Why is the Internet so slow?!

Ilker Nadi Bozkurt[♯], Anthony Aguirre[♣], Balakrishnan Chandrasekaran[§⋆],
P. Brighten Godfrey[†], Gregory Laughlin[‡], Bruce Maggs[♯◦], and Ankit Singla[¶]

[♯]Duke University, [§]TU Berlin, [†]UIUC, [♣]UC Santa Cruz,
[‡]Yale University, [◦]Akamai, [¶]ETH Zürich
{ilker,bmm}@cs.duke.edu, aguirre@scipp.ucsc.edu, balac@inet.tu-berlin.de, pbg@illinois.edu,
greg.laughlin@yale.edu, ankit.singla@inf.ethz.ch

Abstract. In principle, a network can transfer data at nearly the speed of light. Today's Internet, however, is much slower: our measurements show that latencies are typically more than one, and often more than two orders of magnitude larger than the lower bound implied by the speed of light. Closing this gap would not only add value to today's Internet applications, but might also open the door to exciting new applications. Thus, we propose a grand challenge for the networking research community: building a speed-of-light Internet. To help inform research towards this goal, we investigate, through large-scale measurements, the causes of latency inflation in the Internet across the network stack. Our analysis reveals an under-explored problem: the Internet's infrastructural inefficiencies. We find that while protocol overheads, which have dominated the community's attention, are indeed important, reducing latency inflation at the lowest layers will be critical for building a speed-of-light Internet. In fact, eliminating this infrastructural latency inflation, without any other changes in the protocol stack, could speed up small object fetches by more than a factor of three.

1 Introduction

Measurements and analysis by Internet giants have shown that shaving a few hundred milliseconds from the time per transaction can translate into millions of dollars. For Amazon, a 100 ms latency penalty implies a 1% sales loss [18]; for Google, an additional delay of 400 ms in search responses reduces search volume by 0.74%; and for Bing, 500 ms of delay decreases revenue per user by 1.2% [10,13]. The gaming industry, where latencies larger than even 80 ms can hurt gameplay [19], has even tougher latency requirements. These numbers underscore that latency is a key determinant of user experience.

We take the position that the networking community should pursue an ambitious goal: cutting Internet latencies to close to the limiting physical constraint, the speed of light, roughly one to two orders of magnitude faster than today. Beyond the obvious gains in performance and value for today's applications, such a technological leap may help realize the full potential of certain applications that have so far been confined to the laboratory, such as tele-immersion. For some applications, such as massive multi-player online games, the size of the user community reachable within a latency bound plays an important role in user interest and adoption, and linear decreases in communication latency result in super-linear growth in community size [25]. Low latencies on the order of a few tens of milliseconds also open up the possibility of *instant response*, where

[⋆] This work was done when the author was a graduate student at Duke University users are unable to perceive any lag between requesting a page and seeing it rendered in their browsers. Such an elimination of wait time would be an important threshold in user experience.

But the Internet's speed is quite far from the speed of light. As we show later, the time to fetch just the HTML document of the index pages of popular Web sites from a set of generally well-connected clients is, in the median, 37 times the round-trip speed-of-light latency. In the 80^{th} percentile it is more than 100 times slower. Given the promise a speed-of-light Internet holds, *why are we so far from the speed of light?*

While ISPs compete primarily on the basis of peak bandwidth offered, bandwidth is no longer the bottleneck for a significant fraction of the population: for instance, the average Internet connection speed in the US is 15.3 Mbps [9], while the effect of increasing bandwidth on page load time is small beyond as little as 5 Mbps [17]. If bandwidth isn't the culprit, then what is? In our short workshop paper [25], we staked out our vision of a speed-of-light Internet, discussed why it is a worthy goal to pursue, and, provided a preliminary analysis of latency inflation across the network stack. In this work, we present a more thorough analysis of latency inflation using three new data sets. Our contributions are as follows:

1. We quantify the factors that contribute to large latencies today using four sets of measurements: from PlanetLab nodes to Web servers[1]; between a large CDN's servers and end hosts; from volunteer end-user systems[2] to Web servers; and between RIPE Atlas nodes. Our analysis breaks down Internet latency inflation across the network stack, from the physical network infrastructure to the transport layer (including, in some instances, TLS).
2. This work places in perspective the importance of latency inflation at the lowest layers. While in line with the community's understanding that DNS, TCP handshake, and TCP slow-start are all important factors in latency inflation, the Internet's infrastructural inefficiencies are also important. We consider this an under-appreciated piece of the latency puzzle.
3. We find that removing latency inflation in the physical infrastructure and routing without *any* changes at layers above, could improve latencies for fetching small objects by more than 3 times.

