# The Internet at the Speed of Light

Ankit Singla[†], Balakrishnan Chandrasekaran[♯], P. Brighten Godfrey[†], Bruce Maggs[♯*]
[†]University of Illinois at Urbana–Champaign, [♯]Duke University, [*]Akamai
[†]{singla2, pbg}@illinois.edu, [♯]{balac, bmm}@cs.duke.edu

## ABSTRACT

For many Internet services, reducing latency improves the user experience and increases revenue for the service provider. While in principle latencies could nearly match the speed of light, we find that infrastructural inefficiencies and protocol overheads cause today's Internet to be much slower than this bound: typically by more than one, and often, by more than two orders of magnitude. Bridging this large gap would not only add value to today's Internet applications, but could also open the door to exciting new applications. Thus, we propose a grand challenge for the networking research community: a speed-of-light Internet. To inform this research agenda, we investigate the causes of latency inflation in the Internet across the network stack. We also discuss a few broad avenues for latency improvement.

## Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design; C.2.5 [**Computer-Communication Networks**]: Local and Wide-Area Networks—*Internet*

## General Terms

Measurement, Design, Performance

## 1. INTRODUCTION

Reducing latency across the Internet is of immense value — measurements and analysis by Internet giants have shown that shaving a few hundred milliseconds from the time for a transaction can translate into millions of dollars. For Amazon, a 100ms latency penalty implies a 1% sales loss [29]; for Google, an additional delay of 400ms in search responses reduces search volume by 0.74%; and for Bing, 500ms of latency decreases revenue per user by 1.2% [14, 22]. Undercutting a competitor's latency by as little as 250ms is considered a competitive advantage [8] in the industry. Even more crucially, these numbers underscore that latency is a key determinant of user experience.


*HotNets '14,* October 27–28, 2014, Los Angeles, CA, USA.

ACM 978-1-4503-3256-9/14/10 ... $15.00
http://dx.doi.org/10.1145/2670518.2673876.

While latency reductions of a few hundred milliseconds are valuable, in this work, we take the position that the networking community should pursue a much more 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. What would such a drastic reduction in Internet latency mean, and why is it worth pursuing? Beyond the obvious gains in performance and value for today's applications, such a technological leap has truly transformative potential. A speed-of-light Internet may help realize the full potential of certain applications that have so far been limited to the laboratory or have niche availability, such as telemedicine and telepresence. For some applications, such as massive multi-player online games, the size of the user community reachable within a latency bound may play an important role in user interest and adoption, and as we shall see later, linear decreases in communication latency result in super-linear growth in community size. Low latencies on the order of a few tens of milliseconds also open up the possibility of *instant response*, where users are unable to perceive any lag between requesting a page and seeing it rendered on their browsers. Such an elimination of wait time would be an important threshold in user experience. A lightning-fast Internet can also be expected to spur the development of new and creative applications. After all, even the creators of the Internet had not envisioned the myriad ways in which it is used today.

Given the promise a speed-of-light Internet holds, why is today's Internet more than an order of magnitude slower? As we show later, the fetch time for just the HTML for the landing pages of popular Websites from a set of *generally well-connected* clients is, in the median, 34 times the round-trip speed-of-light latency. In the $90^{th}$ percentile it is 169× slower. Why are we so far from the speed of light?

While our ISPs compete primarily on the basis of peak bandwidth offered, bandwidth is not the answer. Bandwidth improvements are also necessary, but bandwidth is no longer the bottleneck for a significant fraction of the population: for instance, the average US consumer clocks in at 5+ Mbps, beyond which, the effect of increasing bandwidth on page load time is small [27]. Besides, projects like Google Fiber and other fiber-to-the-home efforts by ISPs are further improving bandwidth. On the other hand, it has been noted in a variety of contexts from CPUs, to disks, to networks that 'latency lags bandwidth', and is a more difficult problem [32].

How then do we begin addressing the order-of-magnitude gap between today's Internet latencies and the speed of light?

