# Divided at the Edge - Measuring Performance and the Digital Divide of Cloud Edge Data Centers

NOAH MARTIN, Tufts University, USA
FAHAD DOGAR, Tufts University, USA

Cloud providers are highly incentivized to reduce latency. One way they do this is by locating data centers as close to users as possible. These "cloud edge" data centers are placed in metropolitan areas and enable edge computing for residents of these cities. Therefore, which cities are selected to host edge data centers determines who has the fastest access to applications requiring edge compute – creating a digital divide between those closest and furthest from the edge. In this study we measure latency to the current and predicted cloud edge of three major cloud providers around the world. Our measurements use the RIPE Atlas platform targeting cloud regions, AWS Local Zones, and network optimization services that minimize the path to the cloud edge. An analysis of the digital divide shows rising inequality as the relative difference between users closest and farthest from cloud compute increases. We also find this inequality unfairly affects lower income census tracts in the US. This result is extended globally using remotely sensed night time lights as a proxy for wealth. Finally, we demonstrate that low earth orbit satellite internet can help to close this digital divide and provide more fair access to the cloud edge.

CCS Concepts: • **Networks** → *Network measurement*.

Additional Key Words and Phrases: networks, datacenter, edge, digital divide, measurement

**ACM Reference Format:**
Noah Martin and Fahad Dogar. 2023. Divided at the Edge - Measuring Performance and the Digital Divide of Cloud Edge Data Centers. *Proc. ACM Netw.* 1, CoNEXT3, Article 16 (December 2023), 23 pages. https://doi.org/10.1145/3629138

## 1 INTRODUCTION

Latency reduction is a top priority for internet applications. Amazon's frequently-referenced experiment [3] demonstrated a 1% loss in sales for every 100ms of increased latency, with similar trends observed in other experiments as well [2, 16]. Not only does reducing latency lead to more profit for service providers, but it also enables new, emerging applications with strict latency constraints. For example, head-tracking applications, such as virtual reality and 360° video streaming, require a response time of under 20ms to avert motion sickness [75]. Many machine learning (ML) inference applications (e.g., speech and image recognition, large language models (LLMs), etc.) are also latency-sensitive [38]. For instance, processing pipelines for real time video analytics may run at 30 frames per second (FPS) [80]. In these applications, any additional milliseconds saved in latency can enable the use of more powerful ML inference models for improved decision making [53, 95]. Similarly, large LLMs can benefit from additional latency budget to improve accuracy or reduce energy consumption [90, 93]. These ML-based services are increasingly found on the critical path for applications such as healthcare, search, translation and more [71, 77, 92, 97], benefiting both

Authors' addresses: Noah Martin, Tufts University, Boston, MA, USA, noah.martin@tufts.edu; Fahad Dogar, Tufts University, Boston, MA, USA, fahad@cs.tufts.edu.


2834-5509/2023/12-ART16
https://doi.org/10.1145/3629138

the developed countries but also increasingly more important for the developing parts of the world. A common theme in these applications is that they are typically powered by the cloud.

To lower latency for these emerging applications, cloud providers have started offering "edge locations" which are even closer to the end users compared to traditional data centers. This typically entails offering compute (i.e., virtual machines - VMs) at the edges of their private cloud network, the closest they can get to end users. From 2019 to 2023 Amazon has launched over 30 of these edge locations, named Local Zones, around the world, with many more announced [21]. Other cloud providers have announced plans for similar services [8, 15]. These examples of the "cloud edge" are one version of the growing trend towards edge computing [85] that is increasingly becoming popular in industry [20] and the research community [86]. This trend is also supported by the rise of serverless computing [70] to combine the benefits of the edge with the scalability and ease of use of functions as a service [1, 10, 18]. Finally, this cloud edge presents an opportunity to support many traditional networking solutions, including support for content and service centric networking [55–57, 64, 96].

While cloud providers are working to reduce latency, there remains a "digital divide" [11] separating those with quality Internet access from those without. Internet access is known to affect many aspects of a community's socio-economic well-being [17, 49]. It has even been recognized as a key prerequisite for many of the sustainable development goals (SDGs) [29]. A variety of factors contribute to the digital divide, including reduced availability of ISPs [59, 63] and lack of digital literacy [37]. Satellite internet has been able to address the technical limitations of providing access in some of these underserved populations such as remote islands, rural areas, and underdeveloped countries [82, 89], but issues beyond access remain.

Prior work has measured the digital divide's effect on network download speed [81], but there has been no systematic study of its impact on minimum cloud latency, which is becoming an increasingly important factor in determining user experience. Distance between users and the geographic locations of cloud data centers has been observed to be the most important factor in determining latency [51]; therefore, we focus our study of the digital divide on *minimum cloud distance* - which we term "cloud digital divide" (CDD). The data center locations, including ones used for cloud edge, are under control of the major cloud providers. Some important questions regarding cloud edge locations and their impact on the digital divide include: Are the cloud providers considering the CDD in choosing these edge locations? Would the CDD increase or decrease with these cloud locations? What can be done to potentially reduce the CDD?

In this paper, we attempt to answer these questions – we use globally distributed probes to measure the improvements possible with the cloud edge, and combine measures of economic well-being with edge data center locations to analyze the extent of the digital divide with respect to cloud latency. To the best of our knowledge, we are the first to measure latency reductions from commercial cloud edge on all continents and to study how the deployment of this cloud edge affects the digital divide in cloud infrastructure. As part of our study, we make three primary contributions:

First, we perform a measurement study on the widely used measurement platform RIPE Atlas [28] to quantify the reduction in latency due to the cloud edge when compared to the traditional cloud regions. To target cloud edge locations, we initially choose the Local Zone product of AWS, and then broaden our results by using cloud routing optimization services to estimate edge latency for three major cloud providers if they were to offer compute at the edges of their network. Our results in §4 show up to 28% latency reduction at the 80th percentile in North America. Additionally, we show that cloud edge enables *greater than 80% of probes in North America, Oceania, and Europe to achieve under 20ms latency (which is required for head-tracking).* The latency reduction in under-provisioned continents was even higher, significantly improving the minimum cloud latency for users in these

areas. However, there is still a large gap between these continents and more developed regions even when using the cloud edge.

Second, in §5, we demonstrate how cloud edge affects the cloud digital divide by considering two metrics – inequality and unfairness – which are inspired by similar metrics used in the development economics field [41, 47]. We observe that the *inequality* (which we define as the ratio between those furthest and those closest to data centers) doubles for many continents with the introduction of the cloud edge. Oceania, in particular, has the top 90% of users over 200x further from a data center than the bottom 10%. We then analyze the *unfairness* in data center distances, quantifying the extent to which cities hosting cloud edge locations tend to have higher amounts of wealth. This unfairly provides access to applications requiring latency under a low threshold to wealthier populations. Despite existing biases, our results indicate decreasing unfairness as more cloud edge compute is launched, albeit at a slow pace. Our analysis uses economic indicators including census data [33] and night time lights [34]. We offer suggestions for how to select locations which optimize for fairness while reaching the most users.

