cISP: A Speed-of-Light Internet Service Provider

Debopam Bhattacherjee, *ETH Zürich;* Waqar Aqeel, *Duke University;* Sangeetha Abdu Jyothi, *UC Irvine and VMware Research;* Ilker Nadi Bozkurt, *Duke University;* William Sentosa, *UIUC;* Muhammad Tirmazi, *Harvard University;* Anthony Aguirre, *UC Santa Cruz;* Balakrishnan Chandrasekaran, *VU Amsterdam;* P. Brighten Godfrey, *UIUC and VMware;* Gregory Laughlin, *Yale University;* Bruce Maggs, *Duke University and Emerald Technologies;* Ankit Singla, *ETH Zürich*

https://www.usenix.org/conference/nsdi22/presentation/bhattacherjee

This paper is included in the Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation.

April 4–6, 2022 • Renton, WA, USA

978-1-939133-27-4

Open access to the Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation is sponsored by

cISP: A Speed-of-Light Internet Service Provider

Debopam Bhattacherjee* [1], Waqar Aqeel* [2], Sangeetha Abdu Jyothi[3,4], Ilker Nadi Bozkurt[2†], William Sentosa[5], Muhammad Tirmazi[6], Anthony Aguirre[7], Balakrishnan Chandrasekaran[8], P. Brighten Godfrey[5,9], Gregory Laughlin[10], Bruce Maggs[2,11], Ankit Singla[1]

[1]ETH Zürich, [2]Duke University, [3]UC Irvine, [4]VMware Research, [5]UIUC, [6]Harvard University, [7]UC Santa Cruz, [8]VU Amsterdam, [9]VMware, [10]Yale University, [11]Emerald Technologies

Abstract

Low latency is a requirement for a variety of interactive network applications. The Internet, however, is not optimized for latency. We thus explore the design of wide-area networks that move data at nearly the speed of light in vacuum. Our **cISP** design augments the Internet's fiber with free-space microwave wireless connectivity over paths very close to great-circle paths. cISP addresses the fundamental challenge of simultaneously providing ultra-low latency while accounting for numerous practical factors ranging from transmission tower availability to packet queuing. We show that instantiations of cISP across the United States and Europe would achieve mean latencies within 5% of that achievable using great-circle paths at the speed of light, over medium and long distances. Further, using experiments conducted on a nearly-speed-of-light algorithmic trading network, together with an analysis of trading data at its end points, we show that microwave networks are *reliably* faster than fiber networks even in inclement weather. Finally, we estimate that the economic value of such networks would substantially exceed their expense.

1 Introduction

User experience in many interactive network applications depends crucially on achieving low latency. Even seemingly small increases in latency can negatively impact user experience, and, subsequently, revenue for service providers: Google, for example, quantified the impact of an additional 400 ms of latency in search results as 0.7% fewer searches per user [18]. Further, wide-area latency is often the bottleneck, as Facebook's analysis of over a million requests found [21]. Indeed, content delivery networks (CDNs) present latency reduction and its associated increase in conversion rates as one of the key value propositions of their services, citing, e.g., a 1% loss in sales per 100 ms of latency for Amazon [2]. In spite of the significant impact of latency on performance and user experience, the Internet is not designed to treat low latency as a primary objective. This is the problem we address: reducing latencies over the Internet to the lowest possible.

The best achievable latency between two points along the surface of the Earth is determined by their geodesic distance divided by the speed of light, c. Latencies over the Internet, however, are usually much larger than this minimal "c-latency": recent measurement work found that fetching even small amounts of data over the Internet typically takes 37× longer than the c-latency, and often, more than 100× longer [16]. This delay comes from the many round-trips between the communicating endpoints, due to inefficiencies in the transport and application layer protocols, and from each round-trip itself taking 3-4× longer than the c-latency [16]. Given the approximately multiplicative role of network round-trip times (RTTs) when bandwidth is not the main bottleneck, eliminating inflation in Internet RTTs can potentially translate to up to 3-4× speedup, even *without* any protocol changes. Further, as protocol stack improvements get closer to their ideal efficiency of one RTT for small amounts of data, the RTT becomes the singular network bottleneck. Similarly, for well-designed applications dependent on persistent connectivity between two fixed locations, such as gaming, *nothing* other than resolving this 3-4× "infrastructural inefficiency" can improve latency substantially.

Thus, beyond the networking research community's focus on protocol efficiency, reducing the Internet infrastructure's latency inflation is the next frontier in research on latency. While academic research has typically treated infrastructural latency inflation as an unresolvable given, we argue that this is a high-value opportunity, and is much more tractable than may be evident at first.

What are the root causes of the Internet's infrastructural inefficiency, and how do we ameliorate them? Large latencies are partly explained by poor use of existing fiber infrastructure: two communicating sites often use a longer, indirect route because their service providers do not peer over the shortest fiber connectivity between their locations. We find, nevertheless, that even latency-optimal use of *all* known fiber conduits, computed via shortest paths in the InterTubes dataset [34], would leave us 1.98× away from c-latency [17]. This gap stems from the speed of light in fiber being $\sim\frac{2}{3}c$,

* Equal contribution. † Now at Google.

and the unavoidable circuitousness of fiber routes due to topographic and economic constraints of buried conduits.

We thus explore the design of **cISP**, an Internet Service Provider that provides nearly speed-of-light latency by exploiting wireless electromagnetic transmissions, which can be realized with point-to-point microwave antennas mounted on towers. This approach holds promise for overcoming both the aforementioned shortcomings fundamental to today's fiber-based networks: the transmission speed in air is essentially equal to c, and the richness of existing tower infrastructure makes more direct paths possible. Nevertheless, it also presents several new challenges, including:

- overcoming numerous practical constraints, such as tower availability, line-of-sight requirements, and the impact of weather on performance;
- coping with limited wireless bandwidth;
- solving a large-scale cost-optimal network design problem, which is NP-hard; and
- addressing switching and queuing delays, which are more prominent with the smaller propagation delays.

To meet these challenges, we propose a hybrid design that augments the Internet's fiber connectivity with nearly straight-line wireless links. These low-latency links are used judiciously where they provide the maximum latency benefit, and only for the high-impact but small proportion, in terms of bytes, of Internet traffic that is latency-sensitive. We design a simple heuristic that achieves near-optimal results for the network design problem. Our approach is flexible and enables network design for a variety of deployment scenarios; in particular, we show that cISP's design for interconnecting large population centers in the contiguous U.S. and Europe can achieve mean latencies as low as 1.05× c-latency at a cost of under $1 per gigabyte (GB). We show through simulation that such networks can be operated at high utilization without excessive queuing.

To address the practical concerns, we use fine-grained geographic data and the relevant physical constraints to determine where the needed wireless connectivity would be feasible to deploy, and assess our design under a variety of scenarios with respect to budget, tower height and availability, antenna range, and traffic matrices. We also use a year's worth of meteorological data to assess the network's performance during weather disturbances, showing that most of cISP's latency benefits remain intact throughout the year. Our weather simulation and an animation showing how the hybrid network evolves from mostly-fiber to mostly-wireless with increasing budget are available online; see [25] and [26].

But is it feasible to use microwave hardware for low latency in practice? To answer this question, we rented virtual machines in the CME data center in Chicago and the Equinix data center in New Jersey, and, on Saturdays, were given access at these data centers to one of the fastest microwave networks spanning the Chicago – New Jersey algorithmic trading corridor. Experiments conducted on this network show that it successfully operates at a speed extremely close to the speed of light, and that losses can be effectively handled by extremely lightweight forward error correction (FEC). We complement these findings by analyzing real trading data, revealing the minimum latency between the data centers and showing that the network is available in varied weather conditions.

Finally, we explore the application-level benefits for Web browsing and gaming, and present estimates showing that the utility of cISP vastly exceeds its cost, even for web sites already using CDNs to reduce latency.

2 Technology Background

At the highest level, our approach involves using free-space communication between transmitters mounted at a suitable height, e.g., using dedicated towers or existing buildings, and separated from each other by at most a certain limiting distance. Network links longer than this range require a series of such transmitters. Typically, even after accounting for terrain, such a network link can be built close to the shortest path on the Earth's surface between the two end points. Further, the speed of light in air is essentially the same as that in vacuum, c. These properties make our approach attractive for the design of (nearly) c-latency networks.

Technology choices. Several physical layer technologies are amenable for use in our design, including free-space optics (FSO), microwave (MW), and millimeter wave (MMW). At present, we believe MW provides the best combination of range, resilience, throughput, and cost. Future advances in any of these technologies, however, can be easily rolled into our design, and can only improve our cost-benefit analysis.

While hollow fiber [31] could, in the future, also provide c-latency, it would still suffer from the circuitousness of today's fiber conduits. Low Earth orbit satellite networks, as are being currently deployed, could also help, although they currently incur substantially higher latency than cISP (§9).

Switching latency. While long-haul MW networks have existed since the 1940s [10], their use in high-frequency trading starting within the last 10 years [55] has driven innovation in radios so that each MW retransmission only takes a few μs. Thus, even wide-area links with many retransmissions incur negligible switching latency. As an example, the HFT industry operates a MW relay between Chicago and New Jersey comprising ≈ 20 line-of-sight links that operates within 1% of c-latency end-to-end at the *application* layer [58].

Packet loss. Loss occurs for several reasons, including weather disruption and intermittent multi-path fading, especially over bodies of water. In §5.1, using a year's worth of weather data, we analyze the impact of diverting traffic to alternate (fiber or MW) routes during inclement weather. Our active experiments on a microwave network also show that losses experienced could be handled with lightweight forward error correction (FEC).

Spectrum and licensing. We propose the use of MW com-

munication in the 6-18 GHz frequency range. These frequencies are not very crowded, and licensing is generally not very competitive, except at 6 GHz in cities, and along certain routes, like the above mentioned HFT corridor. The licenses are given on a first-come, first-served basis, recorded in a public database, they protect against the deployment of other links that would interfere with licensed links.

Line-of-sight & range. Successive MW towers need line-of-sight visibility, accounting for the Earth's curvature, terrain, trees, buildings and other obstructions, and atmospheric refraction. Attenuation also limits range. A maximum range of around 100 km is practicable, but we show results with maximum allowed range varying between 60-100 km (§5.2).

Bandwidth. Between any two towers, using very efficient encoding (256 QAM or higher), wide frequency channels, and radio multiplexing, a data rate of about 1 Gbps is achievable [45]. This bandwidth is vastly smaller than for fiber, and necessitates a hybrid design using fiber and MW.

Geographic coverage. Connecting individual homes directly to such a MW network would be cost-prohibitive. To maximize cost-efficiency, we focus on long-haul connectivity, with the last mile being traditional fiber. At short distances, fiber's circuitousness and refraction are small overheads.

Cost model. We rely on cost estimates in recent work [55] and based on our conversations with industry participants involved in equipment manufacturing and link provisioning. The cost of installing a bidirectional MW link, on existing towers, is approximately \$75K (\$150K) for 500 Mbps (1 Gbps) bandwidth. The average cost for building a new tower is \$100K, with wide variation by terrain and across cities and rural areas. Any additional towers needed to augment bandwidth for particular links incur this "new tower" cost. The operational costs comprise several elements, including management and personnel, but the dominant operational expense, by far, is tower rent: $\$25-50$K per year per tower. We estimate cost per GB by amortizing the sum of building costs and operational costs over 5 years.

Note that the deployment and operational costs can vary substantially based on the deployment model. For example, imagine that a company like American Tower [7], which has a substantial tower presence across the US (see Fig. 14 in Appendix D), deploys cISP. In such a scenario, not only would the cost of bandwidth augmentation be negligible, but also the cost of maintaining the towers would be drastically reduced. We consider both conservative and optimistic deployment models and conduct an in-depth cost-analysis in this work.

3 cISP Design

At an abstract level, given the tower and fiber infrastructure, a set of n sites (e.g., cities, data centers) to interconnect, and a traffic model between them, we want to select a set of tower-level connections that minimizes network-wide latency while adhering to a budget and the constraints outlined in §2. Our approach comprises the following three broad steps.