2 The Internet is too slow

We pooled the top 500 Web sites from each of 138 countries listed by Alexa [7]. We followed redirects on each URL, and recorded the final URL for use in our measurements; the resulting data set contains 22,800 URLs. We fetched just the HTML at these URLs from 102 PlanetLab locations using cURL [1], and 25% of all fetches in our experiments were over HTTPS[3].

For each connection (or fetch), we geolocated the Web server using six commercial geolocation services, and (since we do not have any basis for deciding which service

[1] Data sets (gathered in 2016) and code are available at `https://cgi.cs.duke.edu/~ilker/cspeed/pam2017-data/`

[2] Explicit volunteer consent was obtained, listing precisely what tests would be run. We have a letter from the IRB stating that our tests did not require IRB approval.

[3] We do not claim this is the percentage of Web sites supporting HTTPS.

is better than another) used the location identified by their majority vote (MV). We computed the time it would take for light to travel round-trip along the shortest path between the same end-points, *i.e.*, the c-latency. Finally, we calculated the Internet's latency inflation as the ratio of the fetch time to c-latency. Fig. 1(a) shows the CDF of inflation over 1.9 million connections. The HTML fetch time is, in the median, 36.5 times the c-latency, while the 80^{th} percentile exceeds 100 times. We note that PlanetLab nodes are generally well-connected, and latency can be expected to be poorer from the network's *true* edge. We verify that this is indeed the case with measurements from end users in §3.7.

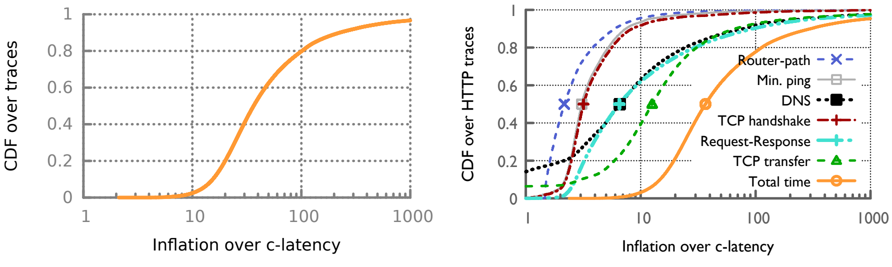
3 Why is the Internet so slow?

To identify the causes of Internet latency inflation, we break down the fetch time across layers, from inflation in the physical path followed by packets to the TCP transfer time.

3.1 Methodology

We use cURL to obtain the time for DNS resolution, TCP handshake, TCP data transfer, and total fetch time for each connection. For HTTPS connections, we also record the time for TLS handshake. TCP handshake is measured as the time between cURL sending the `SYN` and receiving the `SYN-ACK`. The TCP transfer time is measured as the time from cURL's receipt of the first byte of data to the receipt of the last byte. We separately account for the time between cURL sending the data request and the receipt of the first byte as 'request-response' time; this typically comprises one RTT and any server processing time. For each connection, we also run a traceroute from the client PlanetLab node to the Web server. We then geolocate each router in the traceroute path, and connect successive routers with the shortest paths on the Earth's surface as an optimistic approximation for the route the packets follow. We compute the round-trip latency at the speed of light in fiber along this approximate path, and refer to it as the 'router-path latency'. From each client, we also run 30 successive pings to each server, and record the minimum and median across these ping times. We normalize each of these latency components by the c-latency between the respective connection's end-points.

Our experiments yielded 2.1 million page fetches with HTTP status code 200, which corresponds to 94% of all fetches. We also filtered out connections which showed obvious anomalies such as c-latency being larger than TCP handshake time or minimum ping time (probably due to errors in geolocation), leaving us with 1.9 million fetches.



(a) (b)

Fig. 1: *(a) Inflation in fetch time, and (b) its breakdown across various components of HTTP fetches of just the HTML of the landing pages of popular Web sites.*

3.2 Overview of results

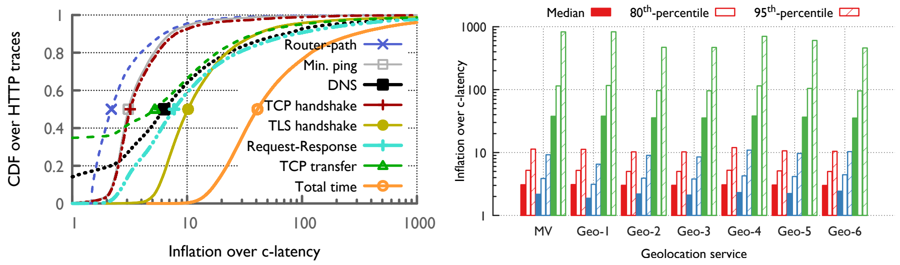
Fig. 1(b) shows the results for all connections over HTTP. DNS resolutions are shown to be faster than c-latency 14% of the time. This is an artifact of the baseline we use—in these cases, the Web server happens to be farther than the DNS resolver, and we always use the c-latency to the Web server as the baseline. (The DNS curve is clipped at the left to more clearly display the other results.) In the median, DNS resolutions are $6.6\times$ inflated over c-latency.

