 10–14, 2020, Virtual Event, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3405796.3405828

1 INTRODUCTION & MOTIVATION

In 2020, after nearly four years, the new QUIC and HTTP/3 (H3) protocol specifications [8, 23] are finally nearing completion. This long period is a testament to their complexity, as they combine decades of best practices, lessons learned from TCP, SCTP and HTTP/2 (H2), and advanced new features (like zero Round-Trip-Time (RTT) connection establishment) into a new Web protocol suite. To help verify

that the protocols' design choices actually hold up in practice and to prepare for deployment, several parties have been continuously updating over 18 different QUIC/H3 implementations [2]. These stacks are regularly tested on their so-called "interoperability", whereby clients from one implementation test features of servers from other codebases. This is done both manually and automatically in projects such as QUIC Tracker and QUIC Interop Runner [4, 37]. Despite this, bugs are still regularly uncovered (several by our research) and more advanced features are often not yet well supported or finetuned.

This is partly because existing tests mainly consider compatibility of the protocols' binary wire image and the mandatory parts of the specifications (i.e., MUST and MUST NOT). There are however many protocol features and situations for which the guidelines are much less clear and up to the developer's choice. These features, such as Flow Control, Congestion Control, Data Multiplexing, Packetization, and 0 RTT are often more difficult to evaluate in an automated fashion, yet arguably can have a large impact on protocol performance and behaviour. A good motivating example of this can be found in H2's highly complex Prioritization setup [45], which controls how bandwidth is distributed across concurrent Web page resource downloads (§3.2). This system was added late in H2's design and poorly validated prior to deployment. Consequently, even today, 5 years after the protocol's standardization, many H2 servers and clients do not properly support this feature [15, 32, 45], and it was decided to fully redesign this for H3 [34]. As such, we feel it is imperative to evaluate implementations of these more loosely defined features.

While there is prior academic work that evaluates some of these aspects [9, 12, 33], most of it is older and outdated, as it considers Google's initial QUIC version (gQUIC) [26] (while gQUIC and IETF QUIC are similar in concepts, their implementation details are fundamentally divergent). Newer work on IETF QUIC does exist [35, 36, 39], but is relatively rare despite the protocol's potential impact on the field. We believe this low academic involvement is partially due to QUIC's rapid evolution/unstableness [24], but also because of its high complexity, making it more difficult to evaluate correctly. This was also remarked on by several critical examinations, which have shown that earlier work often lacks scientific rigor and proper root-cause analysis. For example, Kakhki et al. [24] provide an in-depth discussion on the methodology of four previous papers. They find that prior work often miscalibrated their QUIC implementations, producing "misleading reports of poor QUIC performance". They also show that earlier work "conflates the impact of different workloads on QUIC performance" and only "speculates on the reasons for observed behaviour", rather than uncovering the low-level mechanisms. More recently, Wolsing et al. [46] looked at studies comparing TCP's to QUIC's performance and found that "*all* previous work compares an [unoptimized] out-of-the-box TCP with a highly tuned QUIC Web stack", meaning those works are biased and "do not shed light on the performance of current web stacks". Consequently, we feel researchers need to be able to properly configure their employed implementations and evaluate whether they behave as expected.

We however identified this problem early, recognizing the need for advanced QUIC debugging and analysis techniques. To this end, in 2018 we proposed both qlog and qvis [29]. qlog is a standard JSON-based endpoint logging format [31]. It allows implementations to log critical internal state and debugging information in a structured way, making it more powerful than typical approaches based on packet capture formats. The qlog logging format has found broad uptake in the IETF QUIC community, with 12/18 implementations currently outputting qlog [2] and Facebook logging up to 30 billion qlog events per day in production [30]. On the other end, qvis is an open-source online toolsuite [27] which ingests qlog files to produce powerful interactive visualizations that help analyze several complex protocol features. The qvis tools are being used by many QUIC developers to debug and verify their implementations [13, 30, 41].

As such, as the protocols, their implementations and our tools are reaching maturity, we feel it is now finally time to investigate if they are ready to be deployed, researched and evaluated. In this work, we use qlog and qvis to assess implementation differences between and maturity of 15 of the 18 active IETF QUIC and H3 implementations [2]. Our results for 10 of these stacks are summarized in Table 1. As other previously mentioned projects target interoperability testing [4, 37], we instead focus on protocol aspects that are difficult to automatically measure and that are expected to have a high measure of heterogeneity across implementations (see §3). We indeed identify large differences between implementations and find that many advanced features are not yet finished, tuned or validated in many stacks. Still, we conclude that with the powerful qlog and qvis tooling, proper analysis of implementation behaviour is possible, and thus, with care, researchers can start evaluating IETF QUIC.

2 EXPERIMENTAL METHODOLOGY

To deeply evaluate 15 QUIC implementations in a manageable amount of time, we rely heavily on the structured qlog format [31]. As 12 QUIC stacks output qlog, it is feasible to have always at least one end (and often both) of a cross-implementation connection outputting this format. We then both automatically process qlogs with scripts, and evaluate them manually via the qvis visualizations [27].