1. Identifying a set of links that are likely to be useful by determining, for each pair of sites (s, d), the best feasible tower-level connectivity, if s and d were to be directly connected by a series of towers.
2. Building all $O(n^2)$ direct links, connecting each site to every other, would be prohibitively expensive. Thus, a subset of site-to-site links, together with existing fiber conduits, form our network. Choosing the appropriate subset is the key algorithmic problem.
3. Provisioning capacity beyond 1 Gbps along any link involves building additional tower-level links, e.g., by identifying and using links that are also nearly shortest paths, but were omitted in step 1 above.

Step 1: feasible hops. We first use line-of-sight and range constraints to decide which tower pairs can be connected. Achievable tower-to-tower hop length is limited primarily by the Earth's curvature, which can be treated as a "bulge" of height h_{Earth}. MW hops must clear this curvature and any obstructions in an ellipsoidal region between the sender and the receiver antennae known as the *Fresnel zone*, which has width h_{Fres}. At the midpoint of a hop of length D, using a MW frequency f, we have the following.

$$h_{\text{Fres}} \simeq 8.7\,m \left(\frac{D}{1\,\text{km}}\right)^{1/2} \left(\frac{f}{1\,\text{GHz}}\right)^{-1/2} \quad (1)$$

$$h_{\text{Earth}} \simeq \frac{1\ m}{50\ K} \left(\frac{D}{1\,\text{km}}\right)^{2} \quad (2)$$

In Eq. 2, K accounts for atmospheric refraction [62]. Towers should clear the sum of these heights and any other obstructions. In favorable weather, and with adequately large dish antennae, ranges of up to $D \approx 100$ km are achievable with high availability, provided such line-of-sight clearance [79]. As a specific example, the FCC licensing database [28] indicates that McKay Brothers, LLC (a financial industry provider) operated a $D = 96$ km hop from Chicago, IL (lat. 41.88°, lon. -87.62°) to Galien, MI (lat. 41.81°, lon. -86.47°) as part of a 1183 km MW relay. This example shows that multipath interference issues (associated in this case with a traversal over Lake Michigan) are not an impediment to hop viability.

We assess hop feasibility between each pair of towers by using terrain data made available by NASA [66], which includes buildings and ground clutter, and effectively incorporates the height of the tree canopy.[1] We also require a fully clear Fresnel zone, and adopt $K = 1.3$ and $f = 11$ GHz in the above formulae. The hop engineering routines performing these calculations have been tested in practice: specifically, we have previously used them to design line-of-sight networks, at least 4 of which are now deployed, including ultra-low latency

[1] This NASA data set combines data from the Shuttle Radar Topography Mission (SRTM) [66] and the National Elevation Database (NED) [88], and typically yields acceptably small error ($\sim$ 2 m) against reference, high-accuracy LIDAR measurements.

routes between data centers hosting financial market matching engines. The methodology routinely provided correct clearance assessments when the physical paths were flashed (confirming line-of-sight with an on-site visit, e.g., [33]). It is relatively rare that the hop feasibility assessment is inaccurate; if a problem arises, it is most likely that the locations themselves are not available to rent. For this reason, in §5.2, we explore relaxations of our tower rental assumptions.

After identifying feasible tower-to-tower hops, for each pair of sites, we find the shortest path through a graph containing these hops, which we call a *link*. In line with observations from the tower data around major population centers, we assume each site itself hosts enough towers to use as the starting point for connectivity from that site to many others.

Step 2: topology design. We need to select a subset of site-to-site links to form a nationwide network that minimizes latency, given a limited budget to spend on links. The Steiner-tree problem [41] can be easily reduced to this problem, thereby establishing hardness. Standard approximation algorithms, like linear program relaxation and rounding, yield sub-optimal solutions, which although provably within constant factors of optimal, are insufficient in practice for this setting. Unfortunately, as we show in Appendix A, solving an Integer Linear program "unsplittable flow" formulation is intractable at the scales of interest. We thus propose two heuristics, the combination of which overcomes the scalability challenge, without substantially deteriorating solution quality.

The first observation we make is that the ILP formulation considers some flow variables that will never take non-zero values, allowing us to eliminate them and any resulting null constraints. For instance, if between two end points, a candidate microwave path is of higher latency than a fiber path (which we can always use, at negligible-in-comparison expense), then it will never carry any flow between these two end points. Similar observations apply to individual "distant, off-path" fiber and MW links. This simple observation substantially reduces the problem size. Note that standard network design problems do not typically have this structure available. This is entirely due to the hybrid design using fiber, which is assumed to be cheap, where available. We benefit, in this case, from having an "oracle" that tells us *a priori* when certain flow assignments are "obviously bad" and will not be useful. Further, carefully defined, such constraints preserve optimality; this part of our solution is not an approximation.

Second, we use a fast greedy heuristic to prune out MW links that are unlikely to be chosen. The heuristic operates using a larger budget ($2\times$ in our implementation) than we are ultimately allowed. In each iteration, we add to the solution the MW site-to-site link that decreases average stretch the most, continuing until the total cost reaches the inflated budget; the chosen links are candidates given to the ILP. Intuitively, the other links are uninteresting – they are unlikely to be picked in the final optimization even when a substantially larger budget is available, and so are not presented as options to the ILP. This approach does not provide any guarantees, but we find that on small problem sizes, where the exact ILP can also be evaluated, it obtains the optimal solution.

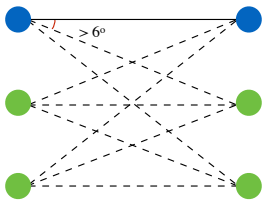


Fig. 1: *k^2 bandwidth with $O(k)$ new towers.*

Step 3: capacity augmentation. In many scenarios, some links require more capacity than a single MW connection. For short distances, this is a non-issue: the MW link can simply be replaced by fiber without a large impact on the network's latency. However, for longer distances, this is not acceptable.

One approach to resolving this problem is simply to build multiple parallel MW links, over multiple series of towers. While tower siting is often a challenging practical problem, with individual sites valued by the HFT industry at as much as $14 million [59], in the cISP context there is a much larger "tolerance" than in HFT, where firms compete for fractions of microseconds. For a 500 km long cISP link, the midpoint diverging 10 km from the geodesic would increase latency by a negligible 0.2%. Thus, the problem of tower siting is substantially simpler. Also, in many cases, tower infrastructure is dense enough already to allow multiple parallel links. For instance, the HFT industry operates nearly 20 parallel networks in the New York-Chicago corridor [55].

We can also employ a simple trick to enhance the effectiveness of parallel series of towers, as shown in Fig. 1. Instead of k parallel series of towers providing merely a $k\times$ bandwidth improvement, connecting multiple antennae on each tower to other towers, we can obtain a $k^2\times$ improvement. Using antennae with overlapping frequencies requires an angular separation of 6° [62], as shown in Fig. 1. Again, the stretch caused by the resulting gap between parallel series of towers is small. For a tower-tower hop distance of 100 km, the minimum distance between two parallel towers should be $100 \cdot \tan(6^\circ) = 10.6$ km, which, as noted above, has a small effect on end-to-end latency for long links.

This approach implies that for site-to-site bandwidths under 1 Gbps, we need just one series of towers; for bandwidths between 1-4 Gbps, we need 2 series; for 4-9 Gbps, 3; etc. While tower siting circumstances are often unique, we are aided by two observations: (a) there is substantial redundancy in existing tower infrastructure, and we can often find existing towers for parallel connections (see Fig. 3b and the related text in §4); and (b) when new towers are needed, there is substantial tolerance in where they are sited, as noted above. Bandwidth may potentially be increased even further through spatial diversity techniques, whereby multiple antennae are placed appropriately on the same tower such that they can adaptively cancel

interference by multiple transmission streams within the same frequency channel [89].

4 A cISP for the United States

We now apply the framework above for a concrete instantiation: designing a cISP for the U.S. mainland. To assess line-of-sight connectivity between existing towers, we use fine-grained data on tower infrastructure, buildings, terrain, and tree canopy. The fiber conduit data is available from past work [34].

Defining the sites and traffic model: To maximize utility while keeping costs low, we connect only the 200 most populous cities in the contiguous United States. In addition, we coalesce suburbs and cities within 50 km of each other, ending up with 120 population centers. (Henceforth, when we refer to "cities", we refer to these population centers.) Based on population data for 2010 [20], we calculate that 85% of the US population lives within 100 km of these 120 cities. For the traffic matrix, we use demands between city pairs that are proportional to their population product.

Step 1: Which city-city links are feasible? We use existing towers listed in FCC's Antenna Structure Registration [39] and databases from American Tower, Crown Castle, and several other tower companies for which we were able to download data. We cull these rather large databases of MW towers to a subset of 12,080 towers as follows: Towers from rental companies are typically suitable for use. From the FCC database, we only use towers over 100 m height. When tower-density exceeds 50 towers per 0.5° square grid cell, we randomly sample towers. (Using all towers could only improve our results, but increases compute time.)

Evaluating link feasibility across tower pairs within range of each other using the aforementioned NASA data [66], we find 261,019 tower-tower hops that satisfy line-of-sight constraints. We find that each city itself has large numbers of suitable towers in its vicinity. We run a shortest path computation on a graph comprising the cities and towers and city-tower and tower-tower hops to find the shortest city-city MW links. This yields both the cost (i.e., number of towers) and latency (i.e., distance along the chosen series of towers) for each city-city link.

For fiber distances, we compute the shortest paths over the InterTubes [34] dataset on US fiber conduits.

Step 2: What subset of links should we build? We use the Gurobi solver [42] to solve our topology design problem. As detailed in Appendix A: (a) both the exact ILP and an LP relaxation approach are too computationally inefficient, while our cISP design heuristic is able to solve the problem at the full scale; and (b) at small scales, where we can also run the exact ILP, our heuristic yields the optimal result.

Fig. 2 shows an example network. Designed with a budget of 3,000 towers and maximum hop length of 100 Km, its average latency is 1.05× c-latency. Fig. 3a shows the reduction of the network's stretch with increases in budget for maximum hop lengths of 70 and 100 Km. Given the similarities with 70 and 100 Km, hereon, we only present results for the latter. An animation, showing how the network structure evolves from mostly-fiber to mostly-MW as the budget increases, is available online [26].

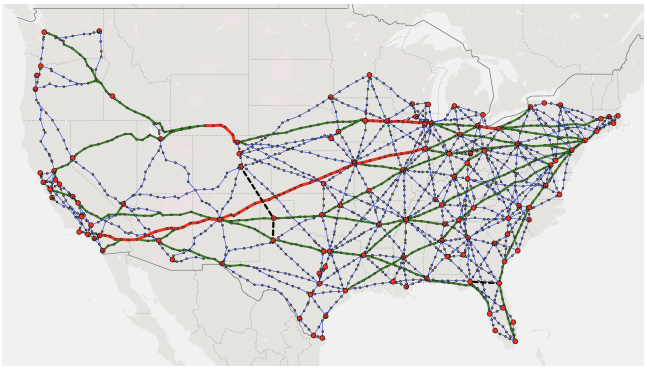
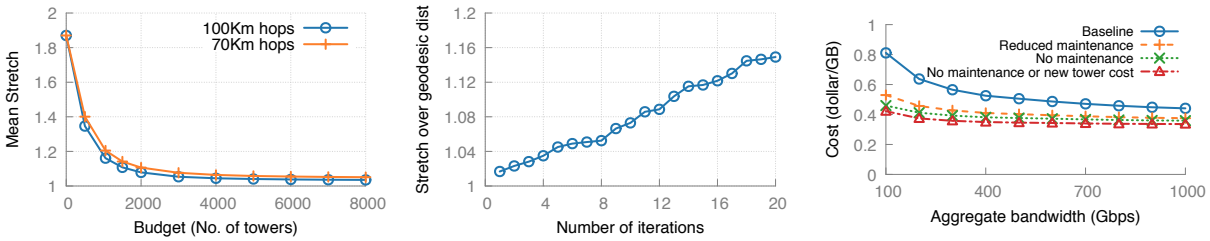


Fig. 2: *A 100 Gbps,* 1.05× *stretch network across 120 cities in the US. Blue links (thin) need no additional towers. Green (thicker) and red links (thickest) need 1 and 2 series of additional towers respectively. Black dashed links are fiber.*

Step 3: Augmenting capacity. We produce a target aggregate demand (i.e., the sum of all site-site traffic demands) by scaling our traffic matrix. Then, each tower-tower MW hop that would be over-utilized (given shortest-path routing and the 1 Gbps capacity from §2) is augmented with additional towers at each end, as described in §3. Fig. 2's topology, when provisioned for an aggregate throughput of 100 Gbps, has 1,660 tower-tower hops that use only already-built towers seen in tower databases, while 552 hops need one additional new tower at each end, and 86 hops need 2 additional towers at each end. Using the cost model described in §2, we find that the cost per GB for this topology, with latency within 1.05× and 100 Gbps throughput, is $0.81. For some context, this is ∼ 10× the cost per GB for content delivery networks [64].

