Measuring Latency Variation in the Internet

Toke Høiland-Jørgensen
Dept of Computer Science
Karlstad University, Sweden
toke.hoiland-jorgensen@kau.se

Bengt Ahlgren
SICS
Box 1263, 164 29 Kista
Sweden
bengta@sics.se

Per Hurtig
Dept of Computer Science
Karlstad University, Sweden
per.hurtig@kau.se

Anna Brunstrom
Dept of Computer Science
Karlstad University, Sweden
anna.brunstrom@kau.se

ABSTRACT

We analyse two complementary datasets to quantify the latency variation experienced by internet end-users: (i) a large-scale active measurement dataset (from the Measurement Lab Network Diagnostic Tool) which shed light on long-term trends and regional differences; and (ii) passive measurement data from an access aggregation link which is used to analyse the edge links closest to the user.

The analysis shows that variation in latency is both common and of significant magnitude, with two thirds of samples exceeding 100 ms of variation. The variation is seen within single connections as well as between connections to the same client. The distribution of experienced latency variation is heavy-tailed, with the most affected clients seeing an order of magnitude larger variation than the least affected. In addition, there are large differences between regions, both within and between continents. Despite consistent improvements in throughput, most regions show no reduction in latency variation over time, and in one region it even increases.

We examine load-induced queueing latency as a possible cause for the variation in latency and find that both datasets readily exhibit symptoms of queueing latency correlated with network load. Additionally, when this queueing latency does occur, it is of significant magnitude, more than 200 ms in the median. This indicates that load-induced queueing contributes significantly to the overall latency variation.

Keywords

Latency; Bufferbloat; Access Network Performance

CoNEXT '16 December 12-15, 2016, Irvine, CA, USA

ACM ISBN 978-1-4503-4292-6/16/12.
DOI: http://dx.doi.org/10.1145/2999572.2999603

1. INTRODUCTION

As applications turn ever more interactive, network latency plays an increasingly important role for their performance. The end-goal is to get as close as possible to the physical limitations of the speed of light [25]. However, today the latency of internet connections is often larger than it needs to be. In this work we set out to quantify how much. Having this information available is important to guide work that sets out to improve the latency behaviour of the internet; and for authors of latency-sensitive applications (such as Voice over IP, or even many web applications) that seek to predict the performance they can expect from the network.

Many sources of added latency can be highly variable in nature. This means that we can quantify undesired latency by looking specifically at the *latency variation* experienced by a client. We do this by measuring how much client latency varies above the minimum seen for that client. Our analysis is based on two complementary sources of data: we combine the extensive publicly available dataset from the Measurement Lab Network Diagnostic Tool (NDT) with a packet capture from within a service provider access network. The NDT data, gathered from 2010 to 2015, comprises a total of 265.8 million active test measurements from all over the world. This allows us to examine the development in latency variation over time and to look at regional differences. The access network dataset is significantly smaller, but the network characteristics are known with greater certainty. Thus, we can be more confident when interpreting the results from the latter dataset. These differences between the datasets make them complement each other nicely.

We find that significant latency variation is common in both datasets. This is the case both within single connections and between different connections from the same client. In the NDT dataset, we also observe that the magnitude of latency variation differs between geographic regions, both between and within continents. Looking at the development over time (also in the NDT dataset), we see very little change in the numbers. This is in contrast to the overall throughput that has improved significantly.

Table 1: Total tests per region and year (millions).

Reg.	2010	2011	2012	2013	2014	2015
AF	0.65	0.63	0.79	0.72	0.76	0.57
AS	7.79	7.45	6.55	5.75	5.83	4.57
EU	35.00	31.12	27.96	22.40	21.23	16.82
NA	11.70	8.51	7.90	7.01	7.06	8.00
SA	2.68	1.73	2.83	2.94	2.05	1.26
OC	1.33	1.33	0.79	0.55	0.65	0.57
Total	59.22	50.81	46.87	39.43	37.60	31.79

One important aspect of latency variation is the correlation between increased latency and high link utilisation. Queueing delay, in particular, can accumulate quickly when the link capacity is exhausted, and paying attention to such scenarios can give insight into issues that can cause real, if intermittent, performance problems for users. We examine queueing delay as a possible source of the observed latency variation for both datasets, and find strong indications that it is present in a number of instances. Furthermore, when queueing latency does occur it is of significant magnitude.

The rest of the paper is structured as follows: Section 2 introduces the datasets and the methodology we have used to analyse them. Section 3 discusses the large-scale variations in latency over time and geography, and Section 4 examines delay variation on access links. Section 5 presents our examination of load-induced queueing delay. Finally, Section 6 discusses related work and Section 7 concludes the paper.

2. DATASETS AND METHODOLOGY

The datasets underlying our analysis are the publicly available dataset from Measurement Lab (M-Lab), specifically the Network Diagnostic Tool (NDT) data [1], combined with packet header traces from access aggregation links of an internet service provider. This section presents each of the datasets, and the methodology we used for analysis.