We use two main qlog sources. Firstly, to assess client-side behaviours, we mostly use the results from the QUIC interop runner [4]. This framework employs an ns-3 [3] network emulation setup between dockerized versions of 10 different QUIC stacks. While these tests do not explicitly consider our targeted behaviours, some of them can be used to obtain the insights we require (e.g., concurrent file transfer tests show Flow Control limits §3.1 and ACK frequencies §3.4). Secondly, to observe server-side behaviours, we use the python-based aioquic implementation [1]. Aioquic is chosen because it supports a wide range of QUIC+H3 features and consistently achieves the highest scores in existing interoperability tests. We adapt the aioquic client slightly [1] to more easily vary configuration parameters and drive automated test runs against QUIC servers. We do not run these servers ourselves, but instead make use of the fact that most implementers already provide public Internet endpoints for manual interoperability tests. This allows us to test even non-open source servers and lets us compare configurations between different endpoints backed by the same implementation. For example, the mvfst stack is deployed on a test server and also on Facebook.com, and both setups show marked differences. To eliminate behavioural artefacts due to real network variations, we run our tests a minimum of 5 times on two different Belgian WAN networks: first the Hasselt University network (1 Gbps downlink/10Mbps up) and second a residential Wi-Fi network (35Mbps/2Mbps).

We were unable to test all targeted features across all QUIC implementations. Firstly, 3/18 stacks were not considered, as they are

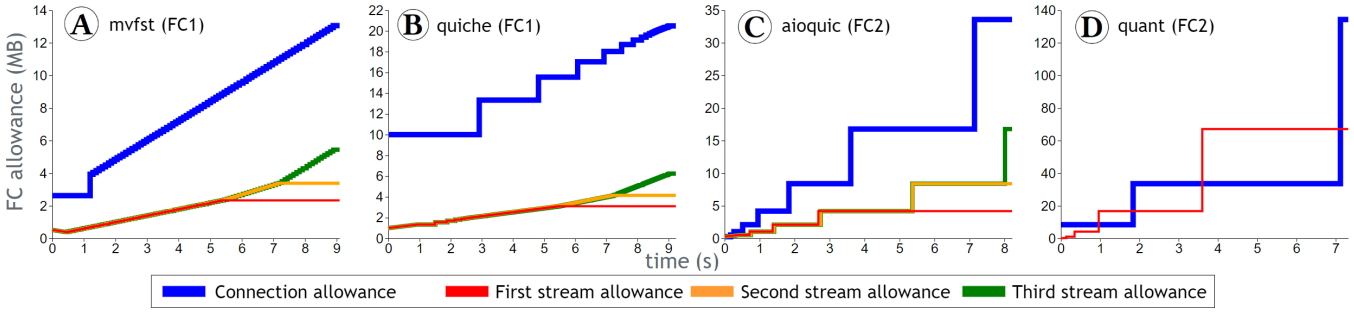

Figure 1: Connection and Stream-level Flow Control allowances for 4 QUIC stacks. A, B and C show concurrent downloads of 3 files (2MB, 3MB, 5MB). D shows a single 10MB download.

not open source, do not provide an endpoint, and/or are not mature enough. For the remaining 15, we focus on the 10 most feature-complete and open source stacks (see Table 1). The other 5 were tested to the extent possible. 8/15 stacks are backed by larger companies, while 7 are from hobbyists or individual implementers [2]. In §3 we indicate the amount of stacks evaluated for each feature. After both automated and manual analysis we further validate our results. Firstly, by performing source code reviews where possible. Secondly, by asking each stack's main implementers to confirm and comment on our conclusions, via the quicdev Slack group [5]. As such, almost none of the results presented here are conjecture, as most have been explicitly validated by their original developers.

Our results were gathered intermittently over a 4-month period (Jan-Apr 2020) and on IETF QUIC draft versions 25-27. As several implementations changed their behaviours over time (partially due to insights from our results), the altered stacks were re-tested in May 2020. Source code for all our tools, full result analysis sheets, source qlog files and other artefacts can be found at https://qlog.edm.uhasselt.be/epiq.

3 RESULTS

3.1 Flow Control

When downloading, an endpoint must reserve a transport-level receive buffer to store incoming data, both because data can arrive out-of-order (but can only be delivered to the application layer in-order) and because the speed at which the application reads from the transport can be lower than the network bandwidth. To prevent overshooting this buffer's capacity, endpoints utilize a Flow Control (FC) system to have the sender match its transmission rate to the speed at which the receive buffer can be emptied. For TCP, which abstracts transported data as a single, ordered byte stream, its singular "receive window" bounds the active bytes in flight allowance and grows and shrinks over time (e.g., a receive window of 0 means a sender should stop sending). In contrast, QUIC allows multiple concurrently active data streams (§3.2), and thus also defines a per-stream FC allowance, in addition to a connection-wide limit. QUIC's limits are expressed in maximum byte stream offsets [23], meaning they can never shrink and only grow in absolute values. Updates to these limits are communicated in MAX(_STREAM)_DATA frames, yet it is up to the developer to decide on the frequency of and allowance amount included in these frames. This is an important aspect to get right, as too few or too low limit updates can stall a fast sender, even if the receive buffer is not fully occupied as the receiver's updates take a RTT to reach the sender [23]. We identify three main approaches (see also Figure 1):

FC1 static allowance Ⓐ: the receive buffer size stays unchanged and the maximum allowance increases linearly. This requires a fixed amount of memory, but can cause stalls if the updates are delayed.

FC2 growing allowance Ⓒ: the receive buffer size grows over time, causing a non-linear relationship. This reduces the risk of being FC stalled, but requires more memory.