Provisioning even more bandwidth would require more new towers. For 1 Tbps, some tower-tower hops would need as many as 8 additional towers at each end. This is not infeasible — latency would not be inflated excessively, and towers could be found or built. In fact, for the long red link in the map in Fig. 2, which spans 2,700 km from Illinois to California, we find that the *longest* of these 8 additional series of towers would be only 5% longer than the shortest MW path, incurring a stretch of 1.07, instead of 1.02.

We can extend this argument even further: for the same Illinois to California link, we compute tower-disjoint shortest paths, i.e., after finding the shortest path, we remove all towers used by it, find the next-shortest tower-path, etc. In this process, we use only existing towers from our databases, and adhere to the same link feasibility constraints. Fig. 3b shows that stretch increases gradually as we keep eliminating towers; nevertheless, even after 20 such iterations, stretch is much smaller (1.15) than with the existing fiber conduit



(a) (b) (c)

Fig. 3: *(a) Network stretch reduces as we add more MW towers. (b) Stretch for* 20 *shortest tower-disjoint purely MW paths along the long red IL-CA link in Fig. 2. (c) Cost per GB for the city-city traffic model decreases with increasing aggregate throughput.*

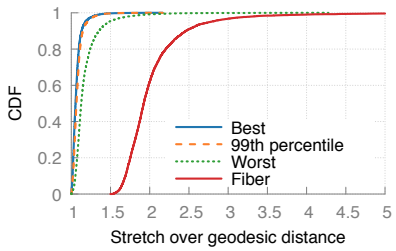


Fig. 4: *Stretch across all city-pairs over a year of weather. The 99th-percentile stretch is comparable to the best stretch.*

(1.75). Note that this route runs through the Rocky mountains and other areas of low tower density. Thus, in accounting for the cost of bandwidth augmentation entirely using the (higher) cost estimates for building new towers, we are substantially overestimating the expense.

There is also another reason our costs are over-estimates: at sufficiently high bandwidth, there is a better option than building many parallel long-distance MW links: one could use the same number of towers to construct a single line of towers with shorter tower-tower distances. This can make shorter-range, but higher-bandwidth technologies like MMW or free-space optics, more cost-effective.

Despite the above two factors, we use parallel MW towers, with all the required additional towers accounted for as new towers, to provide conservative cost-estimates as aggregate bandwidth increases in Fig. 3c.

Routing, queuing, and traffic models. We show in Appendix B that: (a) routing that incurs small (under 10%) latency inflation compared to shortest paths can drive the network at virtually zero loss and minimal queuing delay even at high utilization; and (b) packet pacing addresses the problem of edge links having higher line rates than cISP links. We also discuss evidence for per-MW-hop latency overheads being small enough to ignore. Further, in Appendix C, we show that besides the population-product model, cISP can also be tailored for inter data center traffic, data center to edge traffic, and various combinations of these.

Alternative deployment models. The deployment model and analysis have been conservative in assuming high maintenance and tower installation costs for the provider. What if an incumbent tower company like American Tower [7] deployed cISP? (Fig. 14 in the Appendix shows that American Tower's existing deployment broadly covers areas where our network design of Fig. 2 requires towers.) Besides reduced tower installation costs, maintenance would also be significantly reduced due to the obligation to maintain towers for customers anyway. We evaluated several scenarios of this type, as shown in Fig. 3c. While the solid line represents the baseline deployment model discussed in §2, the dashed lines represent models with reduced maintenance cost ($10K per tower per year), no maintenance cost, and no maintenance or new tower cost (only antenna cost). A network with 3,000 towers offering 100 Gbps bandwidth and 1.05 stretch, built by a company like American Tower, could cost as little as $0.42/GB, thus reducing the baseline cost by almost 50%.

Finally, we note that cISP could be deployed in other geographies besides the US. As discussed in Appendix C.3, we could design a cISP for Europe offering a stretch of 1.04 (vs. 1.05 for the U.S.) with a budget of ∼3k towers.

5 Practical Challenges

Deploying cISP would involve several practical challenges beyond network design and routing, which we now address.

5.1 Impairments due to weather

We use standard equations from MW engineering [48] to calculate signal attenuation due to precipitation. We assume hardware characteristics of a standard low-latency MW radio: an 8-foot dish with a gain of 46.5 dBi at 11 GHz [29, 74, 75]. While antenna gain is determined by the hardware, transmit power and receive power thresholds also depend on the modulation scheme (256 QAM). Following ITU models [48], at ∼11 GHz, precipitation is likely to be the dominant source of attenuation. While the physical layer could trade link bandwidth for higher resilience to weather, we treat the impact of precipitation in a binary manner: if attenuation exceeds a threshold that would degrade bandwidth, we conservatively consider a link to have failed.

We assume that when a link fails, traffic is shifted to the shortest available route, which may use any combination of MW and fiber. The high precipitation that causes failures is easy to predict, especially on the timescale of minutes. Thus, even slow, centralized management would suffice to anticipate failures and reroute accordingly.

We use NASA's precipitation data [65] to determine which links are down when, and what the impact of such failures

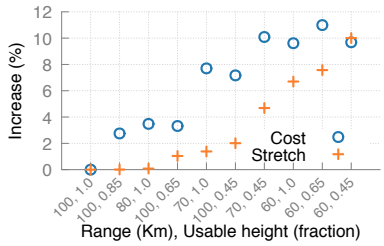


Fig. 5: *As constraints on tower space and range become tighter, the network becomes more expensive, and stretch increases.*

is on the network's latency. For each day over a period of a year (July 2015 - June 2016), we select a 30-minute interval uniformly at random, and identify the links that would fail during it. We then evaluate the latency for each pair of cities end-to-end for each interval. Fig. 4 shows that 99th-percentile latencies are nearly the same as the best fair-weather latencies. In terms of the median across city-pairs, even the worst latencies over the year are 1.7 times lower than those over fiber. Large increases in latency due to weather typically occur only between nearby city-pairs, the fiber route to which runs through a farther-away city, e.g., in Texas, Austin and Killeen fall back to a fiber route through Fort Worth. A more sophisticated analysis allowing dynamic link bandwidth adjustment rather than binary failures can only improve these numbers. Thus, even under significantly adverse weather, most of the latency advantage of cISP remains intact.

We have also created an animated visualization of the network's latency evolving over a year's weather [25].

5.2 Tower height and availability

Our initial design assumed a MW hop to be feasible if it spans a distance of 100 km or less, and satisfies line-of-sight constraints using the tops of the towers. In practice, however, a tower chosen for a route might not have a free spot for a new antenna at the necessary height, especially at the top, where structural concerns for large parabolic antennae are greatest, and where access and maintenance can be problematic. Further, for smaller antennas, insufficient gain margins can decrease the 100 km maximum range. Hence, we evaluate cost and latency of the network with hop-level restrictions modeling these effects.

We test the impact of restricting usable height on towers to three levels, as a fraction of tower height: 0.85, 0.65, and 0.45. Testing for line-of-sight visibility with these restrictions eliminates more towers than using tower tops. We also vary the maximum range, which can necessitate the use of a larger number of towers, thus increasing the cost and potentially making some city-pairs infeasible to connect using MW.

We assess the *percentage increase* in cost and stretch values compared to the baseline values with 100 km range and using the tower tops, i.e., height fraction $= 1$. Fig. 5 shows the results for different combinations of the range and antenna-height constraints, sorted by lowest to highest stretch. The maximum increase in cost is 11% (with the absolute cost per GB under these constraints being \$0.90), while the maximum increase in stretch is 10% (with the absolute stretch compared to the geodesic being 1.16). Thus, even substantial potential problems with mounting antennas do not change our overall conclusions about the viability of cISP.

In our experience designing MW routes, assessments like the ones in this work have yielded accurate estimates of the latency and the number of tower-tower hops that will ultimately be used to connect two sites. The precise set of towers often differs based on real-world constraints, particularly tower unavailability for structural and rental-related reasons. Thus, while accurate in terms of cost and latency, this work does not provide fully engineered routes. In practice, to improve accuracy in preparation for building a MW route, we assign an acquisition probability to each tower in a swath connecting the sites, which depends on a number of factors (e.g., tower type, ownership, and location). Further, for towers that can be acquired, we use a uniform distribution to model height at which space for antennas is available. With this probabilistic model, we compute thousands of candidate MW paths between site pairs, with refinements as acquisitions and height availabilities are confirmed. We make available in video form [24] an example of such refinement.

5.3 Integration into the Internet

We next discuss potential problems cISP may face in terms of integration into the present Internet ecosystem.

Low-hanging fruit: The easiest deployment scenarios involve one entity operating a significant network backbone:

- A CDN could use cISP to carry "back-office" traffic between its locations and content origins, which often supports latency-sensitive user-facing interactions [73]. While the strategies of moving content closer to end users and speeding up the network are orthogonal, on cache misses and when serving uncacheable content, only speeding up the network improves performance.
- Content-providers like Google and Facebook can use cISP to carry lateny-sensitive traffic – such WAN designs already accommodate distinctions between such traffic and background traffic [47, 50].
- Purpose-built networks such as for gaming [40] can easily use cISP between their edge locations and servers.

All of these are interesting and economically viable use cases with minimal deployment barriers, and each *alone* may justify a design like cISP. For instance, while it is tempting to dismiss gaming as a niche, it is a large and growing market: the Steam gaming platform claims 20+ million players worldwide [85]. At a 10 Kbps rate per player [27], this aggregates to 27 Gbps – enough to make cISP viable in this setting. (We present cost-benefit estimates, including for gaming, in §8.)

User-facing deployment: Access ISPs may use cISP as an additional provider, and incorporate a low-latency service

into their broadband plans.[2] Utilizing cISP in this manner can help ISPs to provide and meet the requirements of demanding Service Level Agreements, the case for which was made in recent work [14]. ISPs may use heuristics to classify latency-sensitive traffic and transit it using cISP. Alternatively, software at the user-side may make more informed decisions about which traffic should use the fast-path exposed by the ISP. While this would require significant user-side changes, note that many of today's applications already manage multi-modal WiFi and cellular connectivity.

6 Empirical results

To evaluate the characteristics of long-haul microwave links, we have conducted experiments over one of the most popular nearly-speed-of-light networks deployed in the high-frequency trading corridor between Chicago and New Jersey. We describe these experiments and their results below. The HFT niche is partially characterized by a "winner-takes-all" dynamic which requires these networks to operate at the bleeding edge of low latency. Hence, it is important to quantify the usefulness of these networks in serving more generic low-latency applications on the Internet, which have less-strict latency requirements than HFT, but higher availability and lower packet loss demands.