The TCP transfer time shows significant inflation—12.6 times in the median. With most pages being at most tens of KB (median page size is 73 KB), bandwidth is not the problem, but TCP's slow start causes even small data transfers to require several RTTs. 6% of all pages have transfer times less than the c-latency—this is due to all the data being received in the first TCP window. The TCP handshake (counting only the `SYN` and `SYN-ACK`) and the minimum ping time are 3.2 times and 3.1 times inflated in the median. The request-response time is 6.5 times inflated in the median, *i.e.*, roughly twice the median RTT. However, 24% of the connections use less than 10 ms of server processing time (estimated by subtracting one RTT from the request-response time).The median c-latency, in comparison, is 47 ms. The medians of inflation in DNS time, TCP handshake time, request-response time, and TCP transfer time add up to 28.8 times, lower than the *measured* median total time of 36.5 times, since the distributions are heavy-tailed.

Fig. 2(a) shows the results for fetches over HTTPS. The inflations in DNS resolution and TCP handshake are similar to those for HTTP (6.3 times and 3.1 times in the median respectively). The largest contributor to the latency inflation is the TLS handshake, which is 10.2 times inflated in the median, roughly corresponding to 3 RTTs. Inflation in TCP transfer time, being 5.2 times in the median, is significantly lower than for HTTP connections. This difference is partly explained by the smaller size of pages fetched over HTTPS, with the median fetch size being 43 KB. The median inflation in request-response times increases from 6.5 times for HTTP to 7.7 times for HTTPS.

3.3 Impact of IP geolocation errors

The correctness of our latency inflation analysis crucially depends on geolocation. While we cull data with obvious anomalies, such as when the min. ping time is smaller than c-latency, arising from geolocation errors (some of which may be due to Anycast),



(a) (b)

Fig. 2: *(a) Various components of latency inflation over HTTPS connections, and (b) the median, 80^{th}% and 95^{th}% of inflation in min. ping (red), router-path (blue) and total (green) latency using* 6 *different geolocation databases as well as their majority vote.*

less obvious errors could impact our results. For PlanetLab node locations, we have ground truth data and our tests did not indicate any erroneous location. Retrieving similar ground truth data for the large IP space under consideration appears infeasible. Thus, we focused our efforts on comparing the results we obtained by using 6 different commercial IP geolocation services, as well as a location computed as their majority vote. We computed latency inflation in router-path, minimum ping, and total time using each of these 7 sets of IP geolocations (Fig. 2(b)). As we might expect, router-path latency (blue) is most susceptible to differences in IP geolocation—the result there depends on geolocating not only the Web server, but also each router along the path. Even so, all 6 median inflation values are in the 1.9-2.4 times range. Differences in results for minimum ping time (red) and total time (green) are much smaller. Even the 95^{th}-percentile values for inflation in minimum ping time all lie within 10.4-12.0 times, while the medians lie within 3.0-3.1 times. The results for median inflation in total time all lie between 35.5-38.0 times, but variation at the higher percentiles is larger. Thus, largely, our conclusions, particularly with regards to median values are robust against the significant differences in the geolocations provided by these services. Needless to say, we cannot, without ground truth, account for systematic errors that may impact all geolocation services. Except in Fig. 2(b), we use the majority vote geolocation throughout.

On a related note, small client-server distances can cause a small absolute latency increase to translate into a large inflation over c-latency; effect of geolocation errors can also be more pronounced at short distances. When we restricted our analysis to connections with client-server distances above 100 km, 500 km and 1000 km, we found that the median inflations are relatively close to each other, being 35.5, 33.7 and 31.9 respectively. So, large inflations are not just caused by short distances, and even after limiting ourselves to connections at long distances we observe significant inflation. §3.5 (and Fig. 3(a)) discusses in more detail on the relationship between latency inflation and client-server distances (equivalently, c-latency) and locations.

3.4 Results across page sizes

While we fetch only the HTML for the landing pages of Web sites in our experiments, some of these are still larger than 1 MB. Most pages, however, are much smaller, with the median being 67 KB. To analyze variations in our results across page sizes, we binned pages into 1 KB buckets, and computed the median inflation for each latency component across each bucket. While the median inflation in minimum ping time shows little variation, inflation in TCP transfer time increases over page sizes in an expected linear fashion, also causing an increase in total fetch time.