2.1 The M-Lab NDT data

The M-Lab NDT is run by users to test their internet connections. We use the 10-second bulk transfer from the server to the client, which is part of the test suite. When the test is run, the client attempts to pick the nearest server from the geographically distributed network of servers provided by the M-Lab platform. The M-Lab servers are globally distributed[1], although with varying density in different regions.

The server is instrumented with the Web100 TCP kernel instrumentation [21], and captures several variables of the TCP state machine every 5 ms of the test. Data is available from early 2009, and we focus on the six year period 2010–2015, comprising a total of 265.8 million test runs. Table 1 shows the distribution of test runs for the years and regions we have included in our analysis.

Since the NDT is an active test, the gathered data is not a representative sample of the traffic mix flowing through the internet. Instead, it may tell us something about the *links* being traversed by the measurement flows. Looking at links under load is interesting, because some important effects can be exposed in this way, most notably bufferbloat: Loading up the link causes any latent buffers to fill, adding latency that might not be visible if the link is lightly loaded. This also means that the baseline link utilisation (e.g., caused by diurnal usage patterns) become less important: an already loaded link can, at worst, result in a potentially higher baseline latency. This means that the results may be biased towards showing lower latency variation than is actually seen on the link over time. But since we are interested in establishing a lower bound on the variation, this is acceptable.

[1]See http://www.measurementlab.net/infrastructure for more information on the server placements.

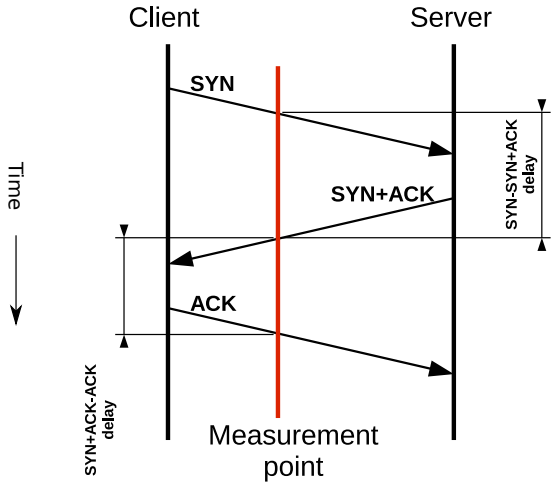


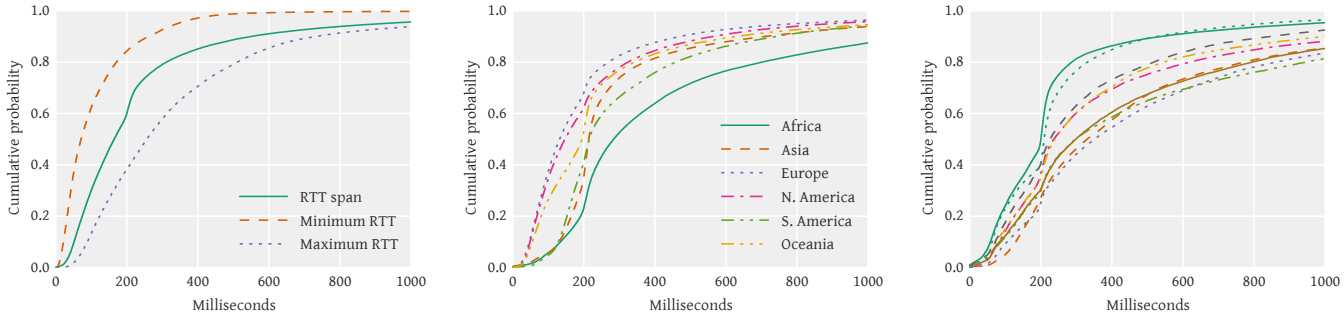
Figure 1: Delays computed from the TCP connection setup.

For the base analysis, we only exclude tests that were truncated (had a total run time less than 9 seconds). We use the TCP RTT samples as our data points, i.e., the samples computed by the server TCP stack according to Karn's algorithm [23], and focus on the *RTT span*, defined as the difference between the minimum and maximum RTT observed during a test run. However, we also examine a subset of the data to assess what impact the choice of min and max observed RTT has on the data when compared to using other percentiles for each flow (see Section 3.3).

2.2 The access aggregation link data

The second dataset comes from two access aggregation links of an internet service provider. The links aggregate the traffic for about 50 and 400 clients, respectively, and connect them to the core network of the service provider. The traffic was captured passively at different times distributed over an eight-month period starting at the end of 2014. The average loads on the 1 Gbps links were, respectively, about 200 and 400 Mbit/s during peak hours. This dataset is an example of real internet traffic, since we are not generating any traffic.