6.1 Active measurements

We conducted active measurements over the microwave link between the Chicago Mercantile Exchange (CME) data center and the Equinix data center in Secaucus, New Jersey, operated by one of the fastest MW networks in the corridor. On weekdays, when the Chicago and New York markets are open, the link carries financial information critical to high-frequency trading that triggers trades worth billions of dollars. The networks are optimized for low latency, with microseconds of advantage [13] providing a significant edge to customers.

We ran experiments for ∼7 hours every Saturday for 11 weeks between Nov. 2019, and Oct. 2020 from one host each located in the CME and Equinix data centers. The microwave link was provided to us without any Forward Error Correction (FEC), thus being exposed to all errors and bit flips expected in radio transmission. We observe that the link behavior tends to be in one of two states: losses are either very low (normal) or very high (degraded). Out of a total of 72 hours of measurements, there are 12 hours during which the link is degraded due to weather, and 4 hours during which it is down due to maintenance or other issues. Note that because there is no FEC at all, very small bit error rates (BER) degrade the link. Also, in our trading data analysis (§6.2), we see that microwave networks stay up in worse weather conditions than these 12 hours. FEC is needed in packet headers to correct for bit errors, which we could not implement as we did not have access to routers on the network.

[2] While large last-mile latencies can overshadow cISP's low latency, this is an entirely orthogonal problem, on which significant progress is being made – 5G prototypes are already showing off sub-millisecond latencies [46].

6.1.1 RTT and bandwidth

The geodesic distance between the CME and Equinix data centers is 1139.5 km. The c-latency for a round-trip, then, is 7.6 ms. In our experiments over 11 weeks, we always observe a round-trip time of 7.7 ms for 32-byte packets, i.e., within 1.5% of c-latency. The RTT goes up to 7.9 ms for 1,499-byte packets because of the limited bandwidth available on the link (or more specifically, the slice of it provided to us).

The 0.1 ms increase in transmission delay as packet size increases by 1,467 bytes gives a bandwidth estimate of 120 Mbps. Our UDP measurements and TCP measurements, in the best case, also give us a bandwidth of 120 Mbps. It is hard for TCP to sustain throughput at this rate in the absence of any FEC because of transmission losses. While the operator did not divulge the exact link capacity, it is likely that our network access was capacity-capped. Hence, these measurements only provide a lower bound on the link bandwidth.

6.1.2 Loss and FEC

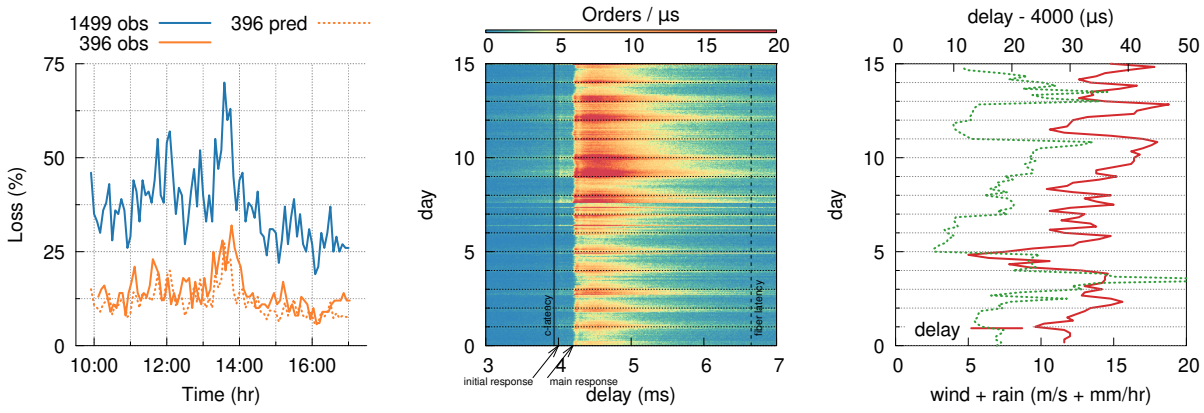
In plain TCP (iperf) and ICMP (ping) probes, we observe high loss rates: typically around 3% to 5% for 32-byte packets. The packet loss rate increases sharply as packet size increases because more bits can potentially be corrupted in transmission. Without FEC, a link with loss rate this high is clearly unsuitable for web traffic [91]. Whether FEC can bring the loss rate down to an acceptable level (say, 0.1%) at reasonable latency and bandwidth overhead depends on two factors: 1. the Bit Error Rate (BER), and 2. the typical length of error bursts, i.e., how many consecutive bits are corrupted in an error burst. We elaborate on these factors below.

First, we derive the underlying BER from observed ping packet loss. For a ping packet of s bytes, a successful response is observed when both the echo request and reply packets are delivered to the respective hosts without any errors. To estimate the BER b_{err}, we first assume that bit errors are uniform and random. Then, for packet loss rate p_{loss}, we get:

$$b_{err} = 1 - (1 - p_{loss})^{1/(2\times 8\times s)}$$

For initial validation of this model, with the possibly unjustified assumption of random and uniform errors, we calculate b_{err} from observed p_{loss} for $s = 1{,}499$ for the 7 hours of measurements on Feb. 15th, 2020. Then, we use the calculated b_{err} to predict p_{loss} for $s = 396$ on the same day. We compare the predicted and observed values in Fig. 6a. While the observed and predicted loss rates for $s = 396$ largely agree, there are some disagreements, e.g., at 12:30, which can be explained by the fact that the observations for $s = 1{,}499$ and $s = 396$ are separated in time by 60 seconds. The underlying BER might change during this interval. For Feb. 15th, the median, 95^{th} percentile, and maximum BER we calculate are 3.6×10^{-5}, 8.2×10^{-5}, and 3.6×10^{-4} respectively.

For a target packet loss rate of 0.1% for packets of size 1,500 bytes, the BER needs to be 4.17×10^{-8} or lower. Extremely lightweight FEC codes, such as Reed-Solomon (255,



(a) **(b)** **(c)**

Fig. 6: *(a) Predicting loss rate of* 396 *byte packets from observed loss rates of 1,499 byte packets on Feb 15th, 2020. Analyzing trading data: (b) Heat map of order book events at delay between Chicago and New Jersey. Response delay never exceeds* 4.3 ms*; (c) A coarse weather signal (max wind speed + max rainfall) is correlated with the observed transmission delay.*

239) can correct from BER of 10^{-4} to 10^{-12} with a bit rate overhead of only 7% [76]. If performed over 255 byte blocks, a 1,500 byte packet can be encoded in 7 blocks with a total redundancy overhead of 112 bytes. At 120 Mbps bandwidth, this incurs a latency penalty of only 7.5 µs. This FEC scheme would break down, however, if errors occurred in bursts of around 8 bytes or more. Now we discuss the earlier assumption of error bursts being short and uniformly distributed.

To analyze bit errors, we sent two sets of UDP probes over the link: the first set consists of 60 byte packets sent at 35 packets per second (slow), and the second consists of 60 byte packets sent at 200,000 packets per second (fast). The slow set characterizes link behavior with no congestion/bandwidth related losses, whereas the fast set provides statistical significance to rare bit flip events. In contrast to ping losses, losses in this experiment are observed through packet captures rather than at the application layer, so a corruption of, e.g., the UDP destination port would not register a loss. For the slow set, we observe a packet loss rate of 0.8%, whereas for the fast set we observe a loss rate of 2.04%.

In the UDP fast set a packet has 4 bytes of payload, 8 bytes of UDP header, 20 bytes of IP header, 14 bytes of Ethernet header, and 14 bytes of padding. A total of 1.6 billion packets were sent, out of which 2.66 million were received on the other end with at least one of the following fields corrupted: source port, destination port, UDP header length field, and payload. We calculate the Hamming distance between the received value and the expected value of the corrupted fields. As Table. 7a shows, there appears to be a linear relationship between field size and number of corruptions, and over 99% of all corruptions consist of 2 bit flips or less. Also, if we extrapolate the errors we observe in these 4 fields to the rest of the 60 byte packet, the expected loss rate due to corruptions in the Ethernet and IP headers and padding matches that observed in the UDP slow set. The other 1.24% packets lost can thus be explained by congestion/bandwidth issues.

6.2 Trading data analysis

To characterize the latency and up-time of the full range of microwave links deployed in the Chicago-New Jersey corridor, we analyze trading data from the Chicago Mercantile Exchange (CME) in Chicago, Illinois, and the CBOE Options Exchange in Secaucus, New Jersey. Information about trades happening at the CME travels over microwave paths and triggers activity at the CBOE [13]. The time difference between stimulus events at the CME and the response at the CBOE represents the network latency between the two exchanges. Laughlin et al. have also used this methodology to estimate latency between financial markets [55].

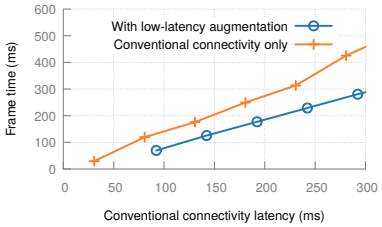
We obtained tick data from CME and CBOE for three weeks of Mar. 2019. The tick data consists of microsecond precision timestamps for events at both ends. Both markets are open simultaneously for 6.5 hours every weekday, which means that we have 97.5 hours of relevant tick data. For each trade executed at the CME at timestamp t, we count the number of order book events at the CBOE at timestamps $t+i$ where $i \in [3000, 7000]$ µs. Fig. 6b plots a heat map of the number of orders per µs for each 10 µs bin in the tick data. The y-axis time is in intervals of 15 minutes. Analysis of the data shows that the main response delay, which reflects the network latency between CME and CBOE, does not exceed 4.3 ms for any 15-minute interval. The lowest fiber latency between the two exchanges is 6.65 ms [60]. This shows that some microwave networks were up through every 15-minute interval over the 3-week period.

In addition to the main response at 4.2 ms, Fig. 6b has a smaller initial response at 4.0 ms. The CME tick data reveals that internal trading algorithms and strategies produce a second stimulus at CME 200 µs after the initial stimulus. The main response in Fig. 6b is triggered by that second stimulus.

We consider the delay between the second stimulus at CME and the main response at CBOE as transmission delay. We calculate the transmission delay for every 1-hour interval

Field	*#corruptions*	*#bits*	*1 bit flip*	*2 bit flips*
src port	873,165	16	84%	15%
dst port	864,955	16	82%	17%
length	914,528	16	85%	14%
payload	1,734,539	32	84%	15%

(a)



(b)

(c)

Fig. 7: *(a) Corruptions observed in the UDP fast set. (b) A substantial reduction in frame time can be obtained by the use of a parallel low-latency augmentation to the present Internet. (c) Mean web page load time (PLT) improvement for each heuristic and its portion of traffic delivered on cISP. PLT can improve substantially by only offloading a small portion of traffic to cISP.*

in the tick data. Fig. 6c plots the moving average of transmission delay over 2 hours. We use the hourly wind speed estimate [30] and rainfall data [22] in the regions through which the MW corridor passes as a coarse weather signal. For each hour, we pick the maximum wind speed and maximum rainfall observed at a granularity of ~ 10km along the geodesic between the end points. Fig. 6c plots wind speed $+$ rainfall $/2$, and shows that there is some correlation. The Pearson correlation coefficient between wind and delay is 0.24, while that between rain and delay is 0.16. Sources of noise in this correlation include the noise inherent in the trading data itself, and issues that may affect transmission delay, such as infrastructure damage or operational downtime. Note that days 3 and 14 have more severe rain and wind than the 12 hours during which the link was degraded in our active measurements (§6.1).

Conclusions: From the active measurements, we conclude that for our MW path, (1) round-trip latency is less than 1.5% inflated over *c*-latency, (2) bandwidth is at least 120 Mbps, (3) error bursts are very short and roughly uniformly distributed under normal link conditions, and (4) errors can be brought down to acceptable levels with extremely lightweight FEC incurring minimal latency and bandwidth overhead.