Lastly, we zoom in on the latency and digital divide when considering ISPs using low Earth orbit (LEO) satellites (e.g., Starlink). We hypothesize that, just like it has improved global access to the Internet (discussed in §2.2), it may be a promising technology to improve inequality in access to the nearest cloud as well. This is based on intuition that delay of the satellite hop dominates the end-to-end delay and is fairly homogeneous across the globe. Our case study in §6 focuses on answering the following question: Can LEO Internet reduce the CDD? Through measurements with RIPE Atlas probes using Starlink connectivity, we first show their feasibility in providing low latency cloud access suitable for tasks requiring the edge. Next, we use a satellite network simulator [73] to measure RTTs across the globe, and find that the differences between the top and bottom 10% of users drops to lower than 10x. By widening the area in which low latency applications can reach a data center, unfairness stemming from the selection of cities also greatly declines.

These results demonstrate that the digital divide, previously observed to disadvantage communities without Internet access, now separates communities by an increasingly important metric - minimum cloud latency. As reliance on the cloud increases for everyday tasks, and applications such as AR [35, 36], remote work [25], and video analytics [38] require lower latency, users on one side of this "cloud digital divide" may be left behind. We believe our work is a first step in highlighting this issue, and provides guidance on promising technologies and deployment paths that could help in reducing this divide. To facilitate reproducibility and follow-up work, we have made our code available at https://github.com/TuftsNATLab/EdgeDivide

## 2 BACKGROUND

In this section we provide background on the technologies and concepts that are the focus of this paper. This includes cloud data centers and their private WANs as well as satellite ISPs.

### 2.1 Cloud Networks

Multiple cloud providers operate large private WANs to connect their regions directly to access ISPs and bypass the public Internet. The extent of these private WANs allow cloud providers to reach over 76% of the Internet without using Tier-1 or 2 ISPs [40]. Due to this "flattening", more and more of the network path between users and cloud resources is falling under the cloud provider's ownership.

These private WANs are known to not have a large effect on latency within continents that have a well provisioned public internet infrastructure [54]. However, they do offer quality advantages over long distances [65–67]. Cloud providers offer this benefit as a commercial product. For example,

AWS Global Accelerator [5] routes ingress through the closest AWS point of presence (PoP) [39]. The IP addresses of Global Accelerator are anycast from the AWS edge locations, similar to strategies used by CDNs [6, 45]. Anycast addressing chooses the best endpoint based on metrics such as number of BGP hops. This can be sub-optimal, but prior measurement studies have shown that anycast CDNs have minimal latency inflation [74]. To further improve quality, Global Accelerator terminates TCP connections at an edge location and proxies them over the private WAN [7]. Throughout this paper, we refer to "edge locations" as any area where the cloud provider has a PoP.

To reduce latency beyond the limits imposed by the physical distances of routing, cloud providers have begun pushing data centers towards the network edge. These new cloud edge data centers are smaller than traditional regions, and offer a subset of services with different pricing rates. Local Zones (LZs) are an AWS edge compute product that has seen significant deployment in recent years [21]. Each LZ is typically a new physical location "owned, managed, and operated by AWS"[19] as opposed to a co-location. LZs are deployed in a particular metropolitan area and connected to a parent AWS region over the private WAN. An example of this deployment is shown in Fig 1. LZ locations (as well as regions) are a subset of edge locations. When using a nearby LZ, data does not need to traverse a long distance in the private WAN, instead compute resources are located at the edge. Requests reach their destination shortly after entering the AWS network. Hybrid workloads can still access the region from the LZ for lower costs or greater resources. Each LZ launch is accompanied by a blog post, which we use to determine the date they became available. The Chicago/NYC post explains these data centers were meant for finance companies to have low latency compute [26], which suggests motivations for choosing cities that might not take the digital divide into account.

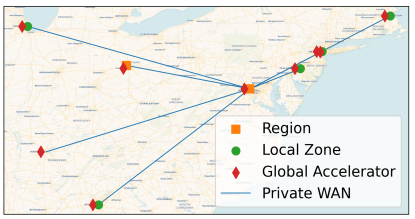



Fig. 1. Amazon's eastern US services include LZs in major cities, linked to regions via a private WAN and located near Global Accelerator-hosting PoPs. Users may be nearest to a Local Zone, a region, or a non-computing PoP. The WAN connectivity is illustrative, not showing actual paths.

Moving compute closer to users is an effective way of reducing latency because distance to data centers has been observed to be the most important factor in determining latency [51]. Other factors causing latency inflation include long fiber distances, inefficient "hairpinning" routes, and queuing [88]. Addressing these issues can decrease latencies to the cloud, but physical distance will still create an inherent divide and determine what applications or optimizations (eg. more complex ML inference) are possible. For these reasons we focus on the distances in our work.

### 2.2 LEO Satellites

Satellite internet is already used to help bridge the digital divide and bring Internet access to people in regions with no ground-based ICT infrastructure [89]. In some places, particularly in African countries where home internet is less prevalent, it is used to offer Internet access to many users at once in public places [82]. A common limitations comes from geosynchronous satellites orbiting at over 35,000 km from Earth and therefore incurring hundreds of milliseconds RTT overhead [73]. Low Earth Orbit (LEO) satellites such as Starlink orbit less than 600 km from Earth and therefore can support Internet with much lower latency [72]. Due to their low orbits, LEO satellites must operate in constellations with thousands of satellites to provide full ground coverage [72]. Many current deployments use the "bent pipe" model which forms a path from a customer satellite dish to one satellite and back to a ground station (GS) connected to the Internet over the terrestrial network. The satellite must be in view of both the customer and the GS. Inter-satellite links (ISLs) create paths through multiple satellites to remove this constraint. While LEO constellations may

have high potential, a number of challenges still exist including scalability, coverage, variability, and affordability [50, 72, 89].

## 3 METHODOLOGY

### 3.1 Latency Measurements

Our latency measurements use RIPE Atlas [28], a global collection of volunteer-hosted Internet probes and de facto standard for Internet measurements [51, 52, 60, 67]. Probes are installed in a variety of networks, including some that are not representative of typical internet users [42]. We exclude probes with the user-provided tag *datacenter* and those in an AS belonging to AWS (16509/14618), Azure (8075) or Google (396982). This does not guarantee exclusion of probes hosted in privileged locations, but helps keep the results more representative of end user latency. In total over 4.8k probes were selected for the study. We ran measurements in May 2023 and group our results into summaries for each continent, as well as a closer look at the US–the focus of Amazon's initial edge compute rollout. Each measurement is configured to send 3 packets per `ping`/hop. Unless otherwise noted, the measurements run 4 times every 4 hours for 4 days. Results for these experiments are presented in §4.

**US Local Zones:** In the continental US, we run `ping` to t3-medium EC2 instances in all 16 Local Zones and four regions - resulting in over 7.8 million data points.

**Global Baseline:** We launch VMs in every region for each of our target cloud providers: t3-medium in AWS, Standard_B1ls in Azure, and e2-micro in GCP. For each probe we determine the three lowest latency regions using 10 pings to all 100 VMs. Then, each probe measures the min RTT to these regions over our measurement period to create a baseline for comparison with the edge. In total over 10.3 million measurements were made to regions.