We analyse the delay experienced by the TCP connection setup packets in this dataset. The TCP connection setup consists of a three-way handshake with SYN, SYN+ACK, and ACK packets, as illustrated in Figure 1.



(a) (b) (c)

Figure 2: RTT values computed over individual flows. (a) Min and max RTT and the span between them, for all flows. (b) Distribution of per-flow RTT span per continent (2015 data). (c) Distribution of per-flow RTT span per country in Africa (2015 data for countries with $n > 10,000$).

For the purpose of this paper, we study the client side of connections made to the public internet. That is, we study the path from the client, over the access link, to the measured aggregation link. This allows us to examine increased delays due to excess queuing in consumer equipment, and ensures that the path we measure is of known length.

We examine the data for outgoing connections, i.e., connections that are initiated from the access side and connect to the public internet, which means we compute the round-trip delay between the SYN+ACK packet and the first ACK packet in the three-way handshake. The variation in these delay values is likely to be caused by queueing, since connection endpoints normally respond immediately. We also compute the instantaneous load at each sample and examine the correlation between delay and load for a few clients.

2.3 Sources of latency variation

Naturally, the observed latency variation can have several causes [6]. These include queueing delay along the path, delayed acknowledgements, transmission delay, media access delays, error recovery, paths changing during the test and processing delays at end-hosts and intermediate nodes. For the main part of our analysis, we make no attempt to distinguish between different causes of latency variation. However, we note that latency variation represents latency that is superfluous in the sense that it is higher than the known attainable minimum for the path. In addition, we analyse a subset of each dataset to examine to what extent queueing latency is a factor in the observed latency variation.

We believe that the chosen datasets complement each other nicely and allow us to illuminate the subject from different angles, drawing on the strengths of them both. The NDT dataset, being based on active measurements, allows us to examine connections that are being deliberately loaded, and the size of the dataset allows us to examine temporal and geographic trends. The access network dataset, on the other hand, has smaller scope but the examined path is known; and so we can rule out several sources of delay and be more confident when interpreting the results.

3. LATENCY VARIATION OVER TIME AND GEOGRAPHY

In this section, we analyse the M-Lab NDT dataset to explore geographic differences, and also look at the development of the RTT span over time. For an initial overview, Figure 2a shows the distribution of the RTT span in the whole M-Lab NDT dataset, along with the minimum and maximum RTTs it is derived from. This shows a significant amount of extra latency: two thirds of samples exceed 100 ms of RTT span, with the 95th percentile exceeding 900 ms.

3.1 Geographic differences

Figure 2b shows the RTT span distributed geographically per continent for the 2015 data. This shows a significant difference between regions, with the median differing by more than a factor of two between the best and the worst region. Looking within these regions, Figure 2c shows the per-country distributions within Africa. Here, the heavy tail of latencies above one second affects as much as 20% of the samples from the country with the highest latency. These high latencies are consistent with previous studies of African internet connections [8]. The data for Europe (omitted due to space constraints) shows that the difference among European countries is of the same magnitude as the difference among continents.

3.2 Development over time

Figure 3a shows the minimum RTT and the RTT span for each of the years in the dataset. While the minimum RTT has decreased slightly over the years, no such development is visible for the span. This is striking when compared to the development in throughput, which has increased consistently, as shown in Figure 3b. Some of this increase in average throughput may be due to other factors than increase in link capacity (e.g. protocol efficiency improvements). Even so, the disparity is clear: while throughput has increased, RTT span has not decreased, and the decrease in minimum RTT is slight.

Looking at this development for different continents, as shown in Figure 3c, an increase in latency span over the

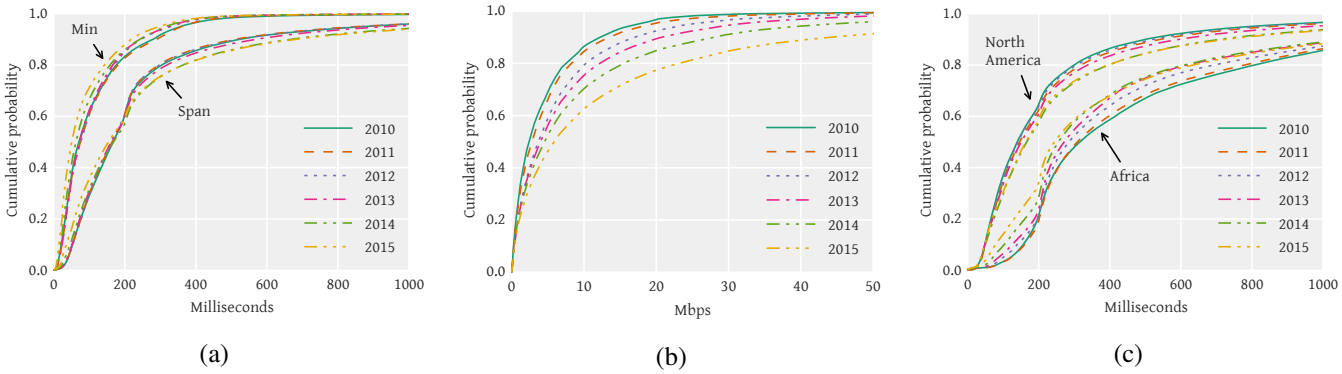


Figure 3: Development over time of per-flow latency and throughput in the NDT dataset. (a) Min RTT and RTT span, per year. (b) Per-flow average throughput, per year. (c) The development in latency span over time for North America and Africa.