From the trading data analysis, we conclude that (1) for the 97.5-hour period, some MW networks, spanning more than 1,000 km, were always up without any significant degradation in latency, and (2) weather events such as high wind speeds and rainfall are correlated with increases in transmission delay by tens of microseconds. This increase may stem from one or more of the following: (a) longer end-to-end paths being picked, (b) shorter tower-to-tower hops leading to higher switching delay, and (c) the network responding to weather changes by ramping up FEC.

7 A Few Potential Applications

Several applications require low latency over the wide area-network. Applications focused on user interactivity, such as augmented and virtual reality, tele-presence and tele-surgery, musical collaboration over long-distances, etc., can all benefit from low-latency connectivity. Likewise, less user-centric applications, such as real-time bidding for Web page advertisements [8] and block propagation in blockchains, would also benefit. While it is beyond the scope of this paper to analyze this in detail, we assess, in simplified environments, the improvements cISP could achieve for two applications.

7.1 Online gaming

We discuss cISP's benefits for both models of online gaming: **thin-client** (where a client essentially streams everything in real-time from a server) and **fat-client** (where the client has the game installed, performs computations, etc., and only relies on the server for updates on the global game state).

Fat-clients are dominant today, and are easy to tackle: communication is almost entirely composed of latency-sensitive player actions and game-state changes, and is low-volume, typically a few Kbps per client for popular games [27]. It can all be transferred over the low-latency network, reducing latency by 3-4$\times$ compared to today's Internet.

Thin-client gaming is still in its infancy, as it depends heavily on the network, with data rates in Mbps. We explore the potential of a speculative approach: the server speculates on the game state and sends data for multiple scenarios in advance over fiber, then on the low-latency network, issues messages indicating which scenario occurred. Such speculation has already shown success for rich games like "Doom 3" [56].

We use a toy thin-client for a multi-player Pacman variant to explore the latency benefit. Our rudimentary implementation speculates on all 4 movement directions possible as user input. In line with the online-gaming literature, we measure "frame-time," which "corresponds to the delay between a user's input and the observed output" [56]. We evaluate frame-time as latency over conventional connectivity increases (emulated by adding latency in software), and for a low-latency network always incurring $1/3$ of the latency of the corresponding conventional network.

As Fig. 7b shows, the speculative approach enabled by the low-latency network augmentation reduces frame-time. This comparison would improve further if non-network overheads from processing and rendering in our naive implementation were smaller. We do not use any heavy graphics on which to evaluate the additional bandwidth overhead on fiber, but even in the sophisticated scenarios examined by prior work [56], this bandwidth overhead can be contained to 2-4.5$\times$.

7.2 Web Browsing

We evaluate the potential impact of cISP's latency improvement on Web page load times (PLTs) (based on the *onLoad* event [71]) using Mahimahi [68] with the addition of content delivery network (CDN) caching. Our emulation supports two levels of the CDN cache hierarchy. The client's request first reaches the edge server. If it is a cache miss, the request will be forwarded to the parent server. In case there is another miss, it will be forwarded to the origin server. This setup thus allows variable request latency, where certain requests can experience more latency.

To realistically recreate the caching behavior, our experiments leverage the Akamai pragma header [3] which is typically used for debugging purposes. We select web pages where at least 75% of the HTTP requests[3], performed when loading a page, are served by Akamai. Overall, we found 27 landing pages and 140 associated internal pages from the Hispar list [9] match this criterion. We record each page's content and the network latency for each (edge) server that a client contacted when loading a page. This recording process is conducted from three different vantage points at three different times. For the CDN server-to-server latency, we estimate the latency by geolocating the IP addresses of the CDNs and origin servers provided by the pragma header[4]. We then replay each page with unmodified network latencies (as a baseline) and with latencies reduced to 0.33× of their original values (as a cISP). No bandwidth limitations are imposed.

Fig. 7c shows the results. Compared to the baseline, a 66% reduction in latencies (all-cISP) results in a mean 42% PLT decrease for both landing and internal pages (an absolute decrease of 600 ms and 651 ms). This PLT reduction is less than the 66% reduction in RTT because loading a Web page also involves significant non-network activity.

If cISP is used only to deliver the CDN's server-to-server (i.e., back-office) traffic, our experiment (backhaul-cISP) suggests that PLT can be improved by 23.7% and 28.5% (331 ms and 447 ms) on landing and internal pages by only sending 13.4% and 22.3% of the overall web-browsing traffic on cISP. Internal pages get better improvement and send a higher proportion of traffic because they experience more cache misses (31.9%) compared to landing pages (13.3%).

While Web-browsing traffic comprises only a small fraction of total Internet traffic[5], we can further reduce the load by carrying only *latency-sensitive* traffic on cISP. Hence, we extend Mahimahi to enable *selective* manipulation of RTTs in the replay, such that some traffic sees lower RTTs than other traffic. We test two heuristics under this setup. First, we try a simple heuristic that only sends uplink traffic to cISP (uplink-cISP). This approach yields a mean PLT improvement of 21.5% (319 ms) by sending only 9.7% of the web-browsing traffic over cISP. Second, we adopt a more advanced PKT-State heuristic [77] (packet-cISP) to distinguish the latency-sensitive traffic (e.g., TCP SYN/ACK packets and small data packets) from the bandwidth-intensive traffic (e.g., data packets). By offloading the latency-sensitive traffic to cISP, we can get a mean PLT improvement of 28.2% (417 ms) by only offloading 10.2% of the traffic.

[3] We assume requests not served by Akamai are served by the edge server.

[4] We geolocate each server, and compute server-to-server c-latency from distance. Then, we estimate baseline latency as 3× c-latency.

[5] Cisco's 2018 estimate puts "Web/Data traffic" at 13% [23] including non-latency sensitive traffic like software updates and some file transfers.

8 Cost-benefit and market analysis

Does cISP's value justify its cost? For three important use cases, we present quantitative lower-bound estimates of cISP's value per GB. cISP would also need enough aggregate demand across one or more use cases to support its total deployment cost, so we estimate market size of each use case.

Web search. *Value per GB:* Putting together Google's quantification of the impact of latency in search [18], their estimated search revenue restricted to the US [63], their search volume [84], estimated data transferred per search[6], and estimated cost per search [53], we estimate that speeding up page load times for 12 Gbps of their US search traffic by only 200 ms (400 ms) would yield an additional yearly profit of $87 ($177) million. This translates to an added value of $1.84 ($3.74) per GB. *Market size:* At 12 Gbps of traffic, Google's search traffic is a nontrivial fraction ($> 10\%$) of a cISP provisioned to provide ∼100 Gbps, but to make cISP viable, it would have to be augmented with other use cases.

E-commerce. *Value per GB:* Using Amazon.com's estimates of number of visits, pages fetched per visit, fraction of US traffic [80], and page size, we arrive at an estimated 480 PB of US traffic per year. Using their US sales [32] and profit margin of 5.5% [61] gives an estimated $16.3 billion in profits per year. Estimates for the effect of PLT on conversion rate vary from 1% [57] to 2.4% (on desktop) and 7% (on mobile) per 100 ms of additional latency [5]. Thus, saving 200 ms by sending only 10% of the data over cISP (§7.2), translates to a value of $6.8-$47.5 per GB, which is much higher than the $0.81 per GB cost of cISP traffic. *Market size:* 10% of 480 PB of Amazon e-commerce annual US traffic translates to 12 Gbps of cISP traffic. But the current (2020) e-Commerce market size of $861 billion [32] (compared to Amazon's ∼$296 billion) proportionately translates to a cISP traffic demand of 35 Gbps. Given the high value per GB, this use case alone could make a 100 Gbps cISP profitable.

Gaming. *Value per GB:* Online gamers often pay for "accelerated VPNs", which promise to lower network latency. Such services cost $4-$10 per client per month [1, 11, 72]. Full-time gaming at 8 hours a day at a 10 Kbps rate (as in §5.3) translates to 1.08 GB / month. Thus, if cISP were priced like a cheap accelerated VPN service at $4 / mo, this would translate to a value of at least $3.7 / GB. A less aggressive model than "full-time gaming" would only improve cISP's value. Another indicator of latency's value in gaming is the

[6] From Firefox desktop's network tools; mobile responses may be smaller.

market for gaming monitors with high screen-refresh rates: the 6-10 ms of latency advantage is valued at over $50 by many gamers, estimated from the pricing of monitors which are exactly the same except in terms of refresh rate [6]. *Market size:* There are more than 350 million [83] Fortnite gamers worldwide. Assuming 20% of the gamers are in the US, each with a demand of 10 Kbps, translates to 700+ Gbps of cISP demand. Even for games with smaller user bases like PUBG (70 million) and Call of Duty Warzone (100 million), cISP demands are high enough to sustain a nationwide network.

Summary. The value per GB obtained from cISP's latency reduction in the above cases – $1.84-$3.74, $6.52-$45.63, and over $3.70 – exceeds its cost estimate of $\leq$ $0.81 per GB, and even leaves room for substantial over-provisioning. Total addressable market demand could greatly exceed a 100 Gbps cISP for the case of gaming, and for web-based use cases could be sufficient to support the infrastructure.

This simplified analysis omits many factors. Not all users would be paying for the infrastructure on day 1, so an incremental roll-out for a smaller set of customers would be important. Also, there are many other applications that can benefit from cISP. CDNs routinely use overlay routing to cut latency for dynamic, non-cacheable content, for which edge replication is difficult or ineffective [4]. Upcoming application areas like virtual and augmented reality can only make the case stronger for cISP. We expect cISP's most valuable impact to be in breaking new ground on user interactivity, as explored in some depth in prior work [16].

9 Related Work

Networking research has made significant progress in measuring latency, as well as improving it through transport, routing, and application-layer changes. However, the underlying infrastructural latency has received little attention and has been assumed to be a given. This work proposes a speed-of-light ISP, demonstrating that improvements are indeed possible.

There are several ongoing Internet infrastructure efforts, including X moonshot factory's project Taara [90], Facebook connectivity's Magma [36], Rural Access [37], Terragraph [38], and the satellite Internet push by Starlink [81], Kuiper [54], Telesat [86], and others. Project Taara consists of networks under deployment in India and Africa, based on free-space optics, and described as "Expanding global access to fast, affordable internet with beams of light". While Facebook's Magma and Rural Access aim to extend connectivity to rural areas by offering a software, hardware, business model, and policy framework, Terragraph aims to extend last-mile connectivity to poorly connected urban and suburbans areas by leveraging short millimeter-wave hops. Free-space networks of this type will likely become more commonplace in the future, and these works are further evidence that many of the concerns with line-of-sight networking can indeed be addressed with careful planning. Further, cISP's design approach is flexible enough to incorporate a variety of media (fiber, MW, MMW, free-space optics, etc.) as the technology landscape changes.

"New Space" satellite networks: While low-Earth orbit (LEO) satellite networks can reduce long-distance latency [12, 44, 52], current deployments are more targeted at last-mile connectivity than long haul [15]. Starlink recently claimed to offer last-mile round-trip latency of 31 ms [82], more than 3.8$\times$ the latency estimated in prior simulations [12], showing that the service is not yet latency optimized.

Despite the apparent differences in objectives — long haul latency for cISP and last-mile connectivity for LEO networks — it is useful to **coarsely** assess how the costs may compare. Starlink, for example, offers uncapped connectivity at $99/month [78]. At an average household consumption of 273.5 GB [35], this translates to $0.36/GB[7]. For cISP, if an incumbent like American Tower were to deploy it, the cost could be as low as $0.33/GB, as shown in Fig. 3c. Thus, a network with costs comparable to cISP (in a per-bit sense; cISP is more than an order of magnitude cheaper in absolute cost, and has commensurately lower bandwidth) is concurrently being deployed, albeit with different goals.

To the best of our knowledge, the only efforts primarily focused on wide-area latency reduction through infrastructural improvements are in niches, such as the point-to-point links for financial markets [55], and isolated submarine cable projects aimed at shortening specific Internet routes [67, 69].

