Beyond Speed Test: Measuring Latency Under Load Across Different Speed Tiers

Kyle MacMillan
University of Chicago

Nick Feamster
University of Chicago

ABSTRACT

A range of performance metrics beyond throughput are increasingly becoming relevant for user experience. Notably, latency under load—the end-to-end latency of an Internet path when the network is loaded with traffic for a period of time—is a distinguishing feature, as increased latency, even for a short period of time, can disrupt connectivity for a wide range of applications. In this brief position paper, we use preliminary experiments from a home broadband measurement testbed across Chicago to demonstrate that latency under load can differ significantly, both for users across different speed tiers and for users within the same speed tier. We use this position paper to present a few compelling examples, to seed a discussion about ways to measure and compare latency under load across subscribers.

1 INTRODUCTION

When evaluating home Internet quality, much attention has been given to network capacity measurements derived from throughput measurements, or "speed tests". Although speed-test results can help to characterize the network, they are not sufficient to characterize end-user experience, particularly for applications that are latency-sensitive, such as video conferencing, gaming, and other interactive applications [4]. Past research has demonstrated that many latency-sensitive applications experience high latency under load when the network is experiencing high traffic load, if the bottleneck link has buffers that are too large (a condition commonly referred to as "bufferbloat" [3] [2] [1]). The bufferbloat phenomenon has motivated the development of various "latency-under-load" tests, which aim to measure network latency under operating conditions, particularly when the network capacity is saturated with other traffic.

As latency under load tests are becoming more prevalent, the importance of this performance metric is becoming apparent. In a preliminary deployment across tens of homes in Chicago, we have begun to measure latency under load for various users under a range of Internet service provider (ISP) speed tiers and service plans. Our initial measurements span only a small number of homes but have been conducted over a period of several months, yet these measurements already have yielded some interesting initial results:

- We see a potential relationship between the latency under load and the participant's Internet speed tier: users in our initial experiment who subscribe to lower throughput Internet service plans also experience higher latency under load.
- We have observed that users can experience dramatically different latency under load, even when they are subscribed to the *same* service plan with the same ISP.

The first result appears to be straightforward and possibly expected: given similar network equipment (and thus similar buffers), a slower bottleneck link would introduce higher latency under load, as the buffer would simply take longer to send queued packets. The second result, however, is more surprising and points to the broader complexities of measuring latency under load, in particular the possibility that the causes of high latency under load could be due to factors that are challenging to measure or isolate, such as the user's equipment, varying load between the different networks, ISP provisioning, and so forth. Given these curious results, we end this position paper with a call for increased attention to these challenges and a request for feedback on our ongoing design of latency under load tests.

2 EXPERIMENT SETUP

To measure network performance in the home, we have developed a network measurement suite, *Netrics*. Netrics is a suite of performance measurements that include various throughput tests, traceroutes, ping tests, and a latency under load test. Netrics can run on any Linux machine. For our initial deployment, we have installed the measurement suite on a Raspberry Pi that is connected directly to each participant's upstream connection (i.e., the user's cable modem or router) over a wired Ethernet connection. Netrics measures many network conditions; in this paper, we focus on the latency under load test design and frequency.

The latency under load test measures the round-trip time (RTT) of a packet to a given destination when either the upstream or downstream link is ostensibly saturated. We saturate the link by initiating an iPerf3 TCP speed test that connects to a server we control on the University of Chicago network. After the iPerf3 test begins, we wait 2 seconds before sending 10 ICMP pings to `8.8.8.8` (Google) every 250

Number of Participants	ISP A 1000/1000 ($n = 1$)	ISP A 1000/40 ($n = 4$)	ISP B 400/400 ($n = 1$)	ISP A 400/25 ($n = 2$)	ISP A 110/5 ($n = 2$)	ISP B 6/1 ($n = 1$)
Upstream						
Number of Tests	537	2705	601	2206	1549	323
Median (ms)	1.88	21.99	30.94	28.04	106.93	238.71
Standard Deviation (ms)	3.71	22.66	13.75	39.63	46.38	156.34
Max (ms)	49.81	242.45	59.31	823.90	334.21	1203.61
Downstream						
Number of Tests	540	2706	601	2210	1551	323
Median (ms)	1.69	15.27	12.37	27.65	61.00	161.62
Standard Deviation (ms)	3.85	8.68	9.38	47.93	24.16	174.48
Max (ms)	51.86	131.55	63.38	1694.212	646.55	1198.25

Table 1: *Latency under load statistics to Google DNS server (8.8.8.8). The maximum response time from each test is used to calculate the statistics. Subscriber plans are shown as (*downstream capacity / upstream capacity*), both in Mbps.*