Is speed-of-light connectivity over the Internet an unachievable fantasy? No! In fact, the high-frequency trading industry has already demonstrated its plausibility. In the quest to cut latency between the New York and Chicago stock exchanges, several iterations of this connection have been built, aimed at successively improving latency by just a few milliseconds at the expense of hundreds of millions of dollars [28]. In the mid-1980s, the round-trip latency was 14.5ms. This was cut to 13.1ms by 2010 by shortening the physical fiber route. In 2012 however, the speed of light in fiber was declared *too slow*: microwave communication cut round-trip latency to 9ms, and later down to 8.5ms [18, 11]. The $c$-latency, *i.e.*, the round-trip travel time between the same two locations along the shortest path on the Earth's surface at the speed of light in vacuum, is only 0.6ms less. A similar race is underway along multiple segments in Europe, including London-Frankfurt [5].

In this work, we propose a 'speed-of-light Internet' as a grand challenge for the networking community, and suggest a path to that vision. In §2, we discuss the potential impact of such an advance on how we use the Internet, and more broadly, on computing. In §3, we measure how latencies over today's Internet compare to $c$-latency. In §4, we break down the causes of Internet latency inflation across the network stack. We believe this to be the first attempt to directly tackle the question *'Why are we so far from the speed of light?'*. Using 20+ million measurements of 28,000 Web URLs served from 120+ countries, we study the impact of both infrastructural bottlenecks and network protocols on latency. In §5, based on our measurements and analysis, we lay out two broad approaches to cutting the large gap between today's Internet latencies and its physical limits.

## 2. THE NEED FOR SPEED

A speed-of-light Internet would be an advance with tremendous impact. It would enhance user satisfaction with Web applications, as well as voice and video communication. The gaming industry, where latencies larger than 50ms can hurt gameplay [31], would also benefit. But beyond the promise of these valuable improvements, a speed-of-light Internet could fundamentally transform the computing landscape.

**New applications.** One of computing's natural, yet unrealized goals is to create a convincing experience of joining two distant locations. Several applications — telemedicine, remote collaborative music performance, and telepresence — would benefit from such technology, but are hampered today by the lack of a low latency communication mechanism. A speed-of-light Internet could move such applications from their limited experimental scope, to ubiquity. And perhaps we will be surprised by the creative new applications that evolve in that environment.[1]

**Illusion of instant response.** A speed-of-light Internet can realize the possibility of *instant response*. The limits of human perception imply that we find it difficult to correctly order visual events separated by less than 30ms [7]. Thus, if responses over the Internet were received within 30ms of the requests, we would achieve the illusion of instant response[2]. A (perceived) zero wait-time for Internet services would greatly improve user experience and allow for richer interaction. Immense resources, both computational and human, would become "instantly" available over a speed-of-light Internet.

**Super-linear community size.** Many applications require that the connected users be reachable within a certain latency threshold, such as 30ms round-trip for instant response, or perhaps 50ms for a massive multi-player online game. The value of low latency is magnified by the fact that *the size of the available user community is a superlinear function of network speed*. The area on the Earth's surface reachable within a given latency grows nearly[3] quadratically in latency. Using population density data[4] reveals somewhat slower, but still super-linear growth. We measured the number of people within a 30ms RTT from 200 capital cities of the world at various communication speeds. Fig. 1(a) shows the median (across cities) of the population reached. If Internet latencies were 20× worse than $c$-latency ($x$-axis=0.05$c$), we could reach 7.5 million people "instantly". A **10**× latency improvement ($x$-axis=0.5$c$) would increase that community size by **49**×, to 366 million. Therefore, the value of latency improvement is magnified, perhaps pushing some applications to reach critical mass.