**Cloud Edge Per Continent:** Our edge latency measurements use the anycast IP addresses provided by routing optimization services we launched on each cloud provider (Global Accelerator, Azure Front Door, and Google Global HTTP Load Balancer) to determine the minimum time to reach the private WAN. The results contain over 4 million data points. Many of the locations hosting Global Accelerator have been announced to have LZs in the future, but some have no public plan to offer compute. Therefore, this what-if analysis is a best case scenario to answer the question: What if cloud providers offered compute at all their edge locations? These measurements may be susceptible to known anycast inefficiencies which are not a factor in the baseline, but this would only underestimate the possible latency improvements.

**Global Accelerator Traceroute:** Lastly, we quantify how much time is spent in the AWS network before reaching our cloud edge endpoint. We run `traceroute` 4 times for all probes to AWS Global Accelerator every 3 hours for 4 days. Resulting in over a million additional measurements. For each hop we record the minimum RTT and lookup the ASN to determine if it is owned by Amazon. The difference in RTT between the last hop and first Amazon hop is the time spent in the AWS network. We also tried using the first AWS hop directly for RTT measurements, but found these addresses do not respond to `ping`. The RTT values for each hop may not be monotonically increasing because each measurement is independent and can vary due to changes such as noise or different forward/reverse paths - resulting in negative time spent in the AWS network. This is the same problem faced in prior studies of cloud connectivity [52]. Since we use only the minimum RTT and a large number of measurements, the effects of noise are limited. We are left with <10% of probes that see significantly higher RTTs to the first AWS hop than the destination.

In all cases we report the minimum RTT for each probe to reach the cloud service we are measuring. All our experiments use ICMP. Prior work has suggested ICMP could see higher latency than TCP due to prioritization [54]. However, prior work on cloud region latency primarily used

ICMP for latency [46, 51, 54] and the prioritization would only under-estimate the edge performance by providing a worst-case latency. We believe this is a fair metric to report on the improvements possible from cloud edge.

### 3.2 Cloud Digital Divide

Our study of the CDD uses two metrics: inequality, $I$, and cloud access indicator, $CAI$. For this analysis, we only consider the AWS cloud. We calculate each metric for the US as well as per-continent. We consider how the CDD takes shape both for regional data centers only, and when compute is offered at the edge of the network based on announced AWS edge locations. For Local Zones that are already available, we show the change in CDD that resulted from each launch. Here we define our metrics and list the datasets used to compute them.

***I*:** Our inequality metric, $I$, is the $p90$ to $p10$ data center distance ratio, demonstrating the cloud edge's benefit to the fastest 10% ($p10$) versus the slowest 10% ($p90$). This metric is commonly used in economics to study income [9, 91] and can be intuitively extended to data center distances. It is a relative metric, so even if the $p90$ and $p10$ are close in value $I$ can be high. This property captures the advantage being close to a data center provides emerging applications such as machine learning at the edge where extra milliseconds makes room for more accurate models. We also present the difference in latency in Appendix (A.1) to paint a full picture.

***CAI*:** We define Cloud Access Indicator ($CAI$) to quantify cloud availability to a population. Unlike $I$, this is based on a threshold, $\sigma$. Data centers closer than the threshold are considered equally useful. This is relevant in cases where an application works equally well with any latency under a fixed threshold. The benefits of multiple available data centers, such as capacity and resilience [22], are also captured in our metric. For a population ($p$) and set of data center locations ($L$) $CAI$ is:

$$CAI(p) = \sum_{l \in L} reachable(p, l) \tag{1}$$

$$reachable(p, l) = \begin{cases} 1 & \text{distance(p, } l\text{) <= } \sigma \\ 0 & \text{distance(p, } l\text{) > } \sigma \end{cases} \tag{2}$$

Where the *distance* function determines the kilometers between a data center location and a population. In §5.2 we use $CAI$ to demonstrate how populations are unfairly disadvantaged based on their economic status.

**US census data** from the American Community Survey was downloaded from ESRI Updated Demographics [33] using ArcGIS Online [4]. US measurements are all at the census tract level and use population as well as median income. Coordinates for each tract is from the US census TIGER FTP archive for 2020 [32] and used to calculate distance from each tract to data centers.

**Administrative 2 level boundaries** define the level for all our global measurements and were downloaded from GADM [13]. Some countries do not have boundaries available, so these were excluded from our analysis.

**Global population** is from the Gridded Population of the World, v4 2020 at 15 arc-minute resolution [61].

**Remotely sensed night time lights** (NTL) have been shown to be a suitable global proxy for economic well-being and not subject to embellishment for political purposes [27, 78, 83]. We retrieve NTL rasters from Annual VNL V2.1 [34] for the year 2020. This is a composite of monthly NTL datasets that removes clouds, sunlight, moonlight, and other outliers. Mean NTL per administrative 2 unit is our proxy for the area's wealth, based on findings in [78].

### 3.3 LEO Satellite ISP

We evaluate how LEO satellites ISPs affect the CDD using the satellite constellation simulator, Hypatia, introduced at IMC 2020 [73]. Our simulations use the same configuration as the original paper: Starlink 5-shell with ISLs. We consider ground stations at each AWS edge location, and simulate RTT from each administrative region to the three nearest regions and edge locations. The RTTs and population per administrative region are used to calculate $I$ with and without cloud edge on each continent over a LEO satellite ISP.

## 4 CLOUD EDGE LATENCY

In this section, we present results on a RIPE Atlas measurement study to answer two questions about cloud edge latency. First, what is the latency reduction possible from using the already supported cloud edge data centers in the United States (§4.1)? Second, how much faster do we expect cloud latency to be when three of the major cloud providers launch their edge compute platforms globally (§4.2)? We find a 1.6x improvement in probes that can currently reach Amazon's cloud in under 20ms. Our what-if analysis of global cloud edge latency estimates a 20% to 60% improvement per continent if cloud providers offer compute at the edge of their private WAN globally. In §4.3 we find the what-if analysis to be accurate for ~95% of probes in each continent.

### 4.1 Current U.S. Cloud Edge Latency

Only AWS had made their cloud edge service publicly available at the time of our experiments, and their Local Zones were most widely deployed in the U.S. To measure expected improvements for current applications that can use cloud edge, we compare AWS regions only vs. all AWS data centers.

Fig 2 shows the latency improvement measured with the methodology described in 3.1. Over 85% (greater than 1.6x more than with regions only) of probes can reach cloud compute in under the 20ms head-tracking threshold when Local Zones are used in addition to regions. This result demonstrates the large potential of cloud edge compute to reduce latency when applications can be distributed across multiple locations.