years is seen in North America while Africa has seen a consistent, but small, reduction in latency span over time. This latter development is most likely due to developments in infrastructure causing traffic to travel fewer hops, thus decreasing the potential sources of extra latency.

3.3 Different measures of latency span

The way the NDT dataset is structured makes the per-flow min and max RTT values the only ones that are practical to analyse for the whole dataset. To assess what effect this choice of metric has on the results, we performed a more detailed analysis for a subset of the data. Figure 4a shows the latency span distribution for the data from August 2013 when using percentiles of the per-flow RTT measurements ranging from 90 to 99 in place of the max. We see that in this case the median measured RTT span drops to between 151 ms and 204 ms, from 250 ms when using the max — a drop of between 17% and 40%. It is not clear that the max is simply an outlier for all flows; but for those where it is, our results will overestimate the absolute magnitude of the RTT span. However, the shape of the distribution stays fairly constant, and using the max simply leads to higher absolute numbers. This means that we can still say something about trends, even for those flows where the max RTT should be considered an outlier.

4. LATENCY VARIATION IN THE ACCESS NETWORK

In this section we analyse the latency variation of TCP 3-way handshakes in the access network dataset. Figure 4b shows the distribution of the per-client RTT variation, computed as the span between the per-client minimum delay and the respective percentiles of samples to that client. To ensure that we do not mistakenly use a too low minimum delay value, only handshakes which did not have any SYN+ACK retransmissions are considered when computing the minimum.

For about half of the client population covered by the link shown in Figure 4b, delay increases substantially over the minimum at times. For example, 20% of the clients experience increased delays of more than about 80 ms, at least 5% of the time. The second link shows similar behaviour, but has *more* clients that are affected by increased delay.

Another interesting feature of the data is that there is a significant difference in the magnitude of the latency span depending on which percentile of latency measurements one looks at. That is, if we consider the per-user 99th percentile of latency rather than the 95th, suddenly more than half the users experience latency variations in excess of 100 ms for the first link, and more than 80% for the second link. This underscores the fact that delay spikes can be a very transient problem, but one that is of significant magnitude when it does occur.

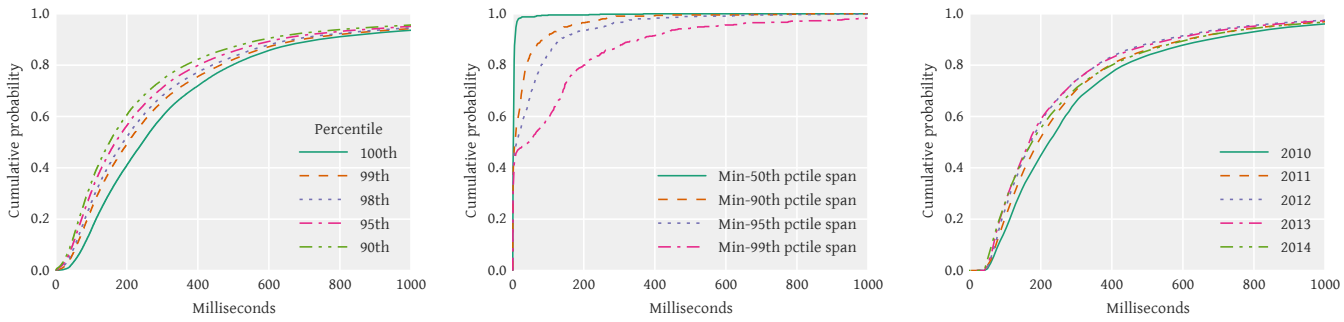
Comparing with the NDT dataset, the analysis of the access link data shows a lower frequency of latency variation, as well as a lower magnitude of the variation when it does occur. However, both datasets show that significant latency variation occurs for a considerable fraction of users. We attribute the difference in magnitude to the difference in measurement methods: the NDT measurements are taken while the link is deliberately loaded, while not all measurements from the access network are taken from saturated links.

5. EXAMINING QUEUEING LATENCY

As mentioned in Section 2.3, latency variations can have many causes, and without having insight into the network path itself it can be difficult to identify which are the most prevalent. However, experience from more controlled environments (such as experiments performed to evaluate AQM algorithms [14]) suggests that queueing delay can be a significant source. Due to the magnitude of the variation we see here, we conjecture that this is also the case in this dataset. To examine this further, in this section we present an analysis of the queueing delay of a subset of the traffic in both datasets. We aim to perform a conservative analysis, and so limit ourselves to tests for which it is possible to identify queueing latency with high certainty.