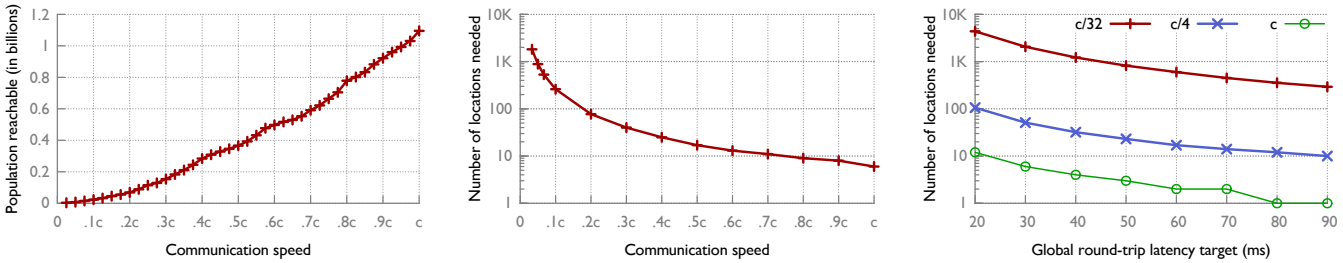
**Cloud computing and thin clients.** Another potential effect of a speedier Internet is further centralization of compute resources. Google and VMware are already jointly working towards the thin client model through virtualization [23]. Currently, their Desktop-as-a-Service offering is targeted at businesses, with the customer centralizing most compute and data in a cluster, and deploying cheaper hardware as workstations. A major difficulty with extending this model to personal computing today is the much larger latency involved in reaching home users. Likewise, in the mobile space, there is interest in offloading some compute to the cloud, thereby exploiting data and computational resources unavailable on user devices [19]. As prior work [25] has argued, however, to achieve highly responsive performance from such applications would today require the presence of a large number of data center facilities. With a speedier Internet, the 'thin client' model becomes plausible for both desktop and mobile computing with far fewer installations. For instance, if the Internet operated at half the speed of light, almost all of

[1] "New capabilities emerge just by virtue of having smart people with access to state-of-the-art technology." — Bob Kahn

[2] This is a convenient benchmark number, but the exact number will vary depending on the scenario. For a 30ms response time, the Internet will actually need to be a little faster because of server-side request processing time, screen refresh delay, etc. And the 'instant response' threshold will differ for audio vs. visual applications.

[3] Because it is a sphere, not a plane.

[4] Throughout, we use population estimates for 2010 [15].




(a) (b) (c)

**Figure 1:** ***The impact of communication speed on computing and people. With increasing communication speed: (a) the population within 30ms round-trip time grows super-linearly; (b) the number of locations (e.g. data centers or CDN nodes) needed for global 30ms reachability from at least one location falls super-linearly; and (c) the tradeoff between the global latency target and the number of locations required to meet it improves.***

the contiguous US could be served instantly from just one location. Fig. 1(b) shows the number of locations needed for 99% of the world's population to be able to instantly reach at least one location — as we decrease Internet latency, the number of facilities required falls drastically, down to only 6 locations with global speed-of-light connectivity. (These numbers were estimated using a heuristic placement algorithm and could possibly be improved upon.) This result is closely related to that in Fig. 1(a) — with increasing communication speed (which, given a latency bound, determines a reachable radius), the population reachable from a center grows super-linearly, and the number of centers needed to cover the entire population falls super-linearly.

**Better geolocation.** As latency gets closer to the speed of light, latency-based geolocation gets better, and in the extreme case of exact $c$-latency, location can be precisely triangulated. While better geolocation provides benefits such as better targeting of services and matching with nearby servers, it also has other implications, such as for privacy.

**Don't CDNs solve the latency problem?** Content distribution networks cut latency by placing a large number of replicas of content across the globe, so that for most customers, some replica is nearby. However, this approach has its limitations. First, some resources simply cannot be replicated or moved, such as people. Second, CDNs today are an expensive option, available only to larger Internet companies. A speedier Internet would significantly cut costs for CDNs as well, and in a sense, democratize the Internet. CDNs make a tradeoff between costs (determined, in part, by the number of infrastructure locations), and latency targets. For any latency target a CDN desires to achieve globally, given the Internet's communication latency, a certain minimum number of locations are required. Speeding up the Internet improves this entire tradeoff curve. This improvement is shown in Fig. 1(c), where we estimate (using our random placement heuristic) the number of locations required to achieve different latency targets for different Internet communication speeds[5]: $\frac{c}{32}$, $\frac{c}{4}$, and $c$. As is clear from these results, while

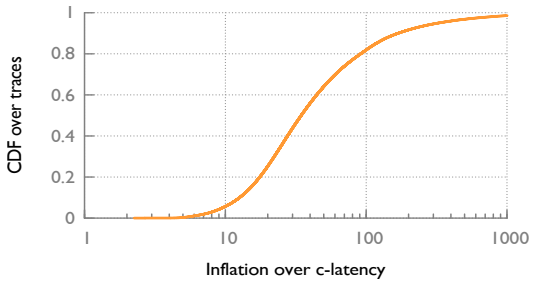



**Figure 2:** ***Fetch time of just the HTML of the landing pages of popular Websites in terms of inflation over the speed of light. In the median, fetch time is 34× slower.***

CDNs will still be necessary to hit global latency targets of a few tens of milliseconds, the amount of infrastructure they require to do so will fall drastically with a speedier Internet.