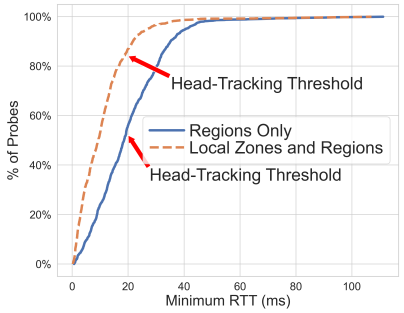


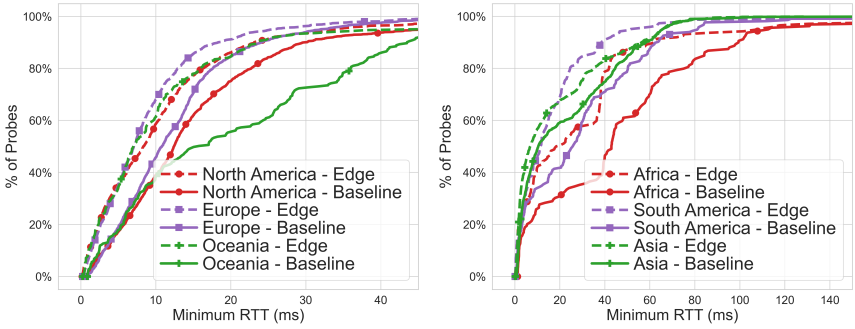

Fig. 2. Reductions in RIPE atlas probe to cloud minimum latency when using regions + Local Zones vs. regions only. When measurement endpoints include Local Zones, 1.6x more probes could reach the cloud in under 20ms.

When each of the three major cloud providers are considered, minimum achievable latency is even further reduced, but requires a new measurement method because launching VMs in new edge data centers was not yet publicly available for Google Cloud or Azure.

### 4.2 What-If Analysis of Cloud Edge Latency Per Continent

Using the endpoints launched in routing optimization services we find the minimum latency to cloud edge in every continent and compare with a region only baseline that consists of two rounds of measurements. The initial set to all regions, and additional measurements to the three with lowest latency. Only 5% of probes had lowest latency to the third data center after the second round. Fig 3 compares minimum edge latency to the region-only baseline. Every continent has significantly lower latency to the cloud edge than data center regions, and a few trends stand out.

First, all three continents in Fig 3a can reach the 20ms threshold for greater than 80% of probes, which was only true for Europe in the baseline. These continents - North America, Europe, and




(a) RTT improvements with cloud edge in North America, Europe, and Oceania.

(b) RTT improvements with cloud edge in Africa, South America, and Asia.

Fig. 3. Cloud edge vs. cloud regions for each continent. Fig 3b has a larger x-axis range than Fig 3a.

| | p80 - Baseline | p80 - Edge | Change | p50 - Baseline | p50 - Edge | Change |
|---|---|---|---|---|---|---|
| NA | 22.72 | 16.21 | 6.51 (28.65%) | 12.56 | 8.31 | 4.25 (33.84%) |
| EU | 17.64 | 12.9 | 4.74 (26.87%) | 10.92 | 7.04 | 3.88 (35.53%) |
| OC | 36.01 | 15.98 | 20.03 (55.62%) | 16.05 | 6.94 | 9.11 (56.76%) |
| AS | 44.65 | 34.79 | 9.86 (22.08%) | 10.65 | 6.78 | 3.87 (36.34%) |
| SA | 50.18 | 24.47 | 25.71 (51.24%) | 25.5 | 11.37 | 14.13 (55.41%) |
| AF | 72.29 | 41.56 | 30.73 (42.51%) | 43.09 | 17.76 | 25.33 (58.78%) |

Table 1. Summary of speedups from using cloud edge on each continent. All measurements are in milliseconds.

Oceania, have sufficient cloud edge infrastructure to support new applications that require the cloud edge for many users. However, as we'll see in §5.1.2, there can still be high proportions of the population without a nearby cloud edge.

Second, Oceania is particularly notable for the dramatic improvement. New Zealand did not have a data center in the baseline, and inter-country latency was high. We ran `traceroute` from New Zealand probes to regions in Australia and observed relatively long latencies which would require traversing an undersea cable. The cloud edge does include data centers within the island, and sub 20ms latency is achievable.

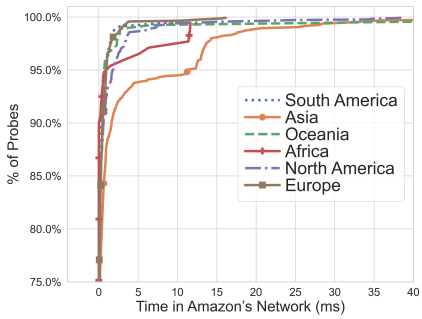
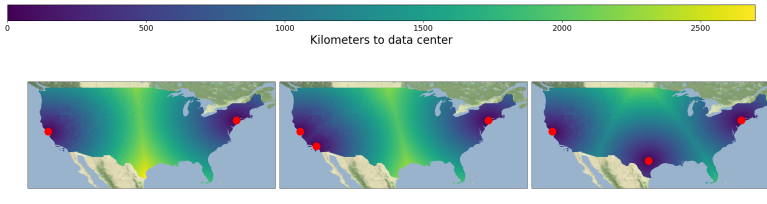



Fig. 4. CDF (top 25%) of latency difference between the first Amazon hop and Global Accelerator. Long tails indicate regions where our method does not capture the lowest possible latency, implying further latency reductions in a completed cloud edge deployment.

Lastly, we see in Fig 3b that the other three continents - Asia, South America, and Africa, see even higher improvements, but also a large gap between the fastest and slowest probes. Each of these continents have a greater baseline p80 latency than any of the other three continents, and Fig 3b has a larger x-axis range to reflect this. Using the cloud edge makes the median more comparable to the other three continents. In fact, it is consistently under the 20ms threshold. However, the 80th percentile is still significantly slower and cannot support emerging edge applications. Africa's CDF stands out as a bi-modal shape, with about 50% reaching the cloud edge in 20ms and another 40% taking >40ms. This indicates a still under-provisioned region that has no nearby cloud edge but high population. We'll return to this idea in §5.1 when demonstrating the unequal access between the top and bottom percentiles. Detailed results for each continent's median and p80, as well as absolute and relative differences between edge and regions, are shown in Table 1.




(a) Initial (b) Los Angeles, CA (c) Austin, TX

| Cities | p10 | p90 | $I$ |
|---|---|---|---|
| Initial | 114 | 1974 | 17 |
| L.A. | 51 | 1906 | 37 |
| Austin | 97 | 1325 | 13 |

Fig. 5. Toy example of different metrics for inequality. 5a starts with a data center in San Francisco and New York City where $I = 17.3$. 5b adds a location in Los Angeles - where a data center was already nearby - and more than halves the p10 as well as increases $I$. 5c adds a location in Austin, TX which greatly reduces the distance to a data center for users who were already far, reducing the inequality metric, $I$.

### 4.3 Private WAN Extent

Finally, we look for exceptions to our assumption that AWS's Global Accelerator [5] can be used to measure the cloud edge. While all Global Accelerator servers are within Amazon's network and are meant to minimize latency, there still could be other factors preventing the Global Accelerator endpoint from being the closest AWS infrastructure to a probe. For example, cloud PoPs may exist in an area to extend the private WAN but not offer Global Accelerator yet due to an incomplete deployment. In this case, `traceroute` between probes and Global Accelerator would contain extra hops not necessary if compute was offered at the private WAN edge. These cases are a limitation in our what-if analysis of the cloud edge, and cause an under-estimation of the speedups.