10 Conclusion

A speed-of-light Internet not only promises significant benefits for present-day applications, but also opens the door to new possibilities, such as *eliminating the perception of wait time* in our interactions over the Internet [16]. We thus present a design approach for building wide-area networks that operate nearly at *c*-latency. Our solution integrates line-of-sight wireless networking with the Internet's fiber infrastructure to achieve both low latency and high bandwidth.

A speed-of-light Internet has not always been clearly viable. The enabling technology of low-latency multi-hop microwave networks was spurred on by HFT only within the last 10 years, and even then it has not been a priori obvious that the challenges of relatively high loss and low bandwidth could be overcome to leverage such links for an Internet backbone. More importantly, the Internet has become increasingly latency-limited due to increasing bandwidths and greater use of interactive applications. Thus, we believe we have reached an exciting point in time when greatly reducing the Internet's infrastructural latency is not only tractable, but surprisingly cost-effective and impactful for applications.

Acknowledgements

This work was supported by National Science Foundation Awards CNS-1763492, CNS-1763742, and CNS-1763841.

[7] Starlink is currently in beta testing, and profit margins are unclear. It is difficult to do a tighter cost analysis for Starlink without more information.

REFERENCES

[1] AAA Internet Publishing, Inc. WTFast. https://www.wtfast.com/en/. [Online; accessed 11-March-2021].

[2] Akamai. Akamai "10for10". https://www.akamai.com/us/en/multimedia/documents/brochure/akamai-10for10-brochure.pdf, July 2015. [Online; accessed 11-March-2021].

[3] Akamai. Using Akamai Pragma headers to Investigate or Troubleshoot Akamai Content Delivery. https://community.akamai.com/customers/s/article/Using-Akamai-Pragma-headers-to-investigate-or-troubleshoot-Akamai-content-delivery?language=en_US, 2015. [Online; accessed 11-March-2021].

[4] Akamai. SureRoute. https://developer.akamai.com/learn/Optimization/SureRoute.html, 2017. [Online; accessed 11-March-2021].

[5] Akamai. The State of Online Retail Performance. https://www.akamai.com/uk/en/multimedia/documents/report/akamai-state-of-online-retail-performance-spring-2017.pdf, 2017. [Online; accessed 11-March-2021].

[6] amazon.com. ASUS VG248QE Gaming Monitor. https://goo.gl/gnFnPv, 2018. [Online; accessed 11-March-2021].

[7] American Tower Global Wireless Solutions. https://www.americantower.com/us/, 2004. [Online; accessed 11-March-2021].

[8] Waqar Aqeel, Debopam Bhattacherjee, Balakrishnan Chandrasekaran, P. Brighten Godfrey, Gregory Laughlin, Bruce Maggs, and Ankit Singla. Untangling header bidding lore: Some myths, some truths, and some hope. In *Passive and Active Measurement*, 2020.

[9] Waqar Aqeel, Balakrishnan Chandrasekaran, Bruce Maggs, and Anja Feldmann. On landing and internal pages: The strange case of Jekyll and Hyde in Internet measurement. In *ACM IMC*, 2020.

[10] AT&T Corporation. AT&T Long Lines Routes March 1960. http://long-lines.net/places-routes/maps/MW6003.html, 2003. [Online; accessed 11-March-2021].

[11] Battleping. Info on our lower ping service. http://www.battleping.com/info.php, 2010. [Online; accessed 11-March-2021].

[12] Debopam Bhattacherjee, Waqar Aqeel, Ilker Nadi Bozkurt, Anthony Aguirre, Balakrishnan Chandrasekaran, P Godfrey, Gregory Laughlin, Bruce Maggs, and Ankit Singla. Gearing up for the 21st century space race. In *ACM HotNets*, 2018.

[13] Debopam Bhattacherjee, Waqar Aqeel, Gregory Laughlin, Bruce M. Maggs, and Ankit Singla. A bird's eye view of the world's fastest networks. In *ACM IMC*, 2020.

[14] Zachary S. Bischof, Fabián E. Bustamante, and Rade Stanojevic. The Utility Argument - Making a Case for Broadband SLAs. In *PAM*, 2017.

[15] Bloomberg. Musk targets telecom for next disruption with Starlink Internet. https://tinyurl.com/wejrv37c, 2021. [Online; accessed 11-March-2021].

[16] Ilker Nadi Bozkurt, Anthony Aguirre, Balakrishnan Chandrasekaran, Brighten Godfrey, Gregory Laughlin, Bruce M. Maggs, and Ankit Singla. Why Is the Internet so Slow?! In *PAM*, 2017.

[17] Ilker Nadi Bozkurt, Waqar Aqeel, Debopam Bhattacherjee, Balakrishnan Chandrasekaran, Philip Brighten Godfrey, Gregory Laughlin, Bruce M. Maggs, and Ankit Singla. Dissecting latency in the Internet's fiber infrastructure, 2018. arXiv:1811.10737.

[18] Jake Brutlag. Speed Matters for Google Web Search. http://goo.gl/vJq1lx, 2009. [Online; accessed 11-March-2021].

[19] Gustavo Carneiro, Pedro Fortuna, and Manuel Ricardo. FlowMonitor: A Network Monitoring Framework for the Network Simulator 3 (NS-3). In *Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools*, VALUETOOLS '09, 2009.

[20] 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. [Online; accessed 11-March-2021].

[21] Michael Chow, David Meisner, Jason Flinn, Daniel Peek, and Thomas F. Wenisch. The mystery machine: End-to-end performance analysis of large-scale internet services. In *USENIX OSDI*, 2014.

[22] CHRS at UC Irvine. PERSIANN-CCS. https://chrsdata.eng.uci.edu/, 2017. [Online; accessed 11-March-2021].

[23] Cisco. Cisco Visual Networking Index: Forecast and Methodology. https://www.reinvention.be/webhdfs/v1/docs/complete-white-paper-c11-481360.pdf, 2017. [Online; accessed 11-March-2021].

[24] cISP authors. MW path refining. https://goo.gl/LwYB5Z. [Online; accessed 11-March-2021].

[25] cISP authors. Impact of rainfall on cISP for a period of 1 year. https://tinyurl.com/a8szcukz, 2021. [Online; accessed 11-March-2021].

[26] cISP authors. The MW+fiber hybrid network evolves with budget. https://tinyurl.com/3vakxccm, 2021. [Online; accessed 11-March-2021].

[27] Mark Claypool, David LaPoint, and Josh Winslow. Network analysis of Counter-Strike and Starcraft. In *IEEE Performance, Computing, and Communications Conference*, 2003.

[28] Federal Communications Commission. Universal Licensing System. http://wireless2.fcc.gov/UlsApp/UlsSearch/searchLicense.jsp. [Online; accessed 11-March-2021].

[29] CommScope. HSX8-107-D3A. https://objects.eanixter.com/PD354739.PDF, 2012. [Online; accessed 28-July-2021].

[30] Copernicus by ECMWF. ERA5 hourly data on single levels from 1979 to present. https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview, 2018. [Online; accessed 11-March-2021].

[31] DARPA. Novel Hollow-Core Optical Fiber to Enable High-Power Military Sensors. http://www.darpa.mil/news-events/2013-07-17, 2013. [Online; accessed 11-March-2021].

[32] Digital Commerce 360. US ecommerce grows 44.0% in 2020. https://www.digitalcommerce360.com/article/us-ecommerce-sales/, 2021. [Online; accessed 28-July-2021].

[33] DragonWave-X. Services & Support / Pre Deployment / Line of Sight. https://www.dragonwavex.com/services/pre-deployment/line-sight, 2021. [Online; accessed 11-March-2021].

[34] Ramakrishnan Durairajan, Paul Barford, Joel Sommers, and Walter Willinger. InterTubes: A study of the US long-haul fiber-optic infrastructure. In *ACM SIGCOMM*, 2015.

[35] Joan Engebretson. Broadband Data Usage Report: Internet-only Homes Use Almost Twice as Much Data as Bundled Homes. https://www.telecompetitor.com/broadband-data-usage-report-internet-only-homes-use-almost-twice-as-much-data-as-bundled-homes/, 2019. [Online; accessed 11-March-2021].

[36] Facebook connectivity. Magma. https://connectivity.fb.com/magma/, 2021. [Online; accessed 11-March-2021].

[37] Facebook connectivity. Rural Access. https://connectivity.fb.com/rural-access/, 2021. [Online; accessed 11-March-2021].

[38] Facebook connectivity. Terragraph. https://connectivity.fb.com/terragraph/, 2021. [Online; accessed 11-March-2021].

[39] Federal Communications Commission. Antenna Structure Registration Database. https://www.fcc.gov/antenna-structure-registration, 2018. [Online; accessed 11-March-2021].

[40] Riot Games. Fixing the Internet for real-time applications. https://goo.gl/SEoxW2, 2016. [Online; accessed 11-March-2021].

[41] M. R. Garey and D. S. Johnson. The rectilinear Steiner tree problem is NP-complete. *SIAM Journal on Applied Mathematics*, 32(4):826–834, 1977.

[42] Inc. Gurobi Optimization. Gurobi optimizer reference manual, 2016.

[43] Nikola Gvozdiev, Stefano Vissicchio, Brad Karp, and Mark Handley. Low-latency routing on mesh-like backbones. *ACM HotNets*, 2017.

[44] Mark Handley. Delay is not an option: Low latency routing in space. In *ACM HotNets*, 2018.

[45] Jonas Hansryd and Jonas Edstam. Microwave capacity evolution. *Ericsson review*, 1:22–27, 2011.

[46] Devindra Hardawar. Samsung proves why 5G is necessary with a robot arm. https://goo.gl/3gZTn8, 2016. [Online; accessed 11-March-2021].

[47] Chi-Yao Hong, Srikanth Kandula, Ratul Mahajan, Ming Zhang, Vijay Gill, Mohan Nanduri, and Roger Wattenhofer. Achieving high utilization with software-driven WAN. In *ACM SIGCOMM*, 2013.

[48] ITU. Specific attenuation model for rain for use in prediction methods. http://www.itu.int/dms_pubrec/itu-r/rec/p/R-REC-P.838-3-200503-I!!PDF-E.pdf, 2005. [Online; accessed 11-March-2021].

[49] Ixia. Measuring Latency in Equity Transactions. http://ixia.cabanday.com/products/_content/wp-measuring-latency.pdf, 2012. [Online; accessed 11-March-2021].

[50] Sushant Jain, Alok Kumar, Subhasree Mandal, Joon Ong, Leon Poutievski, Arjun Singh, Subbaiah Venkata, Jim Wanderer, Junlan Zhou, Min Zhu, Jon Zolla, Urs Hölzle, Stephen Stuart, and Amin Vahdat. B4: experience with a globally-deployed software defined wan. In *ACM SIGCOMM*, 2013.

[51] Srikanth Kandula, Dina Katabi, Bruce Davie, and Anna Charny. Walking the tightrope: Responsive yet stable traffic engineering. In *ACM SIGCOMM*, 2005.

[52] Simon Kassing, Debopam Bhattacherjee, André Baptista Águas, Jens Eirik Saethre, and Ankit Singla. Exploring the "Internet from space" with Hypatia. In *ACM IMC*, 2020.

[53] Kevin Kelly. How much does one search cost? http://kk.org/thetechnium/how-much-does-o/, 2007. [Online; accessed 11-March-2021].

[54] Kuiper Systems LLC. Application of Kuiper Systems LLC for Authority to Launch and Operate a Non-Geostationary Satellite Orbit System in Ka-band Frequencies. https://licensing.fcc.gov/myibfs/download.do?attachment_key=1773885, 2019.

[55] Gregory Laughlin, Anthony Aguirre, and Joseph Grundfest. Information transmission between financial markets in Chicago and New York. *Financial Review*, 2014.

[56] Kyungmin Lee, David Chu, Eduardo Cuervo, Johannes Kopf, Yury Degtyarev, Sergey Grizan, Alec Wolman, and Jason Flinn. Outatime: Using speculation to enable low-latency continuous interaction for mobile cloud gaming. In *ACM MobiSys*, 2015.