FC3 autotuning: the receive buffer size is dynamic, based on RTT estimates and application data consumption rate [40, 44]. This balances memory requirements against the potential for stalls.

Interestingly, we find only one QUIC stack (quiche) employs the more advanced FC3. Just 3/12 do FC2, while most of the 8/12 stacks doing FC1 simply update their absolute FC limits by adding the static buffer size once the receiving application has consumed 50% of the incoming data. Some however do exhibit interesting variations. Firstly, quiche initially used FC1 and updated at the 50% mark, but did not add the full buffer size. Instead it added the amount of bytes the application had consumed, meaning that with every update the allowance *increase* was halved. This in turn led to an increasing update frequency to keep the total allowance static (see Ⓑ), which is bad for goodput efficiency (see §3.3). After reviewing our results, they updated to FC3 [14]. Secondly, quant uses FC1, but allows stream-level allowance to grow beyond the connection-level limit (see Ⓓ), which could stall fast senders on this connection-level allowance.

By the implementers' own admission, the presence of these weird behaviours and absence of smarter schemes, is because most have not yet spent time fine-tuning FC approaches and memory requirements. To prevent stalling the sender, many simply set high initial allowances (e.g., 10MB in Ⓑ, 15MB in Google Chrome) and update early (the 50% mark). Several even asked us for guidance in choosing a better FC approach. However, as QUIC is fundamentally different from TCP in this respect, it is difficult to assess which scheme works best in practice. Facebook's approach (see Ⓐ) does give us an indication, as they have tweaked their behaviour in a real-life deployment. However their setup is also not foolproof (see §3.2), they are biased towards their specific use-case (loading content in native apps), and indicate being limited by existing application layer logic that was originally tweaked for TCP+H2 (e.g., setting higher initial FC limits would cause the app to aggressively preload resources, causing bandwidth contention). In all, we can say QUIC FC is an open problem, though a variant of FC3 is expected to work best.

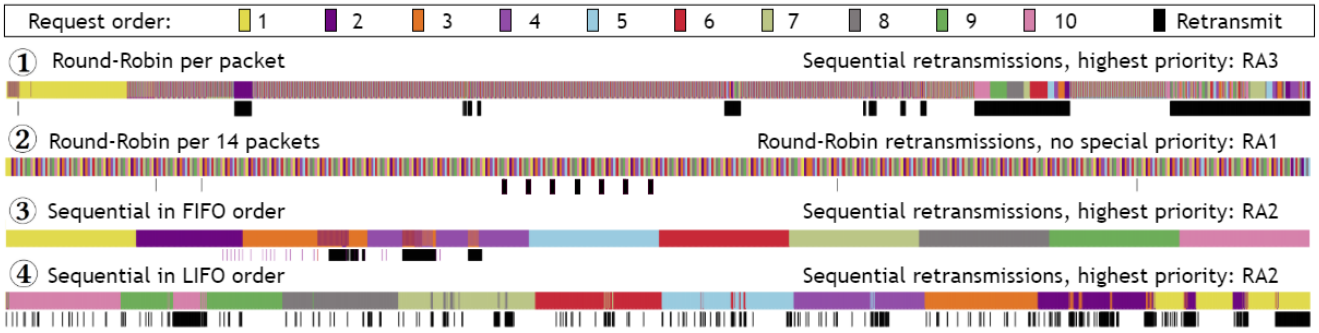

Figure 2: Multiplexing behaviour across different QUIC stacks when downloading 10 1MB files in parallel. Each small colored rectangle is one packet belonging to a file. Long colored areas indicate sequential scheduling. Black areas indicate which frames above them contain retransmitted data. Data arrives from left to right.

3.2 Multiplexing & Prioritization

TCP abstracts its connection as a single, fully ordered and reliable byte stream. This does not perform optimally in situations where multiple, independent data streams can be in progress at the same time (e.g., loading a Web page's resources). H2 attempted to get around this by defining the concept of concurrent byte streams at the application layer, yet this still mapped badly to TCP's single stream viewpoint (e.g., the Head-of-Line Blocking problem [28]). This is one of the motivations behind QUIC, which instead makes streams first-class citizens in the transport layer. A crucial aspect of handling multiple concurrently active byte streams is how to divide a sender's available bandwidth among them. This can be done in two main ways, see Figure 2. Firstly, a Round-Robin (RR) scheduler (①,②) divides bandwidth among various streams (either fairly or with different stream weights) by splitting resources into smaller chunks and interleaving them. Secondly, a sequential scheduler (③,④) sends all (available) data for a single stream before allocating bandwidth to the next. The optimal approach often depends on the application semantics. For example, both H2 and H3 use a "prioritization" system to drive this behaviour and sequential scheduling is thought to work best for Web page loading performance [28, 45]. We did not yet evaluate H3-level prioritization, as it entails a radical departure from H2's approach and is not fully defined or mature at this time [34], which is reflected by only 5/18 providing highly experimental support. Instead, since QUIC is supposed to be a general purpose transport protocol, we evaluated whether stacks provide sensible default transport-layer multiplexing, as the QUIC texts leave it fully up to the developer to determine what this behaviour should be. This is clearly visible in the default approaches taken by the different stacks. We find 9/13 stacks to employ a form of RR (6/9 switching streams each QUIC packet, 2/9 switching every 4-10 packets, and 1 unexpectedly switching only after filling the current congestion window (§3.4)). 4/13 stacks opt for a sequential variant instead, though we originally found 3 of them to erroneously sending data in Last-In First-Out (LIFO) order ④, typically thought to be a worst-case approach [42] (2/3 have since changed their approach to FIFO). The optimal approach is however more difficult to determine at this time and requires further study.