Fig 4 plots CDFs of the time spent in Amazon's network for probes on each continent. Only ~5% of each continent show a significant time in the private WAN, with the exception of Asia's ~10%. Note the y-axis is clipped to the top 25% to zoom in on the relevant area. As discussed in 3.1, less than 10% were significantly negative (under -1ms). This can be caused by asymmetric paths, when earlier hops do not reduce RTT even though they appear closer on the forward path. Plotting the probes with high RTT in the AWS network geographically (not shown for brevity) reveals a few clusters, including the Philippines in Asia, where the private WAN exists but there are no Global Accelerator endpoints. There is a planned LZ in Manila, but latency from nearby probes were consistent with the distance to the Hong Kong region. When the new LZ is available, we expect these probes to have lower latency than predicted by our what-if analysis. A few of the largest results in the long tail were due to probes in Georgia and Kazakhstan with under 5ms RTT to the first AWS hop, but 30-100ms to the final hop. This also indicates the private WAN is near these probes, but covers a large area to the closest Global Accelerator location. Despite these instances, we don't expect this to influence the main results as it is isolated to a few locations with low density of probes.

## 5 CLOUD DIGITAL DIVIDE

With the deployment of cloud edge, we already measured a divide between continents having fast cloud access such as North America, Europe, and Oceania and those with significantly higher RTT (§4). In this section, we analyze the prevalence of a cloud digital divide within continents from two perspectives, using demographic data rather than active probing (such as RIPE Atlas in §4).

First, by treating edge data centers as general purpose compute that can be used from anywhere, we show the existence of a widening gap between those who are closest to a data center and those who are furthest. This is the typical way cloud regions are used, but LZs are marketed as a service to be used from the same city as the data center. However, we believe this is not a fundamental

technical limitation, but a common current use case due to the higher costs of Local Zones. We show in §5.1 the inequality in data center access is exacerbated by the expansion of edge locations.

Second, we juxtapose data center distance with economic indicators to show an unfairness. This analysis uses the fact that some applications require low latency from edge data centers in the same city, which we show in §5.2 to typically be cities with higher amounts of wealth. Despite observing unfairness, we also find the amount of unfairness is gradually decreasing as cloud edge is expanded to more regions.

### 5.1 Inequality in the Edge

Inequality in cloud access occurs when there are significant differences in minimum latency to data centers. The most important factor in determining minimum latency is typically distance [54], and reducing distance is the primary advantage of deploying new edge data centers.

Fig 5 presents a toy example illustrating our method of calculating inequality, using data center locations in the US. Fig 5a shows distances to the initial data center locations in San Francisco and New York City. Fig 5b adds a new data center in Los Angeles. This location increases inequality by reducing distances for locations that were already relatively close to a data center ($p10$), but not for the $p90$. Fig 5c decreases inequality by placing a new data center in Austin - decreasing distances for the furthest population.

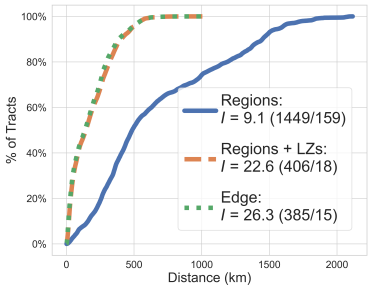



Fig. 6. CDF of minimum distance to regions, regions or Local Zones, and all AWS edge locations, including ones with no compute.

*5.1.1 US Inequality.* As more Amazon edge locations have been added in the US, inequality has been rising. Fig 6 demonstrates this with the min-distance CDF and $I$ (3.2) for different data center distributions. With all regions $I = 9.1$. In this case users in the 90th percentile are less than 10x further from a data center than users in the 10th percentile. With LZs and data center regions, that number rises to 22.6. Jacksonville, Florida is the only edge location in the US that does not already have a Local Zone. This city is relatively close to others with data centers such as Miami and Atlanta, explaining the slight increase in inequality to 26.3 that we expect if it gets a data center.

New data centers at the cloud edge can only decrease the minimum cloud distance, but they often launch relatively nearby existing data centers causing the overall inequality to increase. The CDFs in Fig 6 demonstrate this trend. When edge locations are included, the closest users get even closer, but the users furthest from data centers are left on a relatively longer tail. The absolute distance change is still dramatic for the p90, from over 1400km to less than 400km. However, there is a nearly 30x gap.

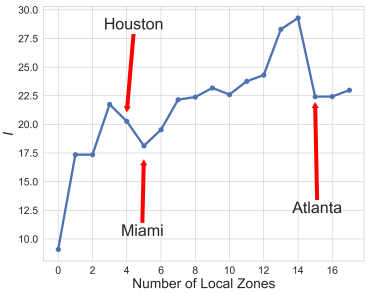



Fig. 7. $I$ as new LZs are launched. The trend is increasing, a few which resulted in sharp decreases are annotated.

Fig 7 illustrates the change in inequality with each new LZ launch – most of these cases increase the $p90/p10$ with some exceptions including Houston, Miami and Atlanta. These cities were much further from existing data centers. The trend points to decreasing latency for an increasing proportion of people while some users are left behind and won't be able to access the same low-latency applications as the rest of the country, widening the CDD.

*5.1.2 World Inequality.* Globally, the deployment of Local Zones is not yet complete, but we use announced locations and existing edge locations to see how inequality will increase as more are launched. Fig 8 shows the change in $I$ for the LZs which are available. For all continents it has increased since the first LZ was introduced. One of the largest jumps was Oceania, which added new locations in Australia and New Zealand. This greatly decreased latency for many in those countries, but did little to help other islands that were already far from AWS regions. There is also a recent slight drop in North America with the first location added in Mexico, helping reduce $I$ by lowering distance for people who were previously located much further away from data centers.

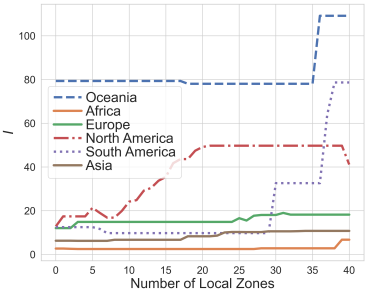



Fig. 8. Change in $I$ per continent as new LZs are launched. For all continents the inequality has increased since the first launch of a LZ.

While the $p90/p10$ looks at differences in the extremes, we can also examine other ratios to see if the increasing trend still applies. Fig 9 plots $p90/p10$ as well as $p80/p20$ for each continent when considering all edge locations (included announced but not launched LZs) and regions only. This gives us an idea of how inequality will change when more Local Zones launch at the edge. Both the $p90/p10$ and $p80/p20$ increase with the addition of edge locations in every continent, which we will see is not the case for satellite networks in §6.

## 5.2 Unfairness in the Edge

While inequalities measure the difference between the closest and furthest populations, unfairness occurs when the furthest populations have this disadvantage due to factors beyond their control. The digital divide has been shown to unfairly affect communities based on attributes including race [63] and poverty [81]. We use a measure of unfairness commonly used for measuring health, the concentration index, to show unfairness in data center locations. This requires defining a metric analogous to those used in health, e.g. life expectancy or number of hospitals. Our metric, cloud access indicator (CAI) (§3.2) captures the availability of edge data centers to serve applications requiring a latency threshold.