## 3. THE INTERNET IS TOO SLOW

We fetched just the HTML for landing pages of 28,000 popular Websites[6] from 400+ PlanetLab nodes using cURL [1]. For each connection, we geolocated the Web server using commercial geolocation services, and 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[7]. Henceforth, we refer to the ratio of the fetch time to $c$-latency as the Internet's latency inflation. Fig. 2 shows the CDF of this inflation over 6 million connections. The time to finish HTML retrieval is, in the median, 34× the $c$-latency, while the $90^{th}$ percentile is 169×. Thus, the Internet is typically more than an order of magnitude slower than the speed of light. We

[5]Per our measurements in §3, $\frac{c}{32}$ is close to the median speed of fetching just the HTML for the landing pages of popular websites today, and $\frac{c}{4}$ is close to the median ping speed.

[6]We pooled Alexa's [9] top 500 Websites from each of 120+ countries and used the unique URLs. We followed redirects on each URL, and recorded the final URL for use in experiments. In our experiments, we ignored any URLs that still caused redirects. We excluded data for the few hundred websites using SSL. We did find, as expected, that SSL incurred several RTTs of additional latency.

[7]We have ground-truth geolocation for PlanetLab nodes — while the PlanetLab API yields incorrect locations for some nodes, these are easy to identify and remove based on simple latency tests.

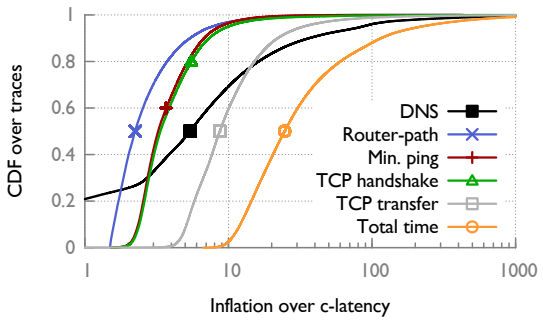



**Figure 3:** ***Various components of latency inflation. One point is marked on each curve for sake of clarity.***

note that PlanetLab nodes are generally well-connected, and latency can be expected to be poorer from the network's *true* edge.

## 4. WHY IS THE INTERNET SO SLOW?

To answer this question, we attempt to break down the fetch time across layers, from inflation in the physical path followed by packets to the TCP transfer time. We use cURL to obtain the time for DNS resolution, TCP handshake, TCP data transfer, and total fetch time for each connection. 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, and connect successive routers with the shortest paths on the Earth's surface as an approximation for the route the packets follow. We compute the roundtrip latency at the speed of light in fiber along this approximate path, and refer to it as the 'router-path latency'. We normalize each latency component by the $c$-latency between the respective connection's end-points.

We limit this analysis to roughly one million connections, for which we used cURL to fetch the first 32KB (22 full-sized packets) of data from the Web server[8]. The results are shown in Fig. 3. It is unsurprising that DNS resolutions are faster than $c$-latency about $20\%$ of the time — in these cases, the server happens to be farther than the DNS resolver. (The DNS curve is clipped at the left to more clearly display the other results.) In the median, DNS resolutions are $5.4\times$ inflated over $c$-latency, with a much longer tail. In fact, we found that when we consider the top and bottom $10$ percentiles of total fetch time inflation, DNS plays a significant role – among the fastest $10\%$ of pages, even the worst DNS inflation is less than $3\times$, while for the slowest $10\%$ of pages, even the median DNS time is worse than $20\times$ inflated.

Fig. 3 also reveals the significant inflation in TCP transfer time — $8.7\times$ in the median. Most of this is simply TCP's slow start mechanism at work — with only 32KB being fetched, bandwidth is not the bottleneck here. The TCP handshake (counting only the SYN and SYN-ACK) is $3.2\times$ worse than $c$-latency in the median, roughly the same as the round trip time (minimum ping latency).