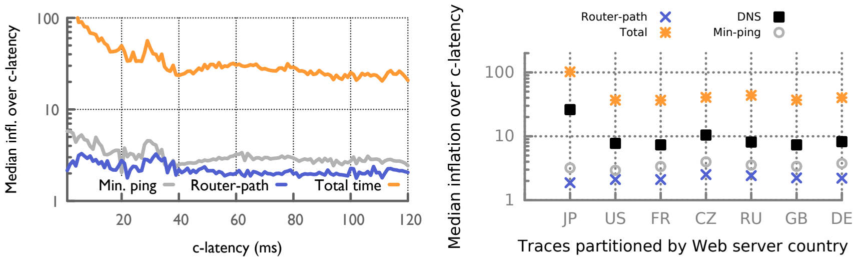
We also examine latency inflation in a narrow range of Web page sizes around the median, using pages within 10% of the median size of 67KB. These pages comprise roughly 7% of our data set. The results of this analysis are similar to the overall results in Fig. 1(b), with expected differences in the transfer time (8% smaller) and total time (5% smaller). The request-response time is 10% larger. Other components of inflation are within 1% of the corresponding values in Fig. 1(b).

3.5 Results across geographies

We fetch pages in 138 countries from 81 unique PlanetLab locations, leading to a wide spread in the pairwise c-latencies observed across these connections. The median c-

latency is 47 ms, with 5^{th} and 95^{th} percentiles being 2 ms and 101 ms respectively. In a manner similar to our analysis across page sizes, we also analyzed latency inflation in router-path latency, minimum ping time, and total time across c-latencies (Fig. 3(a)).

An interesting feature of these results is the inflation bump around a c-latency of 30 ms. It turns out that some countries connectivity to which may be more circuitous than average, are over-represented at these distances in our data. For instance, c-latencies from the Eastern US to Portugal are in the 30 ms vicinity, but all transatlantic connectivity hits Northern Europe, from where routes may go through the ocean or land southward to Portugal, thus incurring significant path 'stretch'. That the differences are largely due to inflation at the lowest layers is also borne out by the inflation in minimum ping and total time following the inflation in the router-path latency.



(a) (b)

Fig. 3: *Inflation in router-path latency, minimum ping time, and total time: (a) as a function of c-latency; and (b) as a function of Web server country.*

An encouraging observation from Fig. 3(a) is that the inflation in minimum ping and total time follows the inflation in the router-path latency. Thus, despite the router-path latency estimation containing multiple approximations (omitting routers that did not respond to traceroutes or we could not geolocate, as well as paths between successive routers themselves potentially being circuitous), it is a useful quantity to measure.

To compare measurements from a geographically balanced set of client locations, we selected 20 PlanetLab hosts such that no two were within 5 degrees of longitude of each other. Then we looked at requests from these PlanetLab clients to Web servers in each country. Fig. 3(b) shows the median inflation in router-path latency, minimum ping time, DNS, and total time across each of the 7 countries for which we had 5,000+ connections. The median c-latencies from these selected PlanetLab hosts to each of these 7 countries all lie in the 48-55 ms range, with the exception of Japan (12 ms). Most of the latencies are fairly consistent across geographies, with the exception of DNS and total time for Japan. We observed that roughly half of the requests to Web servers in Japan come from two PlanetLab nodes in Japan, and it is likely that DNS resolvers are further away than the Web servers causing the larger inflation.

3.6 The role of congestion

Fig. 1(b) shows that TCP transfer time is more than 10 times inflated over c-latency. It is worth considering whether packet losses or large packet delays and delay variations are to blame for poor TCP performance. Oversized and congested router buffers on the path may exacerbate such conditions—a situation referred to as *bufferbloat*.

In addition to fetching the HTML for the landing page, for each connection, we also sent 30 pings from the client to the server's address. We found that variation in ping times is small: the 2^{nd}-longest ping time is only 1.1% larger than the minimum ping time in the median. While pings (using ICMP) might use queues separate from Web traffic, even the TCP handshake time is only 1.6% larger than the minimum ping time in the median. We also used `tcpdump` at PlanetLab clients to analyze the inter-arrival times of packets. More than 92% of the connections we made experienced no packet loss (estimated as packets reordered by more than 3 ms). These results are not surprising—PlanetLab nodes are (largely) well-connected, university-based infrastructure, and likely do not have similar characteristics in terms of congestion and last-mile latency to typical end-user systems.

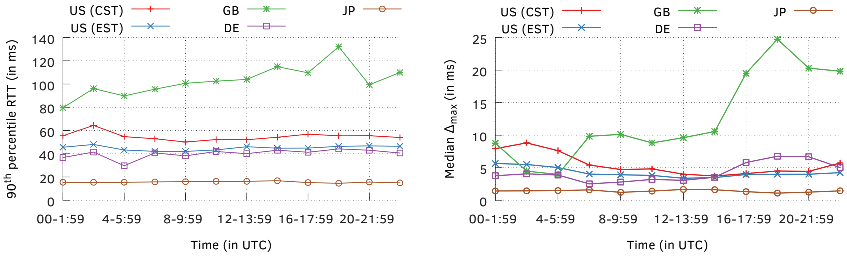
3.7 End-user Measurements

To complement our PlanetLab measurements, in this section we present results from three sets of measurements from the *real edge* of the network.