*5.2.1 Concentration Index.* Concentration curves are used to assess unfairness in a health metric by demonstrating the relationship between the chosen metric and a socio-economic attribute. The curve is defined as the cumulative percentage of the health metric on the y-axis, and the cumulative percentage of population, ranked by socio-economic attribute (income), on the x-axis [79]. The concentration index (CI) quantifies unfairness and is defined as twice the area between the line of equality (y=x) and the concentration curve [79]. This statistic has been used in previous studies to demonstrate unfairness in resource distribution related to health [41, 47], but to the best of our knowledge has not previously been applied to data center locations.

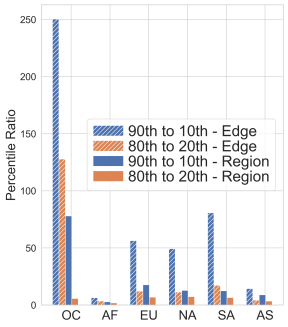
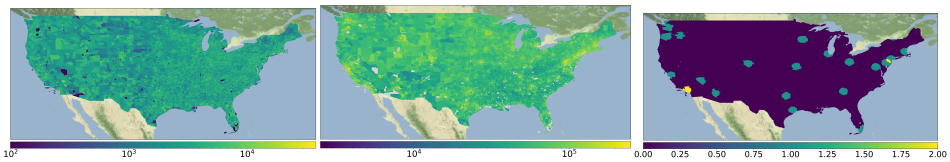



Fig. 9. Inequality increases when compute is expanded to edge locations

By varying the parameter $\sigma$ in CAI, we can measure fairness for different cloud applications. Applications requiring low latency (eg. 20ms for head tracking) will need a lower $\sigma$. Due to the ~300km/ms speed of light, the minimum distance that can support 20ms RTT is 3000km. However, previous work has shown internet speeds when fetching web pages to be 34x (p50) to 169x (p90) inflated over the speed of light [88]. Therefore, we optimistically assume 34x inflation and set $\sigma = 88km$ since applications with strict requirements are likely to be more




(a) Population (b) Median income (c) CAI using all US regions and Local Zones

Fig. 10. variables per census tract used to calculate concentration curves in the U.S.

optimized than the slowest web pages. In §6 we show how satellite networks let us use a larger $\sigma$ and how this changes the results.

A completely fair CI has a value of 0, occurring when the entire population has an equal number of data centers within $\sigma$. The maximum unfairness favoring high income communities is 1, occurring when only the highest income person has nearby data center access. The concentration curve can also be above the line of equality, which is unfairness favoring low income communities and a negative CI.

*5.2.2 U.S. Unfairness.* We perform our analysis using 3 variables per census tract within the continental US: population, median income, and CAI. Fig 10 shows maps of each variable. We consider cases of data center locations in AWS regions, regions and Local Zones, and edge locations.

In all three cases, there is a greater availability of data center resources in higher income communities, creating unfairness. This is graphically represented in Fig 11 with the concentration curve below the line of equality. The cloud edge does help reduce this unfairness, due to additional cities with local zones containing lower income census tracts. The CI for regions is 0.41, for regions + Local Zones it is 0.21, and for the additional city with an edge location it is 0.23.

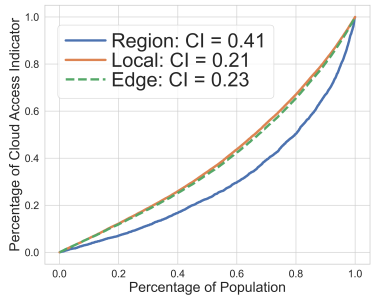
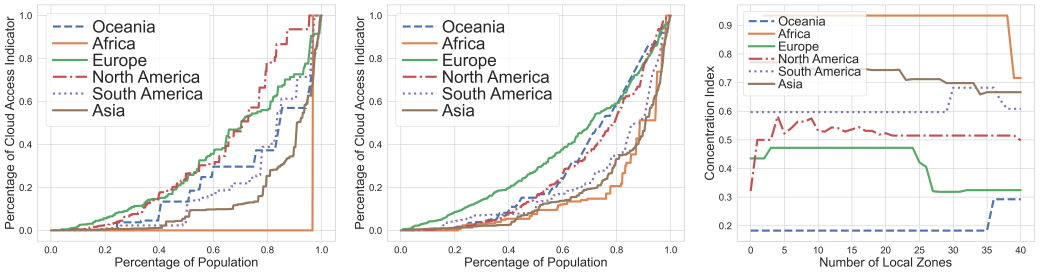



Fig. 11. Concentration curves for regions and LZs.

The deployment of data centers is already unfair before any locations are added on the network edge. A main contributor to this is the us-west-1 region which is located in silicon valley - containing many of the census tracts with the highest median incomes. In §5.3 we elaborate on this with comparisons among major US cities. Local Zones are not particularly fair locations, they are just more fair than this unfair starting point of regions. Fig 12 illustrates the change in CI as more data centers locations are added. This plot compares LZ deployments (the blue lines), and a hypothetical deployment that adds the lowest income US cities [23] (orange lines). The solid lines include the 4 US regions, and the dashed lines are new locations only. The Local Zone only line shows that each city keeps the CI around 0.2. As more Local Zones are added to the region locations, the CI approaches this value. However, the hypothetical low income city deployment demonstrates that a lower CI can be achieved. Without regions, there is a large gap between the fairness of these two deployment strategies. When the lowest income cities are included in the region locations, CI does not diverge from the LZ value until about eight cities are added, overcoming the unfairness of the initial regions.

*5.2.3 Global Unfairness.* Globally our night time light (NTL) dataset is used in place of income, and we again see unfairness on each continent favoring higher income areas because the concentration curves are below the line of equality. Fig 13a shows these concentration curves for each continent




(a) Concentration Curves for AWS Regions Only

(b) Concentration Curves for all AWS Edge Locations

(c) Change in CI when each new LZ is launched

Fig. 13. Concentration curves and the concentration index for each continent

when only considering AWS regions. Africa has the highest CI, 0.93, due to its only data center region being in an area within the 90th percentile of NTL. Fig 13b expanded the data centers considered to include all edge locations and Local Zones. The 3 additional locations in Africa at the network edge drops CI to 0.69. This is similar to what we saw when Local Zones were added in the US.

However, there isn't always a reduction in CI as new data center locations are added. North America in particular, which has the lowest region-only CI, has a higher CI with Local Zones. This can be explained by Fig 13c, which shows the change in CI per-continent as new data centers are added, starting with the first LZ in Los Angeles. This causes a jump in CI since many people in a high NTL area are within the threshold, $\sigma$, to a data center. Overall, we see similar global trends with NTL as we saw in the US with income - new locations tend to decrease the CI, but the unfairness is still high. The largest decrease is observed in Africa with the recent launch of a data center in Lagos, Nigeria.

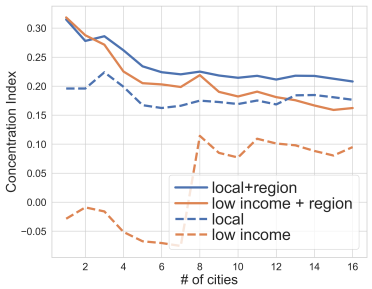



Fig. 12. CI change as new data centers are added.