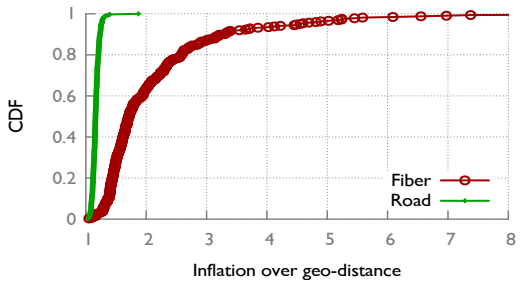
Note that the medians of inflation in DNS, TCP handshake, and TCP transfer time do not add up to the median inflation in total time. This is because of the long tails of the inflations in each of these.

Having analyzed the somewhat easier to examine TCP and DNS factors, we devote the rest of this section to a closer look at inflation in the lower layers: physical infrastructure, routing, and queuing and bufferbloat.

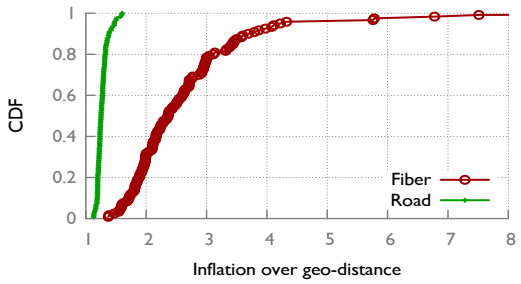
### 4.1 Physical infrastructure and routing

Fig. 3 shows that in the median, the router-path is only $2.3\times$ inflated. (The long tail is, in part, explained by 'hair-pinning', *i.e.*, packets between nearby end-points traversing circuitous routes across the globe. For instance, in some cases, packets between end-points in Eastern China and Taiwan were seen in our traces traveling first to California.) 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 air / vacuum. Excluding this inflation from the median leaves a further inflation of $1.53\times$. While this may appear small, as we discuss below, our estimate is optimistic, and overall, inflation in these lower layers plays a significant role.

We see some separation between the minimum ping time and the router-path latency. This gap may be explained by two factors: (a) traceroute often does not yield responses from all the routers on the path, in which case we essentially see artificially shorter paths — our computation simply as-




(a)




(b)

**Figure 4:** ***Compared to the shortest distance along the Earth's surface, there is significantly more inflation in fiber lengths than in road distances in both (a) Internet2 connections; and (b) GÉANT connections.***

[8]cURL allows explicit specification of the number of bytes to fetch, but some servers do not honor such a request. Measurements from connections that did not fetch roughly 32KB were discarded.

sumes that there is a direct connection between each pair of successive replying routers; 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 two research networks: Internet2 [4] and GÉANT[9]. We obtained point-to-point fiber lengths for these networks and ran an all pairs shortest paths computation on the network maps to calculate fiber lengths between all pairs of end points. We also calculated the shortest distance along the Earth's surface between each pair, and obtained the road distances using the Google Maps API [3]. Fig. 4 shows the inflation in fiber lengths and road distances compared to the shortest distance. Road distances are close to shortest distances, while fiber lengths are significantly larger and have a long tail. Even when only point-to-point connections are considered, fiber lengths are usually 1.5-2× larger than road distances.

While it is tempting to dismiss the 3.2× inflation in the median ping time in light of the larger inflation factors in DNS (5.4×) and TCP transfer (8.7×), each of DNS, TCP handshake, and TCP transfer time suffers due to inflation in the physical and network layers. What if there was no inflation in the lower layers? For an approximate answer, we can normalize inflation in DNS, TCP handshake, and TCP transfer time to that in the minimum ping time. Normalized by the median inflation in ping time (3.2×), the medians are 1.7, 1.0, and 2.7 respectively. Thus, inflation at the lower layers itself plays a big role in Internet latency inflation.

## 4.2 Loss, queuing, and bufferbloat

Fig. 3 shows that the TCP handshake time (time between cURL's sending the SYN and receiving the SYN-ACK) is nearly the same as the minimum ping latency, indicating, perhaps, a lack of significant queuing effects. Nevertheless, 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 propagation 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 in small: the $2^{nd}$-longest ping time is only 1.2% larger than the minimum ping time in the median. However, because pings (using ICMP) might use queues separate from Web traffic, we also used tcpdump [6] at the client to log packet arrival times from the server, and analyzed the inter-arrival gaps between packets. We limited this analysis to the same roughly one million connections as before. More than 95% of these connections experienced no packet loss (estimated as packets re-ordered by more than 3ms).

[9]Data on fiber mileages from GÉANT[2], 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.