Client connections to a CDN For a closer look at congestion, we examined RTTs in a sample of TCP connection handshakes between the servers of a large CDN and clients (end users) over a 24-hour time period, passively logged at the CDN. (Most routes to popular prefixes are unlikely to change at this time-scale in the Internet [24].) We exclude server-client pairs with minimum latencies of less than 3 ms—'clients' in this latency range are often proxy servers in a data center or colocation facility rather than our intended end users.

To evaluate the impact of congestion, we examine our data for both variations across time-of-day (perhaps latencies are, as a whole, significantly larger in peak traffic hours), and within short periods of time for the same server-client pairs (perhaps transient congestion for individual connections is a significant problem). Thus, we discard server-client pairs that do not have repeat measurements. We only look at server-client pairs in the same timezone to simplify the time-of-day analysis. Server locations were provided to us by the CDN, and clients were geolocated using a commercial geolocation service. We include results for a few geographies that have a large number of measurements after these restrictions. We bin all RTT measurements into 12 2-hour periods, separately for each country, and produce results aggregated over these bins.

Time-of-day latency variations across bins: We selected server-client pairs that have at least one RTT measurement in each of the twelve bins. For pairs with multiple RTTs within a bin, we use the median RTT as representative, discarding other measurements. This leaves us with the same number of samples between the same host-pairs in all bins. Fig. 4(a) shows the 90^{th} percentile of RTTs in each 2-hour bin for each of 5 timezones. For the United States (US), we show only data for the central (CST) and eastern (EST) timezones, but the results are similar for the rest. The timezone classification is based on the location of the client; servers can be anywhere in the US and not necessarily restricted to the same timezone as that of the clients. Median latency across our aggregate (not shown) varies little across the day, most timezones seeing no more than 3 ms of variation. The 90^{th} percentile in each bin (Fig. 4(a)) shows similar trends, although with larger variations. In Great Britain, RTTs are higher in the evening (and results for a



(a) (b)

Fig. 4: *Variations in latencies of client-server pairs grouped into 2-hr windows in different geographic regions: (a)* 90^{th} *percentile of RTTs of client-server pairs with measurements in each 2-hr window; and (b) medians of maximum change in RTTs (max - min) in repeat measurements within each time window.*

different 24-hour period look similar.) It is thus possible that congestion is in play there, affecting network-wide latencies. But across other timezones, we see no such effect.

Transient latency variations within bins: To investigate transient congestion, we do not limit ourselves to measurements across the same set of host-pairs across all bins. However, within each bin, only data from host-pairs with multiple measurements inside that time period is included. For each host-pair in each bin, we calculate the maximum change in RTT (Δ_{max})—the difference between the maximum and minimum RTT between the host-pair in that time period. We then compute the median Δ_{max} across host-pairs within each bin. The variation within bins (in Fig. 4(b)) is a bit larger than variations across median latencies across the day, e.g., for US (CST), the median Δ_{max} is as large as 9 ms in the peak hours. That Δ_{max} also shows broadly similar time-of-day trends to median latency is not surprising. Great Britain continues to show exceptionally large latency variations, with a $\Delta_{max}\simeq$25ms at the peak, and also large variations across the day. In summary, in end-user environments, network-wide latency increases in peak hours were largely limited in our data set to one geography (GB). However, individual flows may sometimes experience a few additional milliseconds of latency.

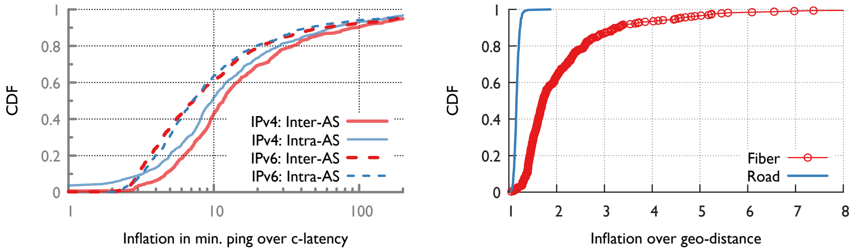
MOOC-recruited end users 678 students in a Massive Open Online Course (MOOC) run by two of the authors volunteered to run experiments for us. The experiments are identical to our PlanetLab experiments, but performed with a smaller list of Web pages. Each volunteer fetched (only the HTML of) 50 pages, with a fixed set of 25 pages for all the participants and another 25 chosen randomly from a handpicked, safe, set of 100 URLs. We deliberately chose a small number of Web sites so that each volunteer could look at the provided descriptions, and make an informed decision to participate. We also asked each volunteer to provide their location and various characteristics of their Internet service such as download speed and connection type.