The changes in CI with each launch also demonstrate the per-region rollout that new locations have been following. The first half of new locations are primarily in the US. Not until all the US locations are launched do we see Europe start to gain new locations, and only after these launched we see significant activity in the rest of the world. There is a gap between the CI values of continents with more developed networking infrastructure (North America, Europe, and Oceania) and the continents that have historically received lower investments in networking infrastructure (Asia, Africa and South America). Globally, unfairness is still very high. Since the first Local Zone it has reached a maximum of 0.73, and is now only slightly lower at 0.65. While countries or continents with more investments from cloud providers are getting more fair, there are still high levels of unfairness globally.

### 5.3 Optimal selection

Cloud providers may try to target cities with the highest population rather than minimizing unfairness. Next, we demonstrate a strategy to optimize for both. With our framework for evaluating fairness, we treat this as a multi-objective optimization problem trading off fairness and number of users. We used GeoNames [14] to list the 200 most populated US cities (excluding any less than $\sigma$ from a previously considered city). Using the same methods as §5.2.2 we find Pareto-optimal cities which lie on the front of high population but low concentration index. While a negative CI is

still unfair (0 is maximum fairness) the cities with a negative CI help balance the high CI from the largest population cities. Fig 14 illustrates this Pareto-optimal selection of cities, as well as the ones actually used for Local Zones and AWS regions.

Most of the largest population cities have data centers, with a few exceptions that are very close to others with data centers. For example, Washington D.C. and Riverside, CA are near existing data centers in Northern Virginia and Los Angeles. The majority of AWS locations have a CI > 0, again demonstrating an unfairness that gives higher income communities closer access to data centers. There are 2 cities on the Pareto front that cover more population than multiple existing locations, but do not have an AWS data center: Tampa and Orlando. Either of these cities could have been selected in place of a city currently hosting an AWS data center to provide greater than or equal population coverage within $\sigma$ while being more fair. This analysis covers just one possible trade-off between desirable metrics for edge locations, AWS likely has their own metrics to optimize for which is why the current Local Zones are not launched in cities that best improve this aspect of the CDD.


Fig. 14. Trade-off between population coverage and concentration index for the largest US cities and AWS data centers. Pareto optimal cities are marked.

## 6 CASE STUDY LEO SATELLITE ISP

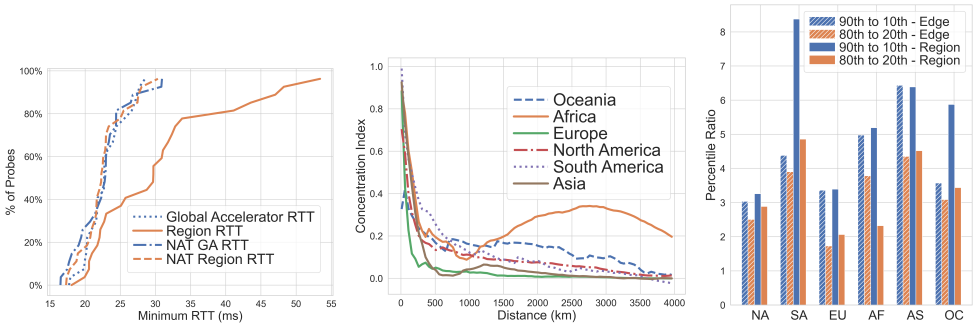
Satellite internet presents a solution for providing Internet access where it is not practical to deploy traditional fiber networks and can help bridge the digital divide in these areas (as explained in §2.2). Due to their increasing popularity in underserved regions, a discussion of the cloud's digital divide would not be complete without looking at it from the perspective of satellite ISPs. In this section, we focus specifically on a new class of satellite internet - low earth orbit (LEO) satellites.

Internet from LEO satellite constellations can already provide low-latency access to the cloud edge. The RTT to reach Amazon's edge network using Global Accelerator is comparable to the latency spent until the end of the satellite hop - when packets reach the ground station. This result is shown in Fig 15a which plots CDFs of RTT to the ground stations and endpoints of Amazon's edge network and regions. Latency to regions are considerably longer which is due to time spent in the private WAN. This tells us satellite ground stations are geographically close and well connected to cloud edge data centers.

Our measurements consisted of 9k traceroutes from all 42 RIPE Atlas probes on the Starlink network (identified using ASN) targeting the Global Accelerator IP address and VMs in AWS regions used in our earlier experiments. Starlink uses a carrier-grade network address translation (NAT) at the exit of the satellite link [76], which allows us to identify the satellite portion of the latency as the 100.64.0.1 hop in the traceroute. These probes are only in North America, Europe, and Oceania - the same continents with the best performing network infrastructure from our previous measurements. At the time of our experiments there was limited deployment in other continents, with coverage expected to expand as more ground stations are built.

While minimum RTT is currently above the 20ms threshold for more than half of the probes, it is expected to drop. Due to the satellites' low orbit, speed of light constraints limit the RTT to only ~7ms. In practice latency will be higher, yet is expected to be under 10ms [31], low enough to be a promising option for new low latency applications running on edge networks.

The minimum RTT we observed is very consistent, with 10ms difference between the fastest and slowest probe. This is due to consistent paths for most probes, regardless of location. Each




(a) Starlink latency. Ground station latency for both regions and the edge is similar to the total RTT for the edge.

(b) Unfairness tends to decrease as $\sigma$ increases because each data center can be used by a larger population.

(c) Overall inequality is much lower than ground distances, in many cases the edge reduces inequality.

Fig. 15. Case study of results using LEO satellite ISPs

path consists of the local network to satellite dish, satellite hop, and the ground station to AWS. Ideally the local network has minimal latency, and we observe the ground stations to be close to cloud edge networks. Based on the traceroute IP path, 80% of our probes could reach Amazon's edge without using another Autonomous System, confirming the homogeneity of this network across different geographies.

This similarity guides our intuition that LEO satellite networks can provide equal and fair access to all. Fig 15c shows $I$ calculated for regions and edge locations accessed over the simulated LEO satellite network. In many continents, the expansion of cloud edge computing actually reduces the inequality for this scenario, in contrast to the previous evaluation where it always increased.

Satellites can also reach a large area with low delay. A single satellite covers a ground area with a radius of 900km [68]. Furthermore, ISLs lower latency due to direct line of sight between satellites and the speed of light in space being faster than in fiber [68]. In this case, a single edge data center can provide low latency access to a wider area, changing our choice of $\sigma = 88km$ in §5.2. Fig 15b plots the CI as $\sigma$ is increased. In general, the value approaches 0 (perfect fairness) the more we can increase $\sigma$. An ideal speed of light internet can reach 3000km with a 20ms RTT, but even 1/3 of that results in much more fair access to cloud compute.

## 7 RELATED WORK

### 7.1 Cloud and Edge Latency

Many previous measurement studies of cloud access latency have focused on traditional cloud regions. Corneo et al. used RIPE Atlas probes to measure the global latency to the cloud [51]. They found that North America, Europe, and Oceania have sufficient data center infrastructure to support emerging applications, but other continents require additional investment. Dang et al. performed a similar study using SpeedChecker [30], which demonstrated much higher latency to the cloud when using wireless probes (Wi-Fi and LTE) [54]. We quantify the improvement from cloud edge by measuring latency to the private WAN, and finding a >50% latency reduction for many continents. We also show how this improvement is not equal, and new edge deployment favors areas that had relatively good prior coverage.