A peculiar interaction between Flow Control (FC, §3.1) and stream scheduling was observed when downloading 10 concurrent 1000000 byte (1MB) files from the mvfst server ①. There, a clearly anomalous sequential period is visible for the yellow (first) stream. This was due to our aioquic testclient setting both the connection and stream-level initial FC limits to 1048576 bytes (1MiB). mvfst processes requests 1-by-1 and fully buffered the first file. For the second request, only 48576 bytes remained of the connection FC limit, so only that much was prepared, while requests 3-10 were placed on-hold. When the RR scheduler kicked in, stream 1 was initially multiplexed with stream 2. Soon the stream 2 data ran out and the scheduler had only stream 1 data available, until the client's connection FC update allowed the server to buffer more stream data. It is clear from this example that FC can have (unintended) impactful interactions with stream multiplexing and can hinder prioritization efforts.

A final aspect is how QUIC arranges retransmissions. As TCP's single byte stream abstraction is fully ordered, its retransmissions are always given the absolute highest precedence. However, QUIC's per-stream loss tracking and delivery ordering means that retransmissions can be scheduled much like "new" stream data. Conceptually, we can define 4 Retransmission Approaches (RAs), see also Figure 2. The following example sequences assume a fair RR multiplexer that needs to schedule 8 packets, 2 for each stream A, B, C, and D, where A and B's packets contain retransmissions, versus C and D's new data:

RA1: retransmissions are seen as "normal" data and sent when the scheduler next selects the stream: ABCDABCD.

RA2: retransmissions are given highest precedence, and use the default RR scheduling approach: ABABCDCD.

RA3: retransmissions are given highest precedence, and use a non-default sequential scheduling approach: AABBCDCD.

RA4: retransmissions explicitly take into account application-layer prioritization (e.g., new data for a high priority H3 stream (C) could get precedence over retransmissions of lower priority H3 streams (A and B, with lowest priority D)): CCABABDD.

Here we find that most implementers do give retransmissions a higher priority: 9/13 do RA2 and 1/13 (mvfst ①) does RA3, while just 2/13 employ RA1. Only quicly currently supports RA4 for H3 transfers, falling back to RA2 if QUIC is used directly without H3. The low occurence of RA4 is likely because only 3/15 stacks integrate QUIC stream scheduling with H3 semantics at this point.

It is unclear which RA performs best in practice, though it seems especially a variant of RA4 could give H3 an edge over H2 in some

situations. Most implementers indicate their current RA is not an intentional, considered choice but rather emergent behaviour of how they globally handle stream data scheduling.

3.3 Packetization

While the binary QUIC and H3 frame and packet structures are well-defined in the specifications, there are many variations in how they can be utilized, sized and combined. For example, H3 defines the HEADERS and DATA frames [8]. These are in turn passed to the QUIC layer, whereby typically QUIC has no knowledge of the H3 semantics: it treats H3-level data on each QUIC stream as an ordered but opaque byte sequence. These bytes are then put inside QUIC-level STREAM frames for transport. Practically, this means that multiple H3-level frames can be aggregated together inside a single QUIC STREAM frame, which is good for efficiency. We find that 9/13 servers consistently do this, but that 4/13 tend to instead pack HEADERS and DATA frames into separate STREAM frames. In stress tests that request hundreds of very small files (<1kB), 6/13 servers started showing even more inefficient behaviour, packing all H3 frames in separate STREAM frames, and even in tiny QUIC packets. If we define goodput efficiency as the amount of useful transported H3-level data (e.g., image file bytes) divided by the total amount of bytes on the wire (including QUIC and H3 framing overhead), we find that most stacks achieve 95-97% efficiency when downloading larger files, which plummets to about 90% for most when downloading many smaller files, with the worst case only achieving 83%. Few implementers indicated that they had actively considered or tuned their packetization overhead at this time.

A part of the goodput efficiency is the sizing of H3 DATA frames. While QUIC STREAM frames cannot span multiple QUIC packets, H3 DATA frames can theoretically be up to about 4600 Petabytes (as their length is 62-bit encoded), and thus span many STREAM frames. For goodput efficiency, fewer and thus larger DATA frames are best. Here, we find a large heterogeneity. When downloading files larger than 1MB, 6/13 have DATA frames larger than 1MB (large), 2/13 between 1MB and 100kB (medium), and 5/13 lower than 100kB (small, of which one has the worst case of generating a new DATA frame for each QUIC packet). Interestingly, we observed 3 stacks that dynamically sized their DATA frames, growing or shrinking over time. While this first seemed like intentional behaviour, it again turned out to be due to (unexpected) cross-layer code interactions. For example, one implementation simply writes as much H3 data (wrapped in a single DATA frame) as allowed by the QUIC send buffer, which is in turn dynamically sized based on the current Congestion Control Window (see §3.4). Another acts as a TCP-QUIC proxy, transforming QUIC streams into individual TCP connections and vice-versa. Here, we observed DATA frame sizes to be highly irregular, as they are dependent on how much data is available for each individual stream, which is in turn internally driven by the separate TCP Congestion and Flow Control dynamics.