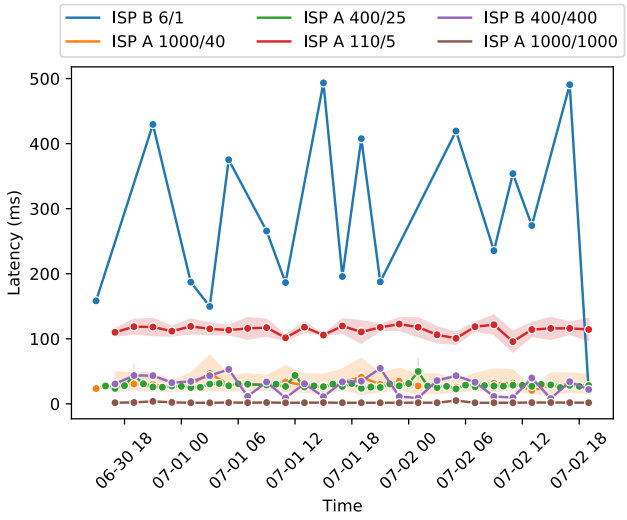
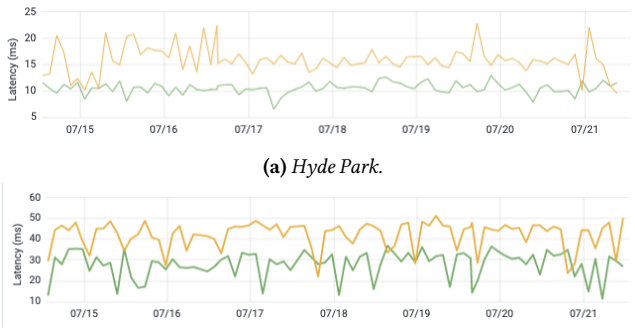


Figure 1: *Mean latency under upstream load over time, aggregated by subscriber plan and ISP with 95% confidence intervals.*



(a) *Hyde Park.*

(b) *South Shore.*

Figure 2: *Latency under load for two subscribers to Comcast's 1 Gbps service plan (an identical service plan), in two different Chicago neighborhoods.*

milliseconds. Upstream tests measure latency while the device is sending at capacity; downstream tests measure latency while the device is receiving at capacity.

We execute this test every two hours on 11 devices installed around Chicago. The tests are run asynchronously across the devices to ensure that the link connected to the iPerf3 server is not saturated by multiple devices on high-bandwidth networks. Each device has been installed for at least 30 days.

3 PRELIMINARY RESULTS

Users subscribed to Internet service plans with lower throughput generally experience higher latency under load than those on higher speed plans. For each latency under load test, we determine the ping with the highest response time of the 10 probes sent and break down the distribution of max response times by subscriber plan and ISP. Table 1 summarizes our findings.

The upstream results show that median latency increases with the provisioned upstream capacity, with the exception of the ISP B 400/400 participant, whose latency exceeds that of both the 1000/40 and 400/25 subscribers on ISP A. We also observe considerable increases in latency among those subscribed to $\leq$ 5Mbps. Figure 1 illustrates a snapshot of the sustained higher latency measured on the $\leq$ 5Mbps subscribers.

Although the latency increases are not as pronounced when the downstream link is saturated, we observe a similar trend, again with the exception that the 400/400 subscriber to ISP B achieves lower latency than the 1000/40 subscriber and less than half that of the 400/25 subscribers on ISP A. (The participant with the symmetric gigabit connection is a device connected to the UChicago network.)

Another interesting preliminary finding is that latency under load can differ significantly even for users subscribed to the same service plan, in the same geography. Figure 2, which shows a timeseries of latency under load, over the same time period, for two different Comcast subscribers *with the same service plan* (in this case, 1 Gbps downstream). For one user, latency under load is consistently in the 10–15 ms range; for the other user, latency under load hovers in the 40–50 ms range. Although these differences are clear, the *causes* of these differences are challenging to uncover.

4 FUTURE WORK

The preliminary results we have presented shed some preliminary light on the importance of measuring latency under load, as this metric becomes increasingly important to the quality of user experience for a broad range of applications. As part of our ongoing work, we are refining, standardizing, and releasing Netrics to allow others to deploy the tests that we have developed, including the latency under load test, under a larger user population and a broader set of participants. We are also actively expanding our study to hundreds of participants across Chicago. In addition to adding more participants, we plan to improve our latency-under-load test design to generate load that mimics that of common network applications such as video streaming or file transfer. Additionally, because ICMP pings yield inaccurate latency measurements, we also plan to measure latency using TCP pings.

REFERENCES

[1] Mark Allman. 2012. Comments on bufferbloat. *ACM SIGCOMM Computer Communication Review* 43, 1 (2012), 30–37.

[2] Vint Cerf, Van Jacobson, Nick Weaver, and Jim Gettys. 2011. BufferBloat: What's Wrong with the Internet? *Queue* 9, 12 (2011), 10–20.

[3] Jim Gettys. 2011. Bufferbloat: Dark buffers in the internet. *IEEE Internet Computing* 15, 3 (2011), 96–96.

[4] Srikanth Sundaresan, Walter De Donato, Nick Feamster, Renata Teixeira, Sam Crawford, and Antonio Pescapè. 2011. Broadband Internet performance: a view from the gateway. *ACM SIGCOMM Computer Communication Review* 41, 4 (2011), 134–145.