[57] Greg Linden. Make Data Useful. https://slideplayer.com/slide/4203392/, 2006. [Online; accessed 11-March-2021].

[58] McKay Brothers LLC. Quincy Extreme Data Latencies. http://www.quincy-data.com/product-page/#latencies, 2017. [Online; accessed 11-March-2021].

[59] Brian Louis. Trading Fortunes Depend on a Mysterious Antenna in an Empty Field. https://goo.gl/82kzXd, 2017. [Online; accessed 11-March-2021].

[60] Donald MacKenzie. *Trading at the Speed of Light: How Ultrafast Algorithms Are Transforming Financial Markets*. Princeton University Press, 2021.

[61] Macrotrends LLC. Amazon Net Profit Margin 2006-2021. https://www.macrotrends.net/stocks/charts/AMZN/amazon/net-profit-margin, 2021. [Online; accessed 28-July-2021].

[62] Trevor Manning. *Microwave Radio Transmission Design Guide*. Artech House, 2009.

[63] Ginny Marvin. Report: Google earns 78% of $36.7B US search ad revenues, soon to be 80%. https://goo.gl/kp4L5X, 2017. [Online; accessed 11-March-2021].

[64] Microsoft Azure. Content Delivery Network pricing. https://azure.microsoft.com/en-us/pricing/details/cdn/, 2018. [Online; accessed 11-March-2021].

[65] NASA. Precipitation Processing System Data Ordering Interface for TRMM and GPM (STORM). https://storm.pps.eosdis.nasa.gov/storm/, 2015. [Online; accessed 11-March-2021].

[66] NASA Jet Propulsion Laboratory. U.S. Releases Enhanced Shuttle Land Elevation Data. https://www2.jpl.nasa.gov/srtm/, 2015. [Online; accessed 11-March-2021].

[67] NEC. SEA-US: Global Consortium to Build Cable System Connecting Indonesia, the Philippines, and the United States. https://tinyurl.com/ybj9nhp3, August 2014. [Online; accessed 11-March-2021].

[68] Ravi Netravali, Anirudh Sivaraman, Somak Das, Ameesh Goyal, Keith Winstein, James Mickens, and Hari Balakrishnan. Mahimahi: Accurate record-and-replay for http. In *USENIX ATC*, 2015.

[69] A. Nordrum. Fiber optics for the far north [news]. *IEEE Spectrum*, 52(1):11–13, January 2015.

[70] ns-3 community. Network simulator ns-3. https://www.nsnam.org, 2011. [Online; accessed 11-March-2021].

[71] Jan Odvarko. Har 1.2 spec. http://www.softwareishard.com/blog/har-12-spec, 2007. [Online; accessed 11-March-2021].

[72] Pingzapper. Pingzapper Pricing. https://pingzapper.com/plans, 2018. [Online; accessed 11-March-2021].

[73] Enric Pujol, Philipp Richter, Balakrishnan Chandrasekaran, Georgios Smaragdakis, Anja Feldmann, Bruce M. Maggs, and Keung-Chi Ng. Back-office web traffic on the Internet. In *ACM IMC*, 2014.

[74] radiowaves. SHPD8-1011. https://www.radiowaves.com/getmedia/b1a7277f-fde0-4c05-a5fc-7c22c29c5b3a/HPD8-1011.aspx, 2018. [Online; accessed 28-July-2021].

[75] radiowaves. SPD8-11. https://www.radiowaves.com/getmedia/f942ec58-9999-4607-a165-fd4db4deef60/SPD8-11.aspx, 2018. [Online; accessed 28-July-2021].

[76] Eduard Sackinger. *Analysis and Design of Transimpedance Amplifiers for Optical Receivers*. John Wiley & Sons, 2017.

[77] William Sentosa, Balakrishnan Chandrasekaran, P. Brighten Godfrey, Haitham Hassanieh, Bruce Maggs, and Ankit Singla. Accelerating mobile applications with parallel high-bandwidth and low-latency channels. In *ACM HotMobile*, 2021.

[78] Michael Sheetz. SpaceX prices Starlink satellite internet service at $99 per month, according to e-mail. https://www.cnbc.com/2020/10/27/spacex-starlink-service-priced-at-99-a-month-public-beta-test-begins.html, 2020. [Online; accessed 11-March-2021].

[79] Shkilko, A. and Sokolov, K. Every Cloud Has a Silver Lining: Fast Trading, Microwave Connectivity and Trading Costs. https://ssrn.com/abstract=2848562, 2016. [Online; accessed 11-March-2021].

[80] SimilarWeb. Overview: amazon.com. https://www.similarweb.com/website/amazon.com/#overview, 2021. [Online; accessed 28-July-2021].

[81] SpaceX Starlink. https://www.spacex.com/webcast, 2017. [Online; accessed 11-March-2021].

[82] Starlink Services. Petition of Starlink Services, LLC for designation as an eligible telecommunications carrier. https://ecfsapi.fcc.gov/file/10203162683 11/Starlink%20Services%20LLC%20Application%20for%20ETC%20Designation.pdf, 2021. [Online; accessed 11-March-2021].

[83] statista. Online gaming - statistics & facts. https://www.statista.com/topics/1551/online-gaming/, 2021. [Online; accessed 28-July-2021].

[84] Internet Live Stats. Google Search Statistics. https://www.internetlivestats.com/google-search-statistics/. [Online; accessed 11-March-2021].

[85] Steam. Steam & game stats, 2017. http://store.steampowered.com/stats/ [Online; accessed 11-March-2021].

[86] Telesat. Telesat: Global Satellite Operators. https://www.telesat.com/, 2020. [Online; accessed 11-March-2021].

[87] Unwired Labs. OpenCelliD Tower Database. https://opencellid.org/, 2018. [Online; accessed 11-March-2021].

[88] USGS. National Elevation Dataset (NED). https://www.usgs.gov/core-science-systems/national-geospatial-program/national-map. [Online; accessed 11-March-2021].

[89] J. H. Winters, J. Salz, and R. D. Gitlin. The impact of antenna diversity on the capacity of wireless communication systems. *IEEE Transactions on Communications*, 42(2/3/4):1740–1751, Feb/Mar/Apr 1994.

[90] X, the moonshot factory. Taara – Expanding global access to fast, affordable internet with beams of light. https://x.company/projects/taara/, 2018. [Online; accessed 11-March-2021].

[91] Xiufeng Xie, Xinyu Zhang, and Shilin Zhu. Accelerating mobile web loading using cellular link information. In *ACM MobiSys*, 2017.

A TOPOLOGY DESIGN

Picking a subset of site-to-site links to connect a set of cities involves solving a typical network design problem. The Steiner-tree problem [41] can be easily reduced to this problem, thereby establishing hardness. Standard approximation algorithms, like linear program relaxation and rounding, yield sub-optimal solutions, which although provably within constant factors of optimal, are insufficient in practice. We develop a simple heuristic, which, by exploiting features specific to our problem setting, obtains nearly optimal solutions.

Inputs: Our network design algorithm requires:

- A set of sites to be interconnected, $v_1, v_2, \ldots, v_n$.
- A traffic matrix H specifying the relative traffic volume $h_{ij} \in [0,1]$ between each pair v_i and v_j.
- The geodesic distance d_{ij} between each v_i and v_j.
- The distance along the shortest, direct MW path between each pair, m_{ij}, as well as its cost, c_{ij}. This is part of the output of step 1.
- The optical fiber distance between each pair, o_{ij}, which we multiply by 1.5 to account for fiber's higher latency.
- A total budget B limiting the maximum number of bidirectional MW links that can be built.

Expected output: The algorithm must decide which direct MW links to pick, i.e., assign values to the corresponding binary decision variables, x_{ij}, such that the total cost of the picked links fits the budget, i.e., $\sum_{ij} x_{ij} c_{ij} \leq B$. Our objective is to minimize, per unit traffic, the mean stretch, i.e., the ratio of latency to c-latency, where c-latency is the speed-of-light travel time between the source and destination of the traffic.

Problem formulation: Expressing such problems in an optimization framework is non-trivial: we need to express our objective in terms of shortest paths in a graph that will itself be the *result*. We use a formulation based on network flows.

Each pair of sites (v_s, v_t) exchanges h_{st} units of flow. To represent flow routing, for each potential link ℓ, we introduce a binary variable $f_{stij,m}$ which is 1 *iff* the $v_s \rightarrow v_t$ flow is carried over the microwave link $v_i \rightarrow v_j$, and a binary variable $f_{stij,o}$ which is 1 *iff* the same flow is carried over the optical link[8] $v_i \rightarrow v_j$. The objective function is:

$$min \sum_{s,t} \frac{h_{st}}{d_{st}} \sum_{i,j} (o_{i,j} f_{stij,o} + m_{i,j} f_{stij,m}) \qquad (3)$$

The h_{st} term achieves our goal of optimizing *per unit traffic*. The $\frac{1}{d_{st}}$ term achieves our goal of optimizing the *stretch*.

[8] A "link" between sites can use multiple physical layer hops, both for MW and fiber. The underlying multi-physical-hop distances are already captured by the inputs o_{ij} and m_{ij} so the optimization views it as a single link.

For brevity, we omit the constraints, which include: flow input and output at sources and sinks; flow conservation; total budget; and the requirement that only links that are built ($x_{ij} = 1$) may carry flow. All variables are binary, so flows are "unsplittable" (carried along a single path) and the overall problem is an integer linear program (ILP).

Note that we have decomposed the problem so that link capacity is *not* a constraint in this formulation: MW links will be built with sufficient capacity in step 3; fiber links are assumed to have plentiful bandwidth at negligible cost relative to MW costs. As a result, the objective function will guide the optimizer to direct each $v_i \to v_j$ flow along the shortest path of built links, which is the direct MW link $v_i \to v_j$ if it happens to be built, or otherwise, a path across some mix of one or more fiber and MW links.

ILP's limited scalability: The exact ILP is not scalable, which is the reason we use multiple heuristics, as discussed in §3. As we show in Fig. 8a, the exact ILP, without using our observations on the problem structure, is too computationally inefficient to scale to this scenario. We use subsets of all 120 cities to assess scalability, with the budget proportional to the number of cities in each test, with a budget of 6,000 towers at the largest scale. Even after 2 days of compute, the exact ILP was unable to obtain a result for sets of cities larger than 50. In contrast, our cISP design heuristic is able to solve the problem at the full scale. Second, as Fig. 8b shows, at small scales, where we can also run the exact ILP, our heuristic yields the optimal result. We also tested a linear program rounding approach, but even the naive LP relaxation followed by rounding did not scale beyond 60 cities, and gave results worse than optimal.

B Routing & Queuing

The HFT industry's point-to-point MW deployments demonstrate end-to-end application layer latencies within 1% of *c*-latency, after accounting for all delays in microwave radios, interfacing with switching equipment and servers, and application stacks. Such low latencies across point-to-point long-distance links place sharp focus on any latencies introduced at routers for switching, queuing, and transmission.

Internet routers can forward packets in a few tens of microseconds, and specialized hardware can hit 100× smaller latencies [49]. Transmitting 1500 B frames at 1 Gbps takes 12 μs. Thus forwarding and transmission even across many long-distance links incur negligible latency. Longer routes and queuing delays, however, can have substantial impact.

To assess the impact of routing and queuing in cISP, we use ns-3 [70]. We use UDP traffic with a uniform packet size of 500 bytes. We use the built-in FlowMonitor [19] to measure delay and loss rate, and add a new monitoring module to track link-level utilization. All experiments simulate 100 Gbps of network traffic for one second of simulated time. An experiment takes approximately 10 hours to complete on a single core of a 3.1 GHz processor. Even achieving this running time requires some compromises: we aggregate the bandwidth of parallel links and remove the individual tower hops to focus on network links between the routing sites.

Routing schemes: Besides ns-3's default shortest path routing, we implement two other schemes – throughput optimal routing, and routing that minimizes the maximum link utilization, a scheme commonly employed by ISPs [51].