Finally, QUIC mandates a minimum UDP payload size of 1200 bytes [23], but it is generally understood that larger sizes significantly improve efficiency [19, 25]. It is best practice to start with a small packet size and perform Path MTU Discovery (PMTUD) [23]. Still, we find that at this time, just 3/14 stacks implement PMTUD, all of them using the naive method of sending a single 1400-1500 byte QUIC packet containing mainly PADDING instead of the more advanced DPLPMTUD approach [17]. The need for PMTUD was emphasized to us by Facebook, who find many networks exhibit higher loss rates if QUIC packets are even a few dozen bytes larger. Other implementers recognize the usefulness of PMTUD, but did not consider it a priority.

3.4 Congestion Control

In terms of recovery (loss detection and congestion control (CC)), QUIC inherits most of TCP's concepts and decades of best practices (e.g., selective acknowledgements, pacing, tail loss probes). The QUIC recovery text [21] aggregates a discussion of all these concepts with how they can be adapted to QUIC peculiarities such as its integrated TLS handshake and unique packet numbers. For a practical example with pseudo-code, the somewhat outdated, yet well-understood New Reno CC [21] is used, even though it is not expected many deployments would want to use this in production. As such, the text provides a good starting point for adapting other CCs to QUIC. We find that while most of the stacks have implemented QUIC's New Reno variant (9/15), especially many of the larger companies indeed also support more modern CCs: 6/15 implement Cubic (4 with hystart [13], 1 with tweaks for satellite networks [20]), 4/15 implement BBRv1 and 3/15 go further with approaches like COPA [7] or BBRv2 [11]. Facebook deploys BBRv1, Cloudflare Cubic [13]. Most indicate that their CC implementation is ongoing work and has not been (extensively) validated for performance or fairness. Some developers not tied to larger companies mention not having enough CC expertise to evaluate their implementation or to move to a more advanced CC algorithm.

An important CC variable is the initial Congestion Window (cwnd), which controls how many bytes an endpoint can send back in the first flight (before growing the cwnd in "slow start"). The QUIC text's advice of an initial cwnd of 12kB-15kB (which is thought to balance the risk of packet loss with a fast enough start [21]) is followed by 11/14 stacks. In contrast, 3/14 (mostly those with roots in the older gQUIC) choose a much larger window of 40kB+, though these values are more heterogeneous in actual deployments [38]. For example, we learned that Facebook uses machine learning to tune their init cwnd, while f5 includes a cwnd estimate in their address validation token for resumed connections (see §3.5).

Additionally, the QUIC text strongly encourages the use of pacing (i.e., spreading out packets over an entire RTT instead of sending them in a single burst with each cwnd increase, which is thought to lower packet loss [6]). Interestingly, only 8/15 currently support pacing. This is mainly due to the complexity of the technique (as many QUIC implementers are not CC experts) and lacking support in the Linux kernel in combination with other optimization techniques (e.g., GSO combined with SO_TXTIME [16, 19, 25]).

Finally, the performance of a CC can be influenced heavily by the frequency with which the receiver acknowledges (ACKs) data. QUIC recommends sending an ACK for every 2 received packets [21]. Just 2/12 do so consistently, with 10/12 ACKing every 1-10+ packets. This latter behaviour is mostly due to implementations reading up to 10 or more packets at a time from the socket (and e.g., pacing can thus influence variability). It is however also understood that ACK processing is expensive in QUIC [18, 25] and 4/15 are experimenting with the ACK Frequency extension to reduce this overhead [22]

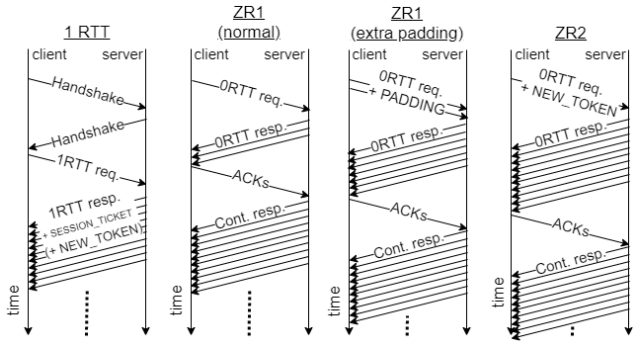

Figure 3: 0RTT request and response size variations. We assume an initial cwnd of 10 packets at the server.

3.5 0RTT

One of the key new features in QUIC is the zero RTT (0RTT) connection setup [26], which allows the exchange of application data (e.g., an H3 GET and its (partial) response) in the first flight (compared to third or fourth in TCP+TLS). This derives from TLS 1.3, which allows exchanging Pre-Shared encryption Keys in Session Tickets during a first "1RTT" connection (where data can only be exchanged from the second flight onwards). These keys are then used to enable 0RTT on a subsequent connection [43], see Figure 3. Despite being a high-profile feature, just 13/18 implement it, of which we tested 9.

One of the reasons for this lower uptake is that 0RTT is complex to implement securely, as it is vulnerable to (HTTP) replay and UDP amplification/reflection attacks [23, 43]. This latter category is possible when the attacker spoofs their IP address and sends a (small) 0RTT request for a (large) resource to the server. If the server simply starts sending the (entire) resource to the spoofed victim IP, it could be used in a (D)DoS attack. To prevent this, a QUIC server MUST NOT [23] send more than 3 times as much data as it has received from the client until the path is validated (confirming the IP was not spoofed). This validation happens in 3 main ways (ZRs) (Figure 3):

ZR1 waits for a reply from the client to the early 0RTT server packets. This has the large downside that the 0RTT response will be rather small (just 5kB-7kB if the client sends its initial request in 1-2 packets). We feel this significantly reduces 0RTT's usefulness for typical Web browsing use cases.