Under normal TCP operation, at this data transfer size, most packets can be expected to arrive with sub-millisecond inter-arrival times, an estimated ∼13% of packets with a gap of one RTT (as the sender waits for ACKs between windows). Only ∼5% of all inter-arrival gaps did not fall into either of those two categories. Further, for more than 80% of all connections, the largest gap was close to one RTT. Based on these numbers, for most connections, we can rule out the possibility of a single large gap, as well as that of multiple smaller gaps additively causing a large delay. We can safely conclude that for most of these connections, bufferbloat cannot explain the large latency inflation observed.

We use the above results from PlanetLab measurements only to stress that even in scenarios where bufferbloat is clearly not the dominant cause of additional latency, significant other problems inflate Internet latencies by more than an order of magnitude. Further, for a peek at bufferbloat in end-user environments, we also examined RTTs in a sample of TCP connection handshakes between Akamai's servers and clients (end-users) over a 24-hour time period, passively logged by Akamai servers. (A large fraction of routes to popular prefixes are unlikely to change at this time-scale in the Internet [35]. The connections under consideration here are physically much shorter, making route changes even more unlikely.) We analyzed all server-client pairs that appeared more than once in our data: ∼10 million pairs, of which 90% had 2 to 5 observations. We computed the inflation over $c$-latency of the minimum (Min), average (Avg) and maximum (Max) of the set of RTTs observed between each pair; for calculating the inflations we had ground truth on the location of the servers, and the clients were geolocated using data from Akamai EdgeScape [13].

Fig. 5 compares the CDFs of inflation in the Min, Avg and Max of the RTTs. In the median, the Avg RTT is 1.9× the Min RTT (*i.e.*, in absolute terms, Avg is 30ms larger than Min). Bufferbloat is certainly a suspect for this difference, although server response times may also play a role. Note however, that in our PlanetLab measurements, where bufferbloat does not play a central role, we observed (in the median) a ping latency of 124ms. If we added an additional 30ms of "edge inflation", it would comprise less than 20%

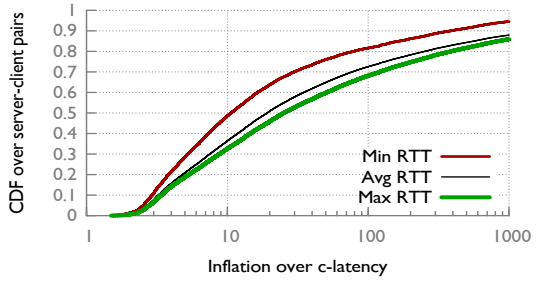



**Figure 5:** ***Latency inflation in RTTs between end users and Akamai servers, and the variation therein. The difference between the minimum and average RTTs could possibly be attributed to bufferbloat.***

of the total inflation in the ping latency, which itself is a fraction of the Internet's latency inflation. Thus, to summarize, loss, queuing, and bufferbloat do not explain most of the large latency inflation in the Internet.

## 5. FAST-FORWARD TO THE FUTURE

In line with the community's understanding, our measurements affirm that TCP transfer and DNS resolution are important factors causing latency inflation. However, inflation at lower layers is equally, if not more important. Thus, below, we lay out two broad ideas for drastically cutting Internet latencies targeting each of these problems.

**A parallel low-latency infrastructure:** Most flows on the Internet are small in size, with most of the bytes being carried in a small fraction of flows [41]. Thus, it is conceivable that we could improve latency for the large fraction of small-sized flows by building a separate low-latency low-bandwidth infrastructure to support them. Such a network could connect major cities along the shortest paths on the Earth's surface (at least within the continents) using a $c$-speed medium, such as either microwave or potentially hollow fiber [20]. Such a vision may not be far-fetched on the time horizon of a decade or two.

As Fig. 4 shows, the road network today is much closer to shortest paths than the fiber network. Road construction is two orders of magnitude costlier per mile than fiber [16, 33]. Further, the additional cost of laying fiber along new roads or roads that are being repaved is even smaller. As the road infrastructure is repaired and expanded over decades, it seems feasible to include fiber outlay in such projects. In fact, along these lines, legislation recently proposed in the United States Congress would make it mandatory to install fiber conduits as part of any future Federal highway projects [17].