5.1 Latency reductions after a drop

Our analysis is based upon a distinct pattern, where the sample RTT increases from the start of a flow until a congestion event, then sharply decreases afterwards. An exam-



(a) (b) (c)

Figure 4: (a) Latency span for all flows in August 2013 when using different percentiles to determine the max RTT. NDT dataset. (b) Client side round trip delay percentiles over all clients relative to the minimum delay. A full day of the first aggregation link in the access network dataset. (c) Distribution of the magnitude of detected queueing delay, per year. Flows with detected queueing latency, NDT dataset.

ple of this pattern is seen in Figure 5. This pattern is due to the behaviour of TCP: The congestion control algorithm will increase its sending rate until a congestion event occurs, then halve it. If a lot of packets are queued when this happens, the queue has a chance to drain, and so subsequent RTT samples will show a lower queueing delay. Thus, it is reasonable to assume that when this pattern occurs, the drop in RTT is because the queue induced by the flow dissipates as it slows down. So when we detect this sharp correlation between a congestion event and a subsequent drop in RTT, we can measure the magnitude of the drop and use it as a lower bound on queueing delay.

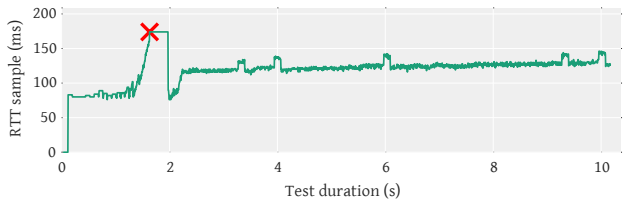


Figure 5: Example of the drop in RTT after a congestion event. The red cross marks the congestion event.

We limit the analysis to flows that have exactly one congestion event, and spend most of its lifetime being limited by the congestion window. Additionally, we exclude flows that are truncated or transfer less than 0.2 MB of data. For the remaining flows, we identify the pattern mentioned above by the following algorithm:

1. Find three values: *first_rtt*, the first non-zero RTT sample; *cong_rtt*, the RTT sample at the congestion event; and *cong_rtt_next*, the first RTT sample after the event that is different from *cong_rtt*.
2. Compute the differences between *first_rtt* and *cong_rtt* and between *cong_rtt* and *cong_rtt_next*. If both of these values are above 40 ms,[2] return the difference between *cong_rtt* and *cong_rtt_next*.

[2]The threshold is needed to exclude naturally occurring variation in RTT samples from the detection. We found 40 ms empirically to be a suitable conservative threshold: It is the lowest value that did not result in a significant number of false positives.

We add a few minor refinements to increase the accuracy of the basic algorithm above:[3]

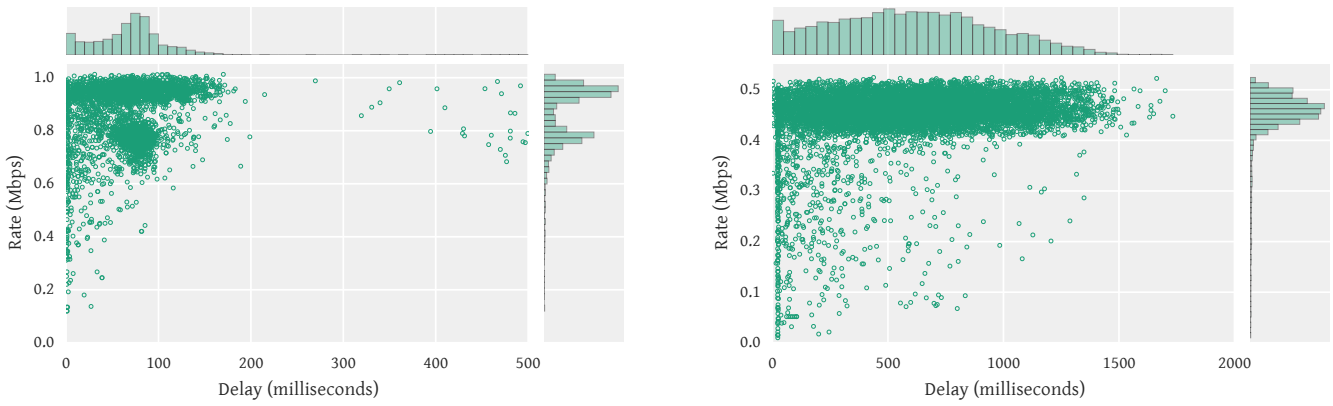
1. When comparing *first_rtt* and *cong_rtt*, use the median of *cong_rtt* and the two previous RTT samples. This weeds out tests where only a single RTT sample (coinciding with the congestion event) is higher than the baseline.
2. When comparing *cong_rtt* and *cong_rtt_next*, use the minimum of the five measurements immediately following *cong_rtt_next*. This makes sure we include cases where the decrease after the congestion event is not instant, but happens over a couple of RTT samples.
3. Compute the maximum span between the largest and smallest RTT sample in a sliding window of 10 data samples over the time period following the point of *cong_rtt_next*. If this span is higher than the drop in RTT after the congestion event, filter out the flow.