Results: When the traffic and routing match the design target, i.e., the population-product traffic routed over shortest paths, we find that the network can be driven to high utilization (95%) with near-zero queuing and loss. Non-shortest-path routing schemes needlessly compromise on latency in such scenarios. (Plots for this easy scenario are omitted.)

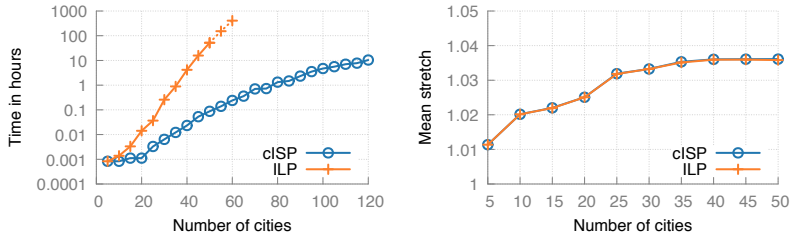
We also test the network's behavior under deviations from the designed-for traffic model. We emulate scenarios where a city produces more or less traffic than expected by allowing, for each city, a "population perturbation" — each city's population is re-weighted by a factor drawn from the uniform distribution $U[1-\gamma, 1+\gamma]$ for a chosen $\gamma \in [0,1]$.

Fig. 9a and Fig. 9b show the results for $\gamma \in \{0.1, 0.3, 0.5\}$. Even for large perturbations, the mean delay does not increase by more than 0.1 ms and the loss rate is zero up to an aggregate load of 70% of the capacity designed for, even with just shortest path routing. Other routing schemes are indeed more resilient to higher load, achieving virtually zero loss and queuing delay even at high utilization, but at the cost of latency. For the tested topology, both the alternative routing schemes incur 10% higher latency on average (not shown in the plots). These results indicate there would be significant value in work that reduces the amount of over-provisioning required by making modest compromises on latency on some routes, e.g., as in [43].

Speed mismatch: The bandwidth disparity between the network core and edge for cISP may seem atypical, in the sense that in most settings, the core has higher bandwidth links compared to the edge, while in cISP, edge links (such as those at large data center end points) may often have much higher line rates when they feed their outgoing traffic into cISP. Thus, we also evaluate if this "speed mismatch" causes persistent congestion at cISP's ingresses.

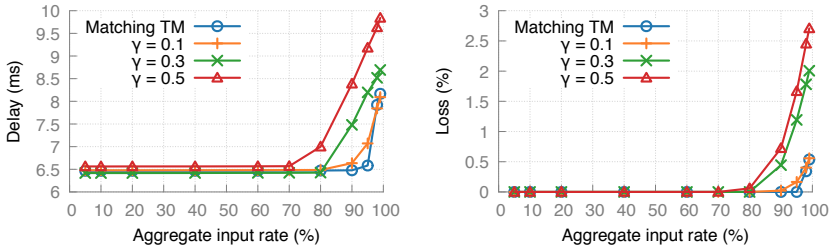
We run ns-3 simulations with several sources (S_i) connected to a sink (D) through the same intermediate node (M). The M-D link rate is fixed at 100 Mbps. We then evaluate settings with every S_i-M link being either 100 Mbps or 10 Gbps. The former is the control, and the latter is the setting with a speed mismatch. M has an unbounded queue. Ten sources send 100 KB TCP flows (small, as is expected in cISP) to the sink, D. The arrival of these TCP flows follows a Poisson process, consuming on average 70% of the I-D link's bandwidth. Each simulation run lasts 10 s and we conduct 100 such runs. We test TCP both with and without pacing.

Fig. 10a shows that the median queue occupancy at M is higher without pacing, especially at the 95$^{\text{th}}$ percentile. However with pacing, queueing behavior is nearly the same.



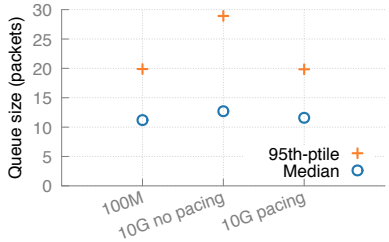
(a) **(b)**

Fig. 8: *cISP's design method is fast-enough and near-optimal: (a) cISP generates an optimized topology within hours for* 120 *cities while the ILP does not yield a result even after* 2 *days for more than* 50 *cities. For the ILP, runtimes for* 50+ *cities are extrapolated by curve fitting. (b) The stretch achieved by cISP matches that of the ILP to two decimal places for instances that can be optimized by the ILP.*

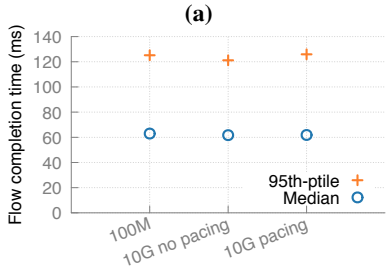


(a) **(b)**

Fig. 9: *(a) Average delay and (b) loss rate remain consistent across perturbations of the city-city traffic model, except under heavy load.*



(a)



(b)

Fig. 10: *TCP pacing addresses the problem of capacity mismatch (a) by reducing persistent queuing (b) without affecting flow completion times.*

The median flow completion times (Fig. 10b) are unaffected both with and without pacing.

C Further Design Considerations

C.1 Is the city-city traffic model special?

Ideally, we would be able to use wide-area traffic matrices from some ISP or content provider for modeling. In the absence of such data, we focus on showing that cISP can be tailored to vastly different deployment scenarios and their corresponding traffic models. Apart from the city-city population product model, we use (a) traffic between a provider's data centers; and (b) traffic between the cities and data centers.

An inter data center cISP: We use Google data centers as an example, considering all 6 publicly available US locations - Berkeley, SC; Council Bluffs, IA; Douglas County, GA; Lenoir, NC; Mayes County, OK; and The Dalles, OR. In the absence of known inter-data center traffic characteristics, we provision equal capacity between each DC-pair.

Data centers to the edge: We also model a scenario where data centers are to be connected to edge locations in cities. Each of the 120 cities connects to its closest Google data center, with traffic proportional to its population.

We show in Fig. 11 that using the same design approach as in §3, both of the above scenarios result in networks with lower cost than the city-city model. Thus, cISP can be tailored to a variety of use cases and traffic models.

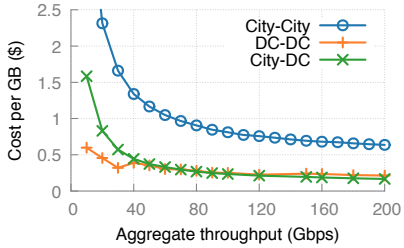


Fig. 11: *Cost per GB for different traffic models: the City-City model, discussed in the most detail, is the most expensive.*

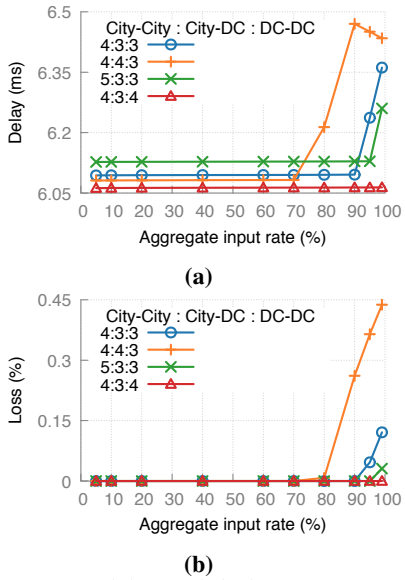


Fig. 12: *(a) Average delay and (b) loss rate remain consistent across deviations from the designed-for traffic mix, except under heavy load.*

C.2 Traffic model mismatches

A cISP may carry a mix of city-city, inter-DC, and DC-edge traffic. How does its performance degrade as the *proportion* of these traffic types departs from the design assumptions?

We design a cISP to carry an aggregate of 100 Gbps with a city-city : DC-edge : inter-DC traffic proportion of 4:3:3. Using ns-3 simulations similar to those in §B, we then test this network under several traffic mixes different from this designed-for mix — 5:3:3, 4:3:4, and 4:4:3.

Fig. 12a and Fig. 12b show that there is a difference of less than 0.05ms in mean delay across different combinations of traffic matrices up to an aggregate load of 70% of the design capacity. Similarly, loss remains nearly 0 until this load. The decrease in delay at high load (4:4:3 for $x > 90$ in Fig. 12b) is due to losses, which are likelier on longer, higher-delay paths.

Mean delay depends more on city-city traffic, as expected: city-city traffic requires a wider infrastructure footprint, and deviations from its design parameters have greater impact.

Thus, as discussed in §B, significant traffic model deviations can be absorbed using some over-provisioning, in line with current ISP practices.

C.3 Is the US geography special?

It is reasonable to ask: are the population distribution and geography of the U.S. especially amenable to this approach, or is it applicable more broadly? The availability of high-quality tower data and geographical information systems data for the U.S. enables a thorough analysis. While similar data is, unfortunately, not available to us for other geographies, we can approximately assess the design of a cISP in Europe using public, crowd-sourced data on cellular towers [87]. Lacking fiber conduit data, we assume that fiber distances between cities are inflated over geodesic distance in the same way as in the US (∼1.9×). Using our methodology in §3, we design a European cISP of similar geographical scale across cities with population more than 300k, targeting the same aggregate capacity and mean latency (1.04× here vs. 1.05× for cISP-US). The cost of this design, shown in Fig. 13, is similar as well, with ∼3k towers. Note that the impact of Europe's higher population density is not seen here, because we explicitly design for the same aggregate throughput. One could, alternatively, normalize throughput per capita, and compare cost per capita, to obtain similar results.

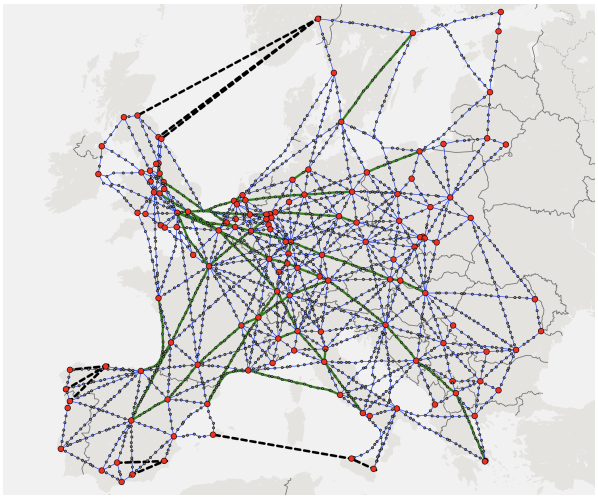


Fig. 13: *A 100 Gbps* 1.04× *stretch cISP across Europe. This network uses several fiber connections (dashed, black lines).*

Admittedly, there is not yet a known approach to bridging large transoceanic distances using MW, limiting our approach to large contiguous land masses that need to be interconnected with fiber. In the distant future, LEO satellite links, hollow-core fiber, or even towers on floating platforms may be of use for such connectivity.

D American Tower Deployment

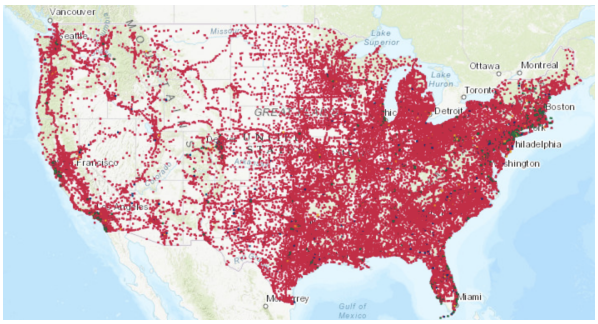


Fig. 14: *American tower deployment as per* 5^{th} *March,* 2021.

American Tower [7] claims to have a presence at more than 42,000 tower sites across the US, as of 5^{th} March 2021. Fig. 14 shows their current deployment. We could not access their database due to legal bindings.