ZR2 alleviates this by sending an Address Validation token in QUIC's NEW_TOKEN frame [23]. This is sent encrypted by the server in the first connection and used by the client for the second, so the server can consider the path validated immediately. This allows it to ignore the 3X limit and send more data, typically up to its initial cwnd (10-40kB, see §3.4), which is superior to ZR1 in most cases.

ZR3 is mainly a legacy equivalent of ZR2 which securely encodes the client's IP address inside the TLS Session Ticket.

ZR1 and ZR2 are both used in 6/13 stacks, but ZR3 only by Facebook, who have plans to migrate to the superior ZR2. Several ZR1 implementers also intend to switch to ZR2 in the future, initially opting for the suboptimal ZR1 mainly for its easier implementation.

One way to improve upon ZR1 would be for the client to send additional data along with the 0RTT request (e.g., in the form of padding), see Figure 3. This would in turn allow the server to reply with more data while still adhering to the 3X limit. While testing whether the stacks would respond well to this, we found several high impact bugs. One stack simply did not adhere to the 3X limit, replying up to their 46kB initial cwnd to a 1.2kB request (a 36X amplification). Another did apply the 3X limit, but forgot to check its initial cwnd for 0RTT responses (e.g., replying with 30kB 0RTT data to a 10kB request even though their cwnd was only 15kB). Finally, one stack forgot to account retransmissions of lost packets in its 3X limit. If an attacking client never replied to anything after its first 1.2kB, this stack sent up to 17kB of (retransmitted) data (14X). Most other servers did adhere to the 3X limit and also sent more data in response to a client sending additional padding. As such, we recommend clients pad their 0RTT requests to about 4kB-5kB (higher values will give diminishing returns as most servers utilize an initial cwnd of about 13kB (§3.4)). This should not be needed for servers employing ZR2 (which should be preferred over ZR3).

4 DISCUSSION & CONCLUSION

In this work, we have discussed 15 different QUIC implementations across a multitude of behaviours (see Table 1). Even though these stacks all implement the exact same QUIC/H3 protocols, we have shown that their low-level implementation choices lead to a large behavioural heterogeneity between them. We believe this has important consequences for future QUIC/H3 research and evaluation.

While not all considered aspects might have a large impact on most types of protocol evaluation results (e.g., H3 DATA frame sizing or PMTUD support will typically matter less for the short lived flows observed in Web page loading performance research [10]), other discussed features, such as Flow Control, Congestion Control, Prioritization and 0 RTT can all lead to significant differences in results. Yet, these are aspects that historically we and others [24, 45, 46] rarely see rigorously evaluated or discussed in related work focusing on H2 and also gQUIC. In order to be able to draw solid conclusions about QUIC/H3 as protocols, we feel that future work should strive to show scientific rigor in two main ways. Firstly, by performing deep root-cause analysis of all observed high-level behaviours. Secondly, by comparing multiple QUIC implementations. This especially holds true in the next few years (2020-2023), as not all implementations will be fully optimized or complete by the time QUIC/H3 are finalized. Even though we expect many implementations to eventually gravitate towards a smaller set of best practices than the approaches we have encountered in this study, we still suspect stacks and especially individual deployments will remain heterogeneous and deep insight will remain key in researching, optimizing and extending QUIC+H3.

We believe that our methodology of using the qlog and qvis tools [27, 31] has proven its potential to form the basis of a framework to both analyze and extend or improve QUIC/H3 stacks. This is also evidenced by the fact that several QUIC implementers have lately started using these tools to validate their approaches [13, 41]. As currently 12/18 QUIC stacks support qlog, these tools are broadly available and ready to use.

Overall, we posit that QUIC stacks are becoming mature enough to be deployed and researched, but results from high-level metrics should be thoroughly root-cause analyzed if researchers want to draw broad conclusions on QUIC/H3 as protocols. There are many opportunities for future research on QUIC behaviour tuning, especially around Flow Control, Multiplexing/H3 Prioritization, and Retransmission approaches.

ACKNOWLEDGEMENTS

Robin Marx is a SB PhD fellow at FWO, Research Foundation Flanders, #1S02717N. The authors would like to thank our shepherd Vaibhav Bajpai for his guidance during the review process. We also appreciate the help of Maarten Wijnants, Jens Bruggemans, Dmitri Tikhonov, Lucas Pardue, Maxime Piraux and Song Zhu in reviewing earlier versions of this work.

