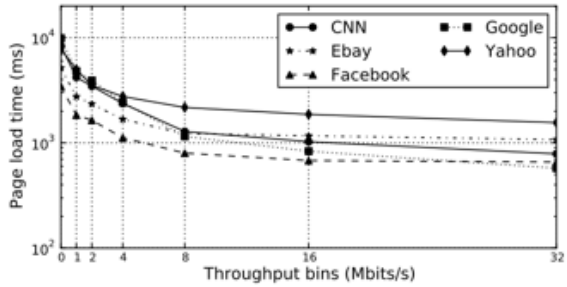
Hidden Sources of Internet Latency

Nick Feamster, *Georgia Tech*

Most Internet Service Providers advertise their performance in terms of downstream throughput. The "speed" that one pays for reflects, effectively, the number of bits per second that can be delivered on the access link into your home network. While this metric makes sense for many applications, it is only one characteristic of network performance that ultimately affects a user's experience. For example, consider the figure below, which shows Web page load times as downstream throughput increases—the time to load many Web pages decreases as throughput increases, but downstream throughput that is faster than about 16 Mbps stops having any effect on Web page load time.



The culprit is *latency*: For short, small transfers (as is the case with many Web objects), the time to initiate a TCP connection and open the initial congestion window is dominated by the round-trip time between the client and the Web server. In other words, the size of the access link no longer matters because TCP cannot increase its sending rate to "fill the pipe" before the connection has completed.

The role of latency in Web performance is no secret to anyone who has spent time studying it. Latency plays a role in the time it takes to complete a DNS lookup, the time to initiate a connection to the server, and the time to increase TCP's congestion window (indeed, students of networking will remember that TCP throughput is inversely proportional to the round-trip time between the client and the server). Thus, as throughput continues to increase, network latency plays an increasingly predominant role in the performance of applications such as the Web. Of course, latency also determines user experience for many latency-sensitive applications as well, including streaming voice, audio, video, and gaming.

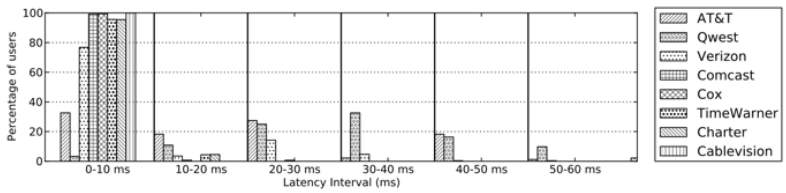
The question, then, becomes how to reduce latency to the destinations that users commonly access. Content providers such as Google and others have taken several approaches: (1) placing Web caches closer to users; (2) adjusting TCP's congestion control mechanism to start sending at a faster rate for the first few round trips. These steps, however, are only part of the story, because the network performance between the Web cache and the user may still suffer, for a variety of reasons:

- First, factors such as bufferbloat and DSL interleaving can introduce significant latency effects in the last mile. Our study from *SIGCOMM* 2011 showed how both access link configuration and a user's choice of equipment (e.g., DSL modem) can significantly affect the latency that a user sess.
- Second, a poor wireless network in the home can introduce significant latency effects; sometimes we see that 20% of the latency for real user connections from homes is within the home itself.
- Finally, if the Web cache is not close to users in the first place (e.g., in the case of developing countries), the paths between the users and their destinations can still be subject to significant latency. These factors can be particularly evident in developing countries, where poor peering and interconnection can result in long paths to content, and where the vast majority of users access the network through mobile and cellular networks.

In the Last Mile

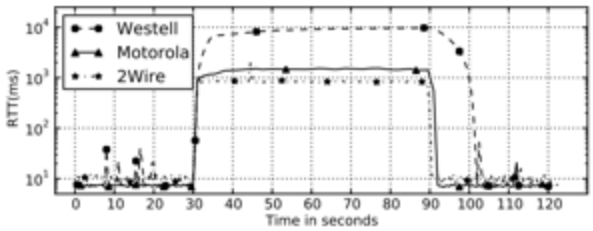
In our *SIGCOMM* 2011 paper "Broadband Internet Performance: A View from the Gateway", we pointed out several aspects of home networks that can contribute significantly to latency. We define a metric called *last-mile latency*, which is the latency to the first hop inside the ISP's network. This metric captures the latency of the access link.