By applying the algorithm to the data from 2010 through 2014[4], we identified a total of 5.7 million instances of the RTT pattern, corresponding to 2.4% of the total number of flows. While this is a relatively small fraction of the flows, in this section we have aimed to be conservative and only pick out flows where we can algorithmically identify the source of the extra latency as queueing delay with a high certainty. This does not mean, however, that queueing delay cannot also be a source of latency variation for other flows.

Figure 4c shows the distribution of the magnitude of the detected queueing delay. We see that this follows a similar heavy-tailed distribution as the total latency variation. In

[3]The full code and dataset is published at http://www.cs.kau.se/tohojo/measuring-latency-variation/

[4]The Measurement Lab dataset was restructured in the middle of 2015, making it difficult to apply the detailed analysis for the 2015 data. For that reason, we have not included 2015 in these results.



(a) (b)

Figure 6: Delay and instantaneous outbound rate for two clients in the access network dataset. The histograms on the axes show the marginal distributions for the rate and delay samples.

addition, of those tests that our algorithm identifies as experiencing self-induced queueing, a significant percentage see quite a lot of it: 80% is above 100 ms, and 20% is above 400 ms. We see a downward trend in the queueing delay magnitude from 2010 to 2012/13, with a slight increase in 2014.

Based on our analysis of this subset of the whole dataset, we conclude that (i) queueing delay is present in a non-trivial number of instances and that (ii) when it does occur, it is of significant magnitude.

5.2 Delay correlated with load

A network is more likely to exhibit queueing delay when it is congested. Thus, delay correlated with load can be an indication of the presence of queueing delay. When analysing the access network dataset, we identified several cases where strong correlation between delay and load existed. In this section we look at two examples of this behaviour.

Figures 6a and 6b show the correlation between client side delay and the instantaneous outbound load during the 200 ms just preceding each delay sample. The sample period is one hour during peak time (20.30–21.30).

For the first client, we see two clusters of delay/load values, indicating two kinds of behaviour. There is one cluster just under a performance ceiling of 1 Mbit/s, but with increased round-trip delays up to almost 200 ms. This behaviour clearly indicates a saturated uplink where the upstream bottleneck limits the throughput and induces queueing latency. The other cluster is centred around about 0.8 Mbit/s and 80 ms increase in round-trip delay. This indicates an equilibrium where the outbound capacity is not the primary bottleneck, and so doesn't induce as much queueing delay. In addition to these clusters, there are some scattered data points up to just over 3 s increase in round-trip delay (not all of which are visible on the figure). The second client is clearly limited by a low uplink capacity, resulting in very large delays — up to about 1.5 s, which is consistent with the low capacity resulting in large queue drain times.

Together, the behaviour of these two client links (along with additional clients we have examined but not included here) clearly show that load-induced queueing delay is one source of the latency variation we have observed in the analysis of the whole dataset. According to the service provider, the access network otherwise does not have the amount of buffering needed for the delays we see in our measurements, pointing to large buffers in consumer equipment as the likely culprit.

5.3 Discussion

The analysis presented in this section indicates that excess queueing latency (i.e., bufferbloat) is indeed a very real and routinely occurring phenomenon. While we do not claim to have a means of accurately quantifying bufferbloat in all instances, we have sought to compensate for the lack of accuracy by erring on the side of caution in identifying bloat. And the fact that signs of bufferbloat so readily appears in both datasets constitutes a strong indicator that bufferbloat is indeed prevalent in real networks.

Another finding is that in the cases where bufferbloat does appear, it tends to be significant: most often on the order of several hundreds of milliseconds. This means that when bufferbloat does appear, it is quite noticeable and a considerable inconvenience for the end-user.

6. RELATED WORK

Several other studies have looked at the latency characteristics of internet traffic. These fall roughly into three categories: studies based on large-scale active measurements, studies based on targeted active measurements of a more limited scope, and passive measurements performed at various vantage points in the network. In this section we provide an overview of each of these categories in turn.

6.1 Large-scale active measurements

The *speedtest.net* measurement tool and the Netalyzr test suite are popular performance benchmark tools in wide use.

Canadi et al [7] perform a study based on 54 million test runs from the former which shows very low baseline (unloaded) latencies, but considers neither latency variation nor development over time. Kreibich et al [18] base their study on 130,000 tests from the latter, and show queueing latency on the order of hundreds of milliseconds, but does not consider differences over time or between regions.