**Latency optimizations by ISPs:** ISPs, by virtue of observing real-time traffic, are in perhaps the best position to make latency optimizations for clients. For instance, an ISP can keep track of the TCP window sizes achieved by flows on a per-prefix basis. It can then direct clients to use these window sizes, thereby reducing the order-of-magnitude slowdown due to TCP transfer time that we see in Fig. 3. Likewise, ISPs can maintain pools of TCP connections to popular web services and splice these on to clients that seek to connect to the services, eliminating the TCP handshake time. A similar optimization is already being used by CDNs — Akamai maintains persistent TCP connections between its own servers as well as from its servers to content providers, and clients only connect to a nearby Akamai server, which may then patch the connection to a distant location [12]. ISPs can also make predictive optimizations. For instance, an ISP may observe that any client that requests a certain Webpage then requests name resolution for certain other domains, or the fetching of certain resources. The ISP can then proactively resolve such names or fetch such resources for the client.

We also observed in §4 that the tail DNS resolution time plays a significant role. Recent work by Vulimiri et al. [38] illustrates a simple and effective method of substantially cutting this tail time – redundancy in queries. This optimization can be deployed either by ISPs, making redundant queries on behalf of clients, or by the clients themselves.

## 6. 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, [34, 42] focus on TCP handshakes; [21] on TCP's initial congestion window; [38] on DNS resolution; [30, 24] on routing inflation due to BGP policy. Other work has discussed results from small scale experiments; for example, [36] presents performance measurements for 9 popular Websites; [26] presents DNS and TCP measurements for the most popular 100 Websites. The WProf [39] project breaks down Webpage load time for 350 Webpages into computational aspects of page rendering, as well as DNS and TCP handshake times. Wang et al. [40] 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 — how would such an advance change computing and its role in our lives? Answering these questions, and thereby helping to set the agenda for networking research in this direction is our work's primary objective.

The 2013 Workshop on Reducing Internet Latency [10] focused on potential mitigation techniques, with bufferbloat and active queue management being among the centerpieces. One interesting outcome of the workshop was a qualitative chart of latency reduction techniques, and their potential impact and feasibility (Fig. 1 in [10]). In a similar vein, one objective of our work is to *quantify* the latency gaps, separating out factors which are fundamental (like the $c$-bound) from those we might hope to improve. The goal of achieving latencies imperceptible to humans was also articulated [37]. We share that vision, and in §2 discuss the possible impacts of that technological leap.

## 7. CONCLUSION

Speed-of-light Internet connectivity would be a technological leap with phenomenal consequences, including 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, we have attempted to quantify the latency gaps introduced by the Internet's physical infrastructure and its network protocols, finding that infrastructural gaps are as significant, if not more than protocol overheads. We hope that these measurements will form the first steps in the networking community's methodical progress towards addressing this grand challenge.

## 8. REFERENCES


[1] cURL. http://curl.haxx.se/.

[2] GÉANT. http://www.geant.net/.

[3] Google Maps API. http://goo.gl/I4ypU.

[4] Internet2. http://www.internet2.edu/.

[5] Quincy Extreme Data service. http://goo.gl/wSRzjX.

[6] tcpdump. http://www.tcpdump.org/.

[7] Temporal Consciousness, Stanford Encyclopedia of Philosophy. http://goo.gl/UKQwy7.

[8] The New York Times quoting Microsoft's "Speed Specialist", Harry Shum. http://goo.gl/G5Ls0O.

[9] Top 500 Sites in Each Country or Territory, Alexa. http://goo.gl/R8HuN6.

[10] Workshop on Reducing Internet Latency, 2013. http://goo.gl/kQpBCt.

[11] J. Adler. Raging Bulls: How Wall Street Got Addicted to Light-Speed Trading. http://goo.gl/Y9kXeS.

[12] Akamai. Accelerating Dynamic Content with Akamai SureRoute. http://goo.gl/bUh1s7.

[13] Akamai. EdgeScape. http://goo.gl/qCHPh1.

[14] J. Brutlag. Speed Matters for Google Web Search. http://goo.gl/t7qGN8, 2009.

[15] Center for International Earth Science Information Network (CIESIN), Columbia University; United Nations Food and Agriculture Programme (FAO); and Centro Internacional de Agricultura Tropical (CIAT). Gridded Population of the World: Future Estimates (GPWFE). http://sedac.ciesin.columbia.edu/gpw, 2005. Accessed: 2014-01-12.