A total of 24,784 pages were fetched in these experiments. The latency inflation measured in these experiments was much larger than in our PlanetLab data set—even after filtering out connections between clients and servers within a 100 km distance of each other, we found that total fetch time is 66 times inflated in the median. One reason for this significantly larger latency inflation is the over-representation of shopping and news Web sites in the handpicked URLs, resulting in larger HTML pages, with the median fetch size being 148 KB. To investigate further, we also computed results over the same set of pages by fetching them from PlanetLab. Over this set, with the same filtering (client-server distances of at least 100 km), median inflation in total fetch time is 49.4 times. This is still smaller than the measurements from the volunteer systems.

Another factor causing this difference is the larger latency inflation in minimum ping time: 4.1 times in the median over the volunteer-runs, compared to 3 times over PlanetLab (over this set of URLs). Of course, if each RTT is longer in this way, the total fetch time will also be longer. In fact, both numbers differ by roughly a factor of $4/3$.

One possible reason of larger inflation in minimum ping time in the end-user experiments is the connection type of the user, affecting the last mile latency. Even though our data is small, we get a glimpse of the situation when we compare different user provided connection types in terms of inflation of minimum ping time over c-latency. The lowest median inflation (3.76) is observed over connections users described as Company/University network, whereas the worst median inflations are observed for mobile and DSL connections, for which minimum ping time is inflated 5.4× and 5.2× respectively in the median.

RIPE Atlas So far, we have limited ourselves to client-server connections, where the server belongs to a popular Web service. In this section, we describe our measurements between RIPE Atlas platform [6] *probes*, which are small network devices that are typically deployed in end-user networks. The locations of the RIPE Atlas probes are known within 1 km resolution, obviating the need for IP geolocation.



(a) (b)

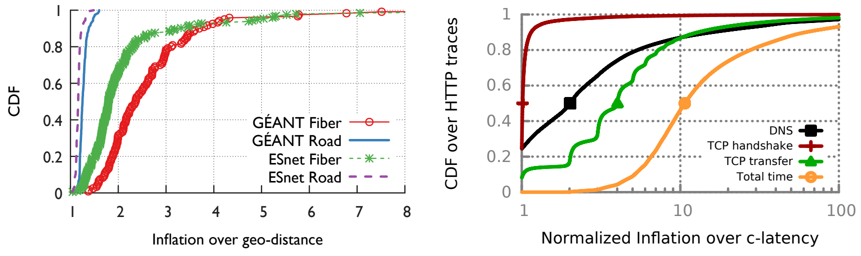
Fig. 5: *(a) Minimum pings between RIPE Atlas nodes are highly inflated regardless of IPv4 or IPv6, inter- or intra-AS connections; (b) comparison of fiber lengths of the Internet2 network to road distances.*

We collected ICMP pings over IPv4 (IPv6) between 935 (1012) sources in 26 (34) countries and 72 (97) destinations in 29 (40) countries every 30 minutes for 24 hours. The data set contains ping measurements between 288,425 (63,884) unique IPv4 (IPv6) endpoint (or source-destination) pairs; 85% (78%) of the IPv4 (IPv6) endpoint pairs are inter-AS pairs with the source and destination belonging to different ASes. To account for the skew in inter-AS and intra-AS pairs, we compute the round-trip distance between the endpoints and bin them into 5 km wide buckets. From each bucket, we uniformly sample an equal number of inter-AS and intra-AS pairs and compute the inflation of the min. pings (minimum across the entire day of measurements) of these endpoint pairs. Fig. 5(a) shows that the median inflation in minimum ping times (ranging from 7.2-11.6 times) is significantly larger than that in our PlanetLab measurements, where median inflation in minimum ping latency was 3.1 times. That the latencies between Atlas probes (typically attached to home networks) are larger than that between PlanetLab nodes and Web servers should not be suprising—home networks surely add more latency than servers in a university cluster or a data center. Perhaps paths from clients to Web servers are also much shorter than between arbitrary pairs of end-points on the Internet, since Web servers are deliberately deployed for fast access, and the Internet's semi-hierarchical nature can make paths between arbitrary end-points long. Interference from concurrent measurements may also be a contributing factor [16], albeit the effect on inflation might be marginal.

4 Infrastructural latency

In line with the community's understanding, our measurements affirm that TCP transfer and DNS resolution are important factors causing latency inflation. However, as we shall detail in this section, our measurements also reveal that the Internet's infrastructural inefficiencies are an equally, if not more important culprit.

In Fig. 1(b), the router-path is only 2.1 times inflated in the median. The long tail is, in part, explained by 'hairpinning', *i.e.*, packets between nearby end points traversing circuitous routes across the globe. Note that 1.5 times inflation would occur even along the shortest path along the Earth's surface because the speed of light in fiber is roughly $2/3^{rd}$ the speed of light in vacuum. In that light, the router-path inflation of 2.1× (which already includes the 1.5× factor) may appear small, but this estimate is optimistic.