Another approach to large-scale active measurements is taken by the BISMark and SamKnows measurement platforms, both of which provide instrumented gateways to users. A study based on this data by Sundaresan et al [26] measures baseline and under-load latency and shows significant buffering in head-end equipment. Chetty et al [8] also use BISMark data (as well as other sources) to measure broadband performance in South Africa. Consistent with our results for this continent, they find that latencies are generally high, often on the order of several hundred milliseconds.

Another large-scale active measurement effort is the Dasu platform [24], which is a software client users install on their machines. This study does not focus on latency measurements, but it includes HTTP latency figures which indicate a large regional variation, not unlike what we observe.

Finally, the M-Lab Consortium has studied ISP interconnections [19] using the same data as we use. However, this study only considers aggregate latency over large time scales.

6.2 Targeted active measurements

Dischinger et al [11] and Choy et al [10], both use active probing of residential hosts to measure network connections. The former study finds that queueing delay in the broadband equipment is extensive, while the latter finds that a significant fraction of users experience too high latency to run games in the cloud. Bischof et al [5] perform a slightly different type of measurements, piggy-backing on the BitTorrent protocol, and find a majority of the users see median last-mile latency between 10 and 100 ms, with the 95th percentile of seeing several hundred milliseconds. A similar conclusion, but specifically targeted at assessing queueing latency, is reached in [9], which estimates that 10% of users experience a 90th percentile queueing latency above 100 ms.

Another type of targeted active measurements are performed by clients under the experimenters' control to examine the access network. These types of experiments are performed by, e.g., Jiang et al [17] and Alfredsson et al [3] to measure bufferbloat in cellular networks. Both studies find evidence of bufferbloat on the order of several hundred ms.

6.3 Passive measurements

Several studies perform passive measurements of backbone or other high-speed links between major sites [12, 15, 16]. They generally find fairly low and quite stable latencies, with the backbone link experiencing latencies dominated by the speed of light, and the others generally seeing median latencies well below 100 ms. Jaiswal et al [15] additionally measure RTT variations and find that the median variation is around 2–300 ms and the 95th percentile variation is on the order of several seconds. In addition, Pathak et al [22] perform passive measurement of latency inflation in MPLS overlay networks and find that inflation is common, mostly due to path changes in the underlying tunnels.

Another technique for passive measurements consists of taking captures at the edge of a network and analysing that traffic. Aikat et al [2] and Allman [4] both employ this technique to analyse the RTT of TCP connections between hosts inside and outside the network where the measurement is performed. Both studies analyse the latency variation, Aikat et al finding it to be somewhat higher than Allman.

Another vantage point for passive measurements is at edge networks. Such studies are performed by Vacirca et al [27] and Maier et al [20] for mobile and residential networks, respectively. The former study finds that RTT can vary greatly over connections, while the latter finds that the baseline latency of the TCP handshake is dominated by the client part.

Finally, Hernandez-Campos and Papadopouli [13] compares wired and wireless traffic by means of passive packet captures. They find that wireless connections experience a much larger RTT variation than wired connections do.

7. CONCLUSIONS

We have analysed the latency variation experienced by clients on the internet by examining two complementary datasets from active measurement tests and from traffic captures from an ISP. In addition, we have analysed a subset of the data to attempt to determine whether load-induced queueing delay in the network can be part of the reason for the large variations. Based on our analysis, we conclude:

– Latency variation is both common and of significant magnitude, both within single connections and between connections to the same client. This indicates that it has a large potential to negatively affect end-user perceived performance.

– The worst affected clients see an order of magnitude larger variation than the least affected, and the choice of per-client percentile for the measured latency significantly affects the resulting conclusions. This indicates that latency spikes are transient and non-uniformly distributed.

– While throughput has increased over the six years we have examined, both minimum latency and latency variation have remained stable, and even increased slightly. This indicates that the improvements in performance afforded by the development of new technology are not improving latency.

– We find that in both datasets load-induced queueing delay is an important factor in latency variation, and quite significant in magnitude when it does occur. This indicates that load-induced queueing does in fact contribute significantly to the variation we see in overall latency behaviour.

8. ACKNOWLEDGEMENTS

We wish to thank Matt Mathis and Collin Anderson for suggesting we look at the Measurement Lab dataset, and for their helpful comments and technical expertise. We also wish to thank Henrik Abrahamsson for the use of his tools for computing instantaneous load from the access link data.

This work was in part funded by The Knowledge Foundation (KKS) through the SIDUS READY project.

9. REFERENCES

[1] NDT test methodology. Wiki page: https://goo.gl/KrqtbQ, 2015.

[2] J. Aikat, J. Kaur, F. D. Smith, and K. Jeffay. Variability in TCP round-trip times. In *3rd ACM SIGCOMM conference on Internet measurement*. ACM, 2003.