REFERENCES

[1] 2020. (2020). https://github.com/rmarx/aioquic.
[2] 2020. Active QUIC implementations. (2020). https://github.com/quicwg/base-drafts/wiki/Implementations.
[3] 2020. ns-3: a discrete-event network simulator for Internet systems. (May 2020). https://www.nsnam.org/.
[4] 2020. QUIC Interop Runner. (2020). https://interop.seemann.io/.
[5] 2020. QUICdev Slack Group. (2020). https://quicdev.slack.com/.
[6] Amit Aggarwal, Stefan Savage, and Thomas Anderson. 2000. Understanding the performance of TCP pacing. In *Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE INFOCOM 2000).*, Vol. 3. IEEE, 1157–1165. https://doi.org/10.1109/INFCOM.2000.832483
[7] Venkat Arun and Hari Balakrishnan. 2018. Copa: Practical Delay-Based Congestion Control for the Internet. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*. USENIX Association, Renton, WA, 329–342. https://www.usenix.org/conference/nsdi18/presentation/arun
[8] Mike Bishop. 2020. *Hypertext Transfer Protocol Version 3 (HTTP/3).* Internet-Draft. https://tools.ietf.org/html/draft-ietf-quic-http-27
[9] Prasenjeet Biswal and Omprakash Gnawali. 2016. Does quic make the web faster?. In *IEEE Global Communications Conference (GLOBECOM 2016)*. IEEE, 1–6. https://doi.org/10.1109/GLOCOM.2016.7841749
[10] Enrico Bocchi, Luca De Cicco, and Dario Rossi. 2016. Measuring the Quality of Experience of Web Users. *SIGCOMM Comput. Commun. Rev.* 46, 4 (Dec. 2016), 8–13. https://doi.org/10.1145/3027947.3027949
[11] Neal Cardwell, Yuchung Cheng, C. Stephen Gunn, Soheil Hassas Yeganeh, and Van Jacobson. 2017. BBR: Congestion-Based Congestion Control. *Commun. ACM* 60, 2 (Jan. 2017), 58–66. https://doi.org/10.1145/3009824
[12] Gaetano Carlucci, Luca De Cicco, and Saverio Mascolo. 2015. HTTP over UDP: An Experimental Investigation of QUIC. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing (SAC '15)*. Association for Computing Machinery, New York, NY, USA, 609–614. https://doi.org/10.1145/2695664.2695706
[13] Junho Choi. 2020. CUBIC and HyStart++ Support in quiche. (2020). https://blog.cloudflare.com/cubic-and-hystart-support-in-quiche/.
[14] Junho Choi. 2020. quiche: update receiver flow control. (May 2020). https://github.com/cloudflare/quiche/pull/529.
[15] Andy Davies and Patrick Meenan. 2018. Tracking HTTP/2 Prioritization Issues. (December 2018). https://github.com/andydavies/http2-prioritization-issues.
[16] Willem de Bruijn and Eric Dumazet. 2018. Optimizing UDP for content delivery: GSO, pacing and zerocopy. In *Linux Plumbers Conference*. http://vger.kernel.org/lpc_net2018_talks/willemdebruijn-lpc2018-udpgso-paper-DRAFT-1.pdf.
[17] Fairhurst et al. 2020. *Packetization Layer Path MTU Discovery for Datagram Transports.* Internet-Draft. https://tools.ietf.org/html/draft-ietf-tsvwg-datagram-plpmtud-20
[18] Gorry Fairhurst and Ana Custura. 2020. Changing QUIC default to ACK 1:10. (2020). https://erg.abdn.ac.uk/users/gorry/ietf/QUIC/QUIC-ack10-24-april-00.pdf.
[19] Alessandro Ghedini. 2020. Accelerating UDP packet transmission for QUIC. (2020). https://blog.cloudflare.com/accelerating-udp-packet-transmission-for-quic/.
[20] Christian Huitema. 2020. Faster slow start for satellite links? (2020). https://huitema.wordpress.com/2020/04/21/faster-slow-start-for-satellite-links/.
[21] Jana Iyengar and Ian Swett. 2020. *QUIC Loss Detection and Congestion Control.* Internet-Draft. https://tools.ietf.org/html/draft-ietf-quic-recovery-27
[22] Jana Iyengar and Ian Swett. 2020. *Sender Control of Acknowledgement Delays in QUIC.* Internet-Draft. https://tools.ietf.org/html/draft-iyengar-quic-delayed-ack-00
[23] Jana Iyengar and Martin Thomson. 2020. *QUIC: A UDP-Based Multiplexed and Secure Transport.* Internet-Draft. https://tools.ietf.org/html/draft-ietf-quic-transport-27
[24] Arash Molavi Kakhki, Samuel Jero, David Choffnes, Cristina Nita-Rotaru, and Alan Mislove. 2017. Taking a Long Look at QUIC: An Approach for Rigorous Evaluation of Rapidly Evolving Transport Protocols. In *Proceedings of the 2017 Internet Measurement Conference (IMC '17)*. Association for Computing Machinery, New York, NY, USA, 290–303. https://doi.org/10.1145/3131365.3131368
[25] Jana Iyengar Kazuho Oku. 2020. Can QUIC match TCP's computational efficiency? (2020). https://www.fastly.com/blog/measuring-quic-vs-tcp-computational-efficiency.
[26] Adam Langley, Alistair Riddoch, Alyssa Wilk, Antonio Vicente, Charles Krasic, Dan Zhang, Fan Yang, Fedor Kouranov, Ian Swett, Janardhan Iyengar, Jeff Bailey, Jeremy Dorfman, Jim Roskind, Joanna Kulik, Patrik Westin, Raman Tenneti, Robbie Shade, Ryan Hamilton, Victor Vasiliev, Wan-Teh Chang, and Zhongyi Shi. 2017. The QUIC Transport Protocol: Design and Internet-Scale Deployment. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17)*. Association for Computing Machinery, New York, NY, USA, 183–196. https://doi.org/10.1145/3098822.3098842