(a) (b)

Fig. 6: *(a) Comparison of fiber lengths of the ESnet and GÉANT network to road distances; (b) various components of latency inflation normalized by minimum ping time.*

The gap between minimum ping time and the router-path latency may be explained by two factors: (a) we perhaps see artificially shorter paths, since traceroute often does not yield responses from all the routers on the path; and (b) even between successive routers, the physical path may be longer than the shortest arc along the Earth's surface. We investigate the latter aspect using data from 3 research networks: Internet2 [5], ESnet [2], and GÉANT[3]. We obtained point-to-point fiber lengths for these networks and calculated end-to-end fiber distances between all pairs of end points in each network. We also computed the shortest distance along the Earth's surface between each pair of end points, and obtained the road distances for comparison using the Google Maps API [4]. In Fig. 6(a), road distances are close to shortest distances (i.e., smaller inflations), while fiber lengths are significantly larger and have a long tail. The median inflation in the three networks, after accounting for the lower speed of light in fiber is 2.6×

(Internet2), 2.7× (ESnet), and 3.6× (GÉANT). A recent analysis of US long-haul fiber infrastructure [12] found results that support ours: even for cities directly connected by fiber conduits, the mean conduit's latency was in the median more than 2× worse than the line-of-sight latency. Of course, we expect end-to-end inflation between cities not connected directly to be higher. Thus, infrastructural inflation (which includes routing sub-optimalities and inflation of end-to-end fiber-distances over geodistance) is likely to be larger than the optimistic estimate from router-path latency (2.1 times), bringing it closer to the inflation in minimum ping latency (3.1 times).

As Fig. 1(b) shows, DNS resolution (6.6× inflated over c-latency), TCP handshake (3.2×), request-response time (6.5×), and TCP transfer (12.6×), all contribute to a total time inflation of 36.5×. With these numbers, it may be tempting to dismiss the 3.1× inflation in the minimum ping time. But this would be incorrect because lower-layer inflation, embodied in RTT, has a *multiplicative* effect on each of DNS, TCP handshake, request-response, and TCP transfer time. The total time for a page fetch (without TLS) can be broken down roughly (ignoring minor factors like the client stack) as: $T_{total} = T_{DNS} + T_{handshake} + T_{request} + T_{serverproc} + T_{response} + T_{transfer}$. If we changed the network's RTTs as a whole by a factor of x, everything on the RHS except the server processing time (which can be made quite small in practice) changes by a factor of x (to an approximation; TCP transfer time's dependence on RTTs is a bit more complex), thus changing T_{total} by approximately a factor of x as well.

What if there was no inflation in the lower layers, *i.e.*, RTTs were the same as c-latencies? For an approximate answer, we can normalize DNS, TCP handshake, request-response (excluding the server processing time, *i.e.*, only the RTT) and TCP transfer time by the minimum ping time instead of c-latency, as shown in Fig. 6(b).

The medians are 2 times (DNS), 1.02 times (TCP handshake), 4 times (TCP transfer), and 10.7 times (Total time) respectively. (Request-response is excluded because processing time at the server does not depend on the RTT.) When the 3.1 times inflation in minimum ping time is compared to these numbers, instead of the medians without such normalization, it appears much more significant. Also consider that if, for example, TCP transfer could be optimized such that it happens within an RTT, the Internet would still be more than ∼25 times slower than the c-latency in the median, but if we could cut inflation at the lower layers from 3.1 times to close to 1, even if we made no transport protocol improvements, we would get to around ∼10.7 times.

5 Related Work

There is a large body of work on reducing Internet latency. However, this work has been limited in its scope, its scale, and most crucially, its ambition. Several efforts have focused on particular pieces; for example, [23,31] focus on TCP handshakes; [11] on TCP's initial congestion window; [28] on DNS resolution; [20,14] on routing inflation due to BGP policy. Other work has discussed results from small scale experiments; for example, [26] presents performance measurements for 9 popular Web sites; [15] presents DNS and TCP measurements for the most popular 100 Web sites. The WProf [29] project profiles 350 pages and produces a break down of time spent in various browser activities. Wang et al. [30] investigate latency on mobile browsers, but focus on the compute aspects rather than networking.

The central question we have not seen answered, or even posed before, is *'Why are we so far from the speed of light?'*. Even the ramifications of a speed-of-light Internet have not been explored in any depth. The 2013 Workshop on Reducing Internet Latency [8] focused on potential mitigation techniques, with bufferbloat and active queue management being among the centerpieces. The goal of achieving latencies imperceptible to humans was also articulated [27]. Our measurements and analysis put the focus on an aspect of the latency problem that has been largely ignored so far: infrastructural inefficiencies. We hope that our work urges greater consideration for latency in efforts for expanding Internet's reach to under-served populations. However, so far, infrastructural latency has only garnered attention in niche scenarios, such as the financial markets, and isolated submarine cable projects aimed at shortening specific routes [22,21].