[3] S. Alfredsson, G. Del Giudice, J. Garcia, A. Brunstrom, L. De Cicco, and S. Mascolo. Impact of tcp congestion control on bufferbloat in cellular networks. In *IEEE 14th International Symposium and Workshops on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2013.

[4] M. Allman. Comments on bufferbloat. *ACM SIGCOMM Computer Communications Review*, January 2013.

[5] Z. S. Bischof, J. S. Otto, and F. E. Bustamante. Up, down and around the stack: Isp characterization from network intensive applications. *ACM SIGCOMM Computer Communication Review*, 2012.

[6] B. Briscoe, A. Brunstrom, A. Petlund, D. Hayes, D. Ros, I.-J. Tsang, S. Gjessing, G. Fairhurst, C. Griwodz, and M. Welzl. Reducing internet latency: A survey of techniques and their merits. *Communications Surveys Tutorials, IEEE*, 2014.

[7] I. Canadi, P. Barford, and J. Sommers. Revisiting broadband performance. In *2012 ACM conference on Internet measurement*. ACM, 2012.

[8] M. Chetty, S. Sundaresan, S. Muckaden, N. Feamster, and E. Calandro. Measuring broadband performance in south africa. In *4th Annual Symposium on Computing for Development*. ACM, 2013.

[9] C. Chirichella and D. Rossi. To the moon and back: are internet bufferbloat delays really that large? In *Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on*. IEEE, 2013.

[10] S. Choy, B. Wong, G. Simon, and C. Rosenberg. The brewing storm in cloud gaming: A measurement study on cloud to end-user latency. In *11th annual workshop on network and systems support for games*. IEEE Press, 2012.

[11] M. Dischinger, A. Haeberlen, K. P. Gummadi, and S. Saroiu. Characterizing residential broadband networks. In *Internet Measurement Conference*, 2007.

[12] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and S. C. Diot. Packet-level traffic measurements from the sprint ip backbone. *Network, IEEE*, 2003.

[13] F. Hernandez-Campos and M. Papadopouli. Assessing the real impact of 802.11 wlans: A large-scale comparison of wired and wireless traffic. In *14th IEEE Workshop on Local and Metropolitan Area Networks*. IEEE, 2005.

[14] T. Høiland-Jørgensen, P. Hurtig, and A. Brunstrom. The Good, the Bad and the WiFi: Modern AQMs in a residential setting. *Computer Networks*, Oct. 2015.

[15] S. Jaiswal, G. Iannaccone, C. Diot, J. Kurose, and D. Towsley. Inferring tcp connection characteristics through passive measurements. In *INFOCOM 2004*. IEEE, 2004.

[16] H. Jiang and C. Dovrolis. Passive estimation of tcp round-trip times. *ACM SIGCOMM Computer Communication Review*, 2002.

[17] H. Jiang, Y. Wang, K. Lee, and I. Rhee. Tackling bufferbloat in 3g/4g networks. In *2012 ACM conference on Internet measurement*. ACM, 2012.

[18] C. Kreibich, N. Weaver, B. Nechaev, and V. Paxson. Netalyzr: illuminating the edge network. In *10th ACM SIGCOMM conference on Internet measurement*. ACM, 2010.

[19] M. Lab. ISP interconnection and its impact on consumer internet performance. Technical report, Measurement Lab Consortium, October 2014.

[20] G. Maier, A. Feldmann, V. Paxson, and M. Allman. On dominant characteristics of residential broadband internet traffic. In *9th ACM SIGCOMM conference on Internet measurement*. ACM, 2009.

[21] M. Mathis, J. Heffner, and R. Raghunarayan. TCP Extended Statistics MIB. RFC 4898 (Proposed Standard), May 2007.

[22] A. Pathak, M. Zhang, Y. C. Hu, R. Mahajan, and D. Maltz. Latency inflation with mpls-based traffic engineering. In *2011 ACM SIGCOMM conference on Internet measurement*. ACM, 2011.

[23] V. Paxson, M. Allman, J. Chu, and M. Sargent. Computing TCP's Retransmission Timer. RFC 6298 (Proposed Standard), June 2011.

[24] M. A. Sánchez, J. S. Otto, Z. S. Bischof, D. R. Choffnes, F. E. Bustamante, B. Krishnamurthy, and W. Willinger. Dasu: Pushing experiments to the internet's edge. In *NSDI*, 2013.

[25] A. Singla, B. Chandrasekaran, P. Godfrey, and B. Maggs. The internet at the speed of light. In *13th ACM Workshop on Hot Topics in Networks*. ACM, 2014.

[26] S. Sundaresan, W. De Donato, N. Feamster, R. Teixeira, S. Crawford, and A. Pescapè. Broadband internet performance: a view from the gateway. In *ACM SIGCOMM computer communication review*. ACM, 2011.

[27] F. Vacirca, F. Ricciato, and R. Pilz. Large-scale RTT measurements from an operational UMTS/GPRS network. In *First International Conference on Wireless Internet*. IEEE, 2005.