We found in this study that last-mile latencies are often quite high, varying from about 10 ms to nearly 40 ms (ranging from 40–80% of the end-to-end path latency). Variance is also high. One might expect that variance would be lower for DSL, since it is not a shared medium like cable. Surprisingly, we found that the opposite was true: Most users of cable ISPs have last-mile latencies of 0–10 ms. On the other hand, a significant proportion of DSL users have baseline last-mile latencies more than 20 ms, with some users seeing



last-mile latencies as high as 50 to 60 ms. Based on discussions with network operators, we believe DSL companies may be enabling an interleaved local loop for these users. ISPs enable interleaving for three main reasons: (1) the user is far from the DSLAM; (2) the user has a poor quality link to the DSLAM; or (3) the user subscribes to "triple play" services. An interleaved last-mile data path increases robustness to line noise at the cost of higher latency. The cost varies between two to four times the baseline latency. Thus, cable providers in general have lower last-mile latency and jitter. Latencies for DSL users may vary significantly based on physical factors such as distance to the DSLAM or line quality.

Customer provided equipment also plays a role. Our study confirmed that excessive buffering is a widespread problem afflicting most ISPs (and the equipment they provide). We profile different modems to study how the problem affects each of them. We also see the possible effect of ISP policies, such as active queue and buffer management, on latency and loss. For example, when measuring *latency under load* (the latency that a user experiences when the access link is saturated due to an upload or a download), we see more than an order of magnitude of difference between modems. The 2Wire modem we tested had the lowest worst-case last-mile latency, 800 ms. Motorola's was about 1.6 seconds, and the Westell modem we tested had a worst case latency of more than 10 seconds.



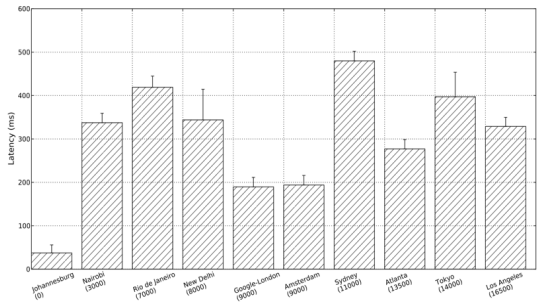
Last-mile latency can also be high for particular technologies such as mobile. In a recent study of fixed and mobile broadband performance in South Africa, we found that, although the mobile providers consistently offer higher throughput, the latency of mobile connections is often 2–3x higher than that of fixed-line connectivity in the country.

In the Home Wireless Network

Our recent study of home network performance found that a home wireless network can also be a significant source of latency. We have recently instrumented home networks with a passive monitoring tool that determines whether the access link or the home wireless network (or both) are potential sources of performance problems. One of the features that we explored in that work was the TCP round-trip time between wireless clients in the home network and the wireless access point in the home. In many cases, due to wireless contention or other sources of wireless bottlenecks, the TCP round-trip latency in home wireless networks was a significant portion of the overall round-trip latency. In a study across about 65 homes over one month we found that for 30% of devices in those homes, at least half of the flows have end-to-end latencies where the home wireless network contributes more than 20% of the overall end-to-end latency.

In Developing Regions

Placing content in a Web cache has little effect if the users accessing the content still have high latency to those destinations. Our study of latency from access networks in South Africa showed that peering and interconnectivity within the country still has a long way to go: in particular, the plot below shows the average latency from 16 users of fixed-line access networks in South Africa to various Internet destinations. The bars are sorted in order of increasing distance from Johannesburg, South Africa. Notably, geographic distance from South Africa does not correlate with latency—the latency to Nairobi, Kenya is almost twice as much as the latency to London. In our study, we found that users in South Africa experienced average round-trip latencies exceeding 200 ms to five of the ten most popular websites in South Africa: Facebook (246 ms), Yahoo (265 ms), LinkedIn (305 ms), Wikipedia (265 ms), and Amazon (236 ms). Many of these sites only have data centers in Europe and North America.



People familiar with Internet connectivity may not find this result surprising: indeed, many ISPs in South Africa connect to one another via the London Internet Exchange (LINX) or the Amsterdam Internet Exchange (AMS-IX) because it is cheaper to backhaul connectivity to exchange points in Europe than it is to connect directly at an exchange point on the African continent. The reasons for this behavior appears to be both regulatory and economic, but more work is needed, both in deploying caches and improving Internet interconnectivity to reduce the latency that users in developing regions see to popular Internet content.