[16] Columbia Telecommunications Corporation. Brief Engineering Assessment: Cost Estimate for Building Fiber Optics to Key Anchor Institutions. http://goo.gl/ESqVPW.

[17] Congressional Bills 112th Congress. Broadband Conduit Deployment Act of 2011. http://goo.gl/9kLQ4X.

[18] C. Cookson. Time is Money When it Comes to Microwaves. http://goo.gl/PspDwl.

[19] E. Cuervo. *Enhancing Mobile Devices through Code Offload*. PhD thesis, Duke University, 2012.

[20] DARPA. Novel Hollow-Core Optical Fiber to Enable High-Power Military Sensors. http://goo.gl/GPdb0g.

[21] N. Dukkipati, T. Refice, Y. Cheng, J. Chu, T. Herbert, A. Agarwal, A. Jain, and N. Sutin. An Argument for Increasing TCP's Initial Congestion Window. *SIGCOMM CCR*, 2010.

[22] Eric Schurman (Bing) and Jake Brutlag (Google). Performance Related Changes and their User Impact. http://goo.gl/hAUENq.

[23] Erik Frieberg, VMWare. Google and VMware Double Down on Desktop as a Service. http://goo.gl/5quMU7.

[24] L. Gao and F. Wang. The Extent of AS Path Inflation by Routing Policies. *GLOBECOM*, 2002.

[25] K. Ha, P. Pillai, G. Lewis, S. Simanta, S. Clinch, N. Davies, and M. Satyanarayanan. The Impact of Mobile Multimedia Applications on Data Center Consolidation. *IC2E*, 2013.

[26] M. A. Habib and M. Abrams. Analysis of Sources of Latency in Downloading Web Pages. *WEBNET*, 2000.

[27] Ilya Grigorik (Google). Latency: The New Web Performance Bottleneck. http://goo.gl/djXp3.

[28] G. Laughlin, A. Aguirre, and J. Grundfest. Information Transmission Between Financial Markets in Chicago and New York. *arXiv:1302.5966v1*, 2013.

[29] J. Liddle. Amazon Found Every 100ms of Latency Cost Them 1% in Sales. http://goo.gl/BUJgV.

[30] W. Mühlbauer, S. Uhlig, A. Feldmann, O. Maennel, B. Quoitin, and B. Fu. Impact of Routing Parameters on Route Diversity and Path Inflation. *Computer Networks*, 2010.

[31] L. Pantel and L. C. Wolf. On the Impact of Delay on Real-Time Multiplayer Games. *NOSSDAV*, 2002.

[32] D. A. Patterson. Latency Lags Bandwidth. *Communications of the ACM*, 2004.

[33] Planning & Markets, University of Southern California. Highway Construction Costs under Government Provision. http://goo.gl/pJHFSB.

[34] S. Radhakrishnan, Y. Cheng, J. Chu, A. Jain, and B. Raghavan. TCP Fast Open. *CoNEXT*, 2011.

[35] J. Rexford, J. Wang, Z. Xiao, and Y. Zhang. BGP Routing Stability of Popular Destinations. *ACM SIGCOMM Workshop on Internet Measurment*, 2002.

[36] S. Sundaresan, N. Magharei, N. Feamster, and R. Teixeira. Measuring and Mitigating Web Performance Bottlenecks in Broadband Access Networks. *IMC*, 2013.

[37] D. Täht. On Reducing Latencies Below the Perceptible. *Workshop on Reducing Internet Latency*, 2013.

[38] A. Vulimiri, P. B. Godfrey, R. Mittal, J. Sherry, S. Ratnasamy, and S. Shenker. Low Latency via Redundancy. *CoNEXT*, 2013.

[39] X. S. Wang, A. Balasubramanian, A. Krishnamurthy, and D. Wetherall. Demystify Page Load Performance with WProf. *NSDI*, 2013.

[40] Z. Wang. Speeding Up Mobile Browsers without Infrastructure Support. Master's thesis, Duke University, 2012.

[41] Y. Zhang, L. Breslau, V. Paxson, and S. Shenker. On the Characteristics and Origins of Internet Flow Rates. *ACM CCR*, 2002.

[42] W. Zhou, Q. Li, M. Caesar, and P. B. Godfrey. ASAP: A Low-Latency Transport Layer. *CoNEXT*, 2011.