[27] Robin Marx. 2020. qvis toolsuite live. (2020). https://qvis.edm.uhasselt.be.
[28] Robin Marx., Tom De Decker., Peter Quax., and Wim Lamotte. 2019. Of the Utmost Importance: Resource Prioritization in HTTP/3 over QUIC. In *Proceedings of the 15th International Conference on Web Information Systems and Technologies (WEBIST 2019)*. INSTICC, SciTePress, 130–143. https://doi.org/10.5220/0008191701300143
[29] Robin Marx, Wim Lamotte, Jonas Reynders, Kevin Pittevils, and Peter Quax. 2018. Towards QUIC Debuggability. In *Proceedings of the Workshop on the Evolution, Performance, and Interoperability of QUIC (EPIQ'18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3284850.3284851
[30] Robin Marx, Maxime Piraux, Peter Quax, and Wim Lamotte. 2020. Debugging Modern Web Protocols with qlog. In *Proceedings of the Applied Networking Research Workshop (ANRW 2020)*. https://qlog.edm.uhasselt.be/anrw/
[31] Robin Marx, Marten Seemann, and Jeremy Lainé. 2019. The IETF I-D documents for the qlog format. (2019). https://github.com/quiclog/internet-drafts.
[32] Patrick Meenan. 2019. Better HTTP/2 Prioritization for a Faster Web. (2019). https://blog.cloudflare.com/better-http-2-prioritization-for-a-faster-web/.
[33] Péter Megyesi, Zsolt Krämer, and Sándor Molnár. 2016. How quick is QUIC?. In *IEEE International Conference on Communications (ICC 2016)*. IEEE, 1–6. https://doi.org/10.1109/ICC.2016.7510788
[34] Kazuho Oku and Lucas Pardue. 2020. *Extensible Prioritization Scheme for HTTP.* Internet-Draft. https://tools.ietf.org/html/draft-ietf-httpbis-priority-00
[35] Mirko Palmer, Thorben Krüger, Balakrishnan Chandrasekaran, and Anja Feldmann. 2018. The QUIC Fix for Optimal Video Streaming. In *Proceedings of the Workshop on the Evolution, Performance, and Interoperability of QUIC (EPIQ'18)*. Association for Computing Machinery, New York, NY, USA, 43–49. https://doi.org/10.1145/3284850.3284857
[36] James Pavur, Martin Strohmeier, Vincent Lenders, and Ivan Martinovic. 2020. QPEP: A QUIC-Based Approach to Encrypted Performance Enhancing Proxies for High-Latency Satellite Broadband. *arXiv preprint* (2020). https://arxiv.org/pdf/2002.05091.pdf.
[37] Maxime Piraux, Quentin De Coninck, and Olivier Bonaventure. 2018. Observing the Evolution of QUIC Implementations. In *Proceedings of the Workshop on the Evolution, Performance, and Interoperability of QUIC (EPIQ'18)*. Association for Computing Machinery, New York, NY, USA, 8–14. https://doi.org/10.1145/3284850.3284852
[38] Jan Rüth, Ike Kunze, and Oliver Hohlfeld. 2019. TCP's Initial Window—Deployment in the Wild and Its Impact on Performance. *IEEE Transactions on Network and Service Management (TNSM 2019)* 16, 2 (2019), 389–402. https://doi.org/10.1109/TNSM.2019.2896335
[39] Darius Saif, Chung-Horng Lung, and Ashraf Matrawy. 2020. An Early Benchmark of Quality of Experience Between HTTP/2 and HTTP/3 using Lighthouse. *arXiv preprint* (2020). https://arxiv.org/pdf/2004.01978.pdf.
[40] Robbie Shade. 2016. Flow Control in Google QUIC. (2016). https://docs.google.com/document/d/1F2YfdDXKpy20WVKJueEf4abn_LVZHhMUMS5gX6Pgjl4.
[41] Daniel Stenberg. 2020. qlog with curl. (2020). https://daniel.haxx.se/blog/2020/05/07/qlog-with-curl/.
[42] Ian Swett and Robin Marx. 2019. HTTP Priority design team update - IETF 107. (18 November 2019). https://github.com/httpwg/wg-materials/blob/gh-pages/ietf106/priorities.pdf.
[43] Martin Thomson and Sean Turner. 2020. *Using TLS to Secure QUIC.* Internet-Draft. https://tools.ietf.org/html/draft-ietf-quic-tls-27
[44] Eric Weigle and Wu-chun Feng. 2002. A comparison of TCP automatic tuning techniques for distributed computing. In *Proceedings 11th IEEE International Symposium on High Performance Distributed Computing (HPDC 2002)*. IEEE, 265–272. https://doi.org/10.1109/HPDC.2002.1029926
[45] Maarten Wijnants, Robin Marx, Peter Quax, and Wim Lamotte. 2018. HTTP/2 Prioritization and Its Impact on Web Performance. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. ACM, 1755–1764. https://doi.org/10.1145/3178876.3186181
[46] Konrad Wolsing, Jan Rüth, Klaus Wehrle, and Oliver Hohlfeld. 2019. A Performance Perspective on Web Optimized Protocol Stacks: TCP+TLS+HTTP/2 vs. QUIC. In *Proceedings of the Applied Networking Research Workshop (ANRW '19)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3340301.3341123