6 Discussion & Conclusion

Speed-of-light Internet connectivity would be a technological leap with the potential for new applications, instant response, and radical changes in the interactions between people and computing. To shed light on what's keeping us from this vision, in this work, we quantify the latency gaps introduced by the Internet's physical infrastructure and its network protocols. Our analysis suggests that the networking community should, in addition to continuing efforts for protocol improvements, also explore methods of reducing latency at the lowest layers.

Acknowledgments

Dhruv Diddi helped process the ESnet data. Data on fiber mileages from GÉANT, the high-speed pan-European research and education network, was obtained through personal communication with Xavier Martins-Rivas, DANTE. DANTE is the project coordinator and operator of GÉANT.

References

1. cURL. http://curl.haxx.se/
2. ESnet. http://www.es.net/
3. GÉANT. http://www.geant.net/
4. Google Maps API. http://goo.gl/I4ypU
5. Internet2. http://www.internet2.edu/
6. RIPE Atlas. https://atlas.ripe.net
7. Top 500 Sites in Each Country or Territory, Alexa. http://goo.gl/R8HuN6
8. Workshop on Reducing Internet Latency, 2013. http://goo.gl/kQpBCt
9. Akamai: State of the Internet, Q1 2016. https://goo.gl/XQt324
10. Brutlag, J.: Speed Matters for Google Web Search. http://goo.gl/t7qGN8 (2009)
11. Dukkipati, N., Refice, T., Cheng, Y., Chu, J., Herbert, T., Agarwal, A., Jain, A., Sutin, N.: An Argument for Increasing TCP's Initial Congestion Window. SIGCOMM CCR (2010)
12. Durairajan, R., Barford, P., Sommers, J., Willinger, W.: Intertubes: A study of the us long-haul fiber-optic infrastructure. In: ACM SIGCOMM (2015)
13. Eric Schurman (Bing) and Jake Brutlag (Google): Performance Related Changes and their User Impact. http://goo.gl/hAUENq
14. Gao, L., Wang, F.: The Extent of AS Path Inflation by Routing Policies. GLOBECOM (2002)
15. Habib, M.A., Abrams, M.: Analysis of Sources of Latency in Downloading Web Pages. WEBNET (2000)

16. Holterbach, T., Pelsser, C., Bush, R., Vanbever, L.: Quantifying interference between measurements on the RIPE Atlas platform (2015)
17. Ilya Grigorik (Google): Latency: The New Web Performance Bottleneck. http://goo.gl/djXp3
18. Liddle, J.: Amazon Found Every 100ms of Latency Cost Them 1% in Sales. http://goo.gl/BUJgV
19. Maynard-Koran, P.: Fixing the Internet for real time applications: Part II. http://goo.gl/46EiDC
20. Mühlbauer, W., Uhlig, S., Feldmann, A., Maennel, O., Quoitin, B., Fu, B.: Impact of Routing Parameters on Route Diversity and Path Inflation. Computer Networks (2010)
21. NEC: SEA-US: Global Consortium to Build Cable System Connecting Indonesia, the Philippines, and the United States. http://goo.gl/ZOV3qa
22. Nordrum, A.: Fiber optics for the far North [News]. Spectrum, IEEE (2015)
23. Radhakrishnan, S., Cheng, Y., Chu, J., Jain, A., Raghavan, B.: TCP Fast Open. CoNEXT (2011)
24. Rexford, J., Wang, J., Xiao, Z., Zhang, Y.: BGP Routing Stability of Popular Destinations. ACM SIGCOMM Workshop on Internet Measurment (2002)
25. Singla, A., Chandrasekaran, B., Godfrey, P.B., Maggs, B.: The Internet at the Speed of Light. In: HotNets. ACM (2014)
26. Sundaresan, S., Magharei, N., Feamster, N., Teixeira, R.: Measuring and Mitigating Web Performance Bottlenecks in Broadband Access Networks. IMC (2013)
27. Täht, D.: On Reducing Latencies Below the Perceptible. Workshop on Reducing Internet Latency (2013)
28. Vulimiri, A., Godfrey, P.B., Mittal, R., Sherry, J., Ratnasamy, S., Shenker, S.: Low Latency via Redundancy. CoNEXT (2013)
29. Wang, X.S., Balasubramanian, A., Krishnamurthy, A., Wetherall, D.: Demystify Page Load Performance with WProf. NSDI (2013)
30. Wang, Z.: Speeding Up Mobile Browsers without Infrastructure Support. Master's thesis, Duke University (2012)
31. Zhou, W., Li, Q., Caesar, M., Godfrey, P.B.: ASAP: A low-latency transport layer. CoNEXT (2011)