Prior work that does quantify edge compute has primarily covered hypothetical deployments or focused on one region. Corneo et al. use traceroutes from RIPE Atlas probes to cloud data centers in

the US to identify routers best suited for nearby edge locations [52]. They find cloud edge latency improvements of up to 30% ( ~3ms absolute difference). Their study does not extend to other countries, but the authors suggest Europe and Oceania would see similar reductions while other continents have more to gain. We confirmed their US findings in our North America measurements from §4.2, but also find large improvements in Oceania due to the additional data centers in New Zealand, which are not present in our baseline.

Previous studies have also targeted 6k+ Akamai edge servers, more prevalent than the cloud edge in our measurements, to show these servers are lower latency than cloud data centers for over 90% of probes [46]. We take a different approach by quantifying p80/p50 latency improvements with a focus on reaching thresholds that enable new applications. Additionally, Alibaba's edge compute service in China was compared to their cloud compute service based on factors including latency and application QoE [94]. To the best of our knowledge, there have not been prior global measurement studies on latency to commercially available cloud edge.

Recent work has also measured performance characteristics of networks between public clouds. Rotman et al. measure latency between cloud regions of the same three cloud providers as our measurements [84]. The Skyplane system [69] optimizes large transfers between cloud regions by measuring the highest bandwidth path. In both cases, these measurements are for inter-cloud links, while we focus on the edge connecting users to cloud networks.

### 7.2 Internet Access Inequality

Prior work has studied the availability, or lack thereof, of ISPs at the level of US counties [59]. Additionally, examining Internet providers in LA has shown that low income and minority census blocks have less access to broadband upgrades - demonstrating the inequality in Internet access is unfairly leaving these communities behind [63]. Paul et al. conducted a study across California using crowdsourced speed measurements to show how several demographic attributes such as income and education relate to internet quality, not just access [81]. Previous work has also surveyed digital literacy in developing countries, which contributes to the digital divide [37]. Our study uses global demographic data to characterize the unfairness in data center locations - which determines the minimum latency for cloud applications and introduces this metric as a new aspect affecting the digital divide. Finally, other work in the context of developing regions have looked at performance implications of free services [58, 87].

## 8 DISCUSSION

**Compute Inside Satellite Networks.** Our work demonstrates the selection of cities for cloud edge data centers creates a CDD and as new applications are developed we expect the consequences of this divide to result in unfair access to technologies that support health, education, or economic activity. This observation may be used to motivate deployments of new networks such as satellite ISPs, as was highlighted in the case study. These networks help level the playing field but the resulting latency (which is similar for all users) may be too high for certain applications such as when sub-millisecond latency matters [43]. However, there have been other proposed systems such as allowing compute to move with satellites rather than be fixed in one place [44] which we expect to help reduce inequalities and latency.

**Other Considerations in Data Center Selection.** There are other aspects to data center locations that affect the CDD which we plan to consider in future work. For instance, data sovereignty laws including GDPR may require applications to run in select countries, limiting availability of edge compute. There are also sustainability concerns when selecting a data center region such as the carbon intensity of the local electric grid and the environmental impact of construction. Providing cloud services without exceeding emission targets may further limit available data center

locations. Networking infrastructure can also affect cloud latency, and has been shown to cause high intra-continent latencies in Africa that are a barrier to adopting cloud computing [48, 62]. This infrastructure can be improved through dedicated links such as Direct Connect [12] and Megaport [24]; however, they could also further the CDD if available only in areas with relatively low latency.

**Recommendations** Our study reveals the impact of CDD to be increasing with the adoption of cloud edge. Based on this we have recommendations for developers, cloud providers, and regulators.

*Developers:* To avoid applications using the cloud edge only being available to those with the lowest latency, we recommend developers design user interactions that are adaptive to various network conditions. This has been noted before in the context of AR [36], and we suggest designing with latency variations in mind improves usability of applications across the CDD.

*Cloud providers:* We recommend cloud providers measure the CDD when prioritizing new data center locations. This metric is only one of many factors affecting location selection, but without measurements it is difficult to improve the status quo - so we suggest cloud providers keep track of how their decisions affect these metrics.

*Regulators:* For regulators, we suggest considering the CDD as one aspect in a framework for evaluating investments in data centers and internet infrastructure. This could take many forms including incentives for bringing cloud services closer to underserved communities, or deployment of satellite networks as an ISP that improves fairness.

## 9 CONCLUSION

Our study quantifies the current and expected latency improvements with the cloud edge. The cloud edge is not serving everyone equally, and widens the CDD between those with closest and furthest data centers. Furthermore, we measure the locations selected for data centers to demonstrate unfairness. Using census datasets and NTL satellite images, we show the high concentration of data centers in high income communities, even while cities with higher populations are lacking data centers. We believe our work looks at the digital divide from a new and important perspective – latency to the nearest cloud location – and provides guidance on promising technologies and deployment paths that could help in reducing the cloud digital divide.

## ACKNOWLEDGEMENTS

Thanks to the Tufts NAT Lab, the anonymous CoNEXT and JCSS reviewers, and Dr. Ihsan Ayyub Qazi for their valuable feedback and support and to the RIPE Atlas community for generous donations of credits. This work was partially funded by NSF CNS award: 2106797.

## REFERENCES


[1] Serverless Computing with Akamai EdgeWorkers | Akamai. https://www.akamai.com/products/serverless-computing-edgeworkers, .

[2] Akamai Online Retail Performance Report: Milliseconds Are Critical. https://www.akamai.com/newsroom/press-release/akamai-releases-spring-2017-state-of-online-retail-performance-report, .

[3] Amazon Found Every 100ms of Latency Cost them 1% in Sales. https://www.gigaspaces.com/blog/amazon-found-every-100ms-of-latency-cost-them-1-in-sales.

[4] ArcGIS Online. https://www.arcgis.com/index.html.

[5] Network Acceleration Service - AWS Global Accelerator - AWS. https://aws.amazon.com/global-accelerator/, .

[6] AWS Global Accelerator FAQs. https://aws.amazon.com/global-accelerator/faqs/, .

[7] AWS Global Accelerator launches TCP Termination at the Edge. https://aws.amazon.com/about-aws/whats-new/2020/03/aws-global-accelerator-launches-tcp-termination-at-the-edge/, .

[8] Azure Edge Zones: Microsoft's Plan to Dominate Edge Computing and 5G. https://www.datacenterknowledge.com/microsoft/azure-edge-zones-microsoft-s-plan-dominate-edge-computing-and-5g.

[9] Increase in Income Inequality Driven by Real Declines in Income at the Bottom. https://www.census.gov/library/stories/2022/09/income-inequality-increased.html.

[10] Cloudflare Workers. https://workers.cloudflare.com.

[11] America's Digital Divide. https://www.pewtrusts.org/en/trust/archive/summer-2019/americas-digital-divide.

[12] Dedicated Network Connections - AWS Direct Connect. https://aws.amazon.com/directconnect/.

[13] GADM. https://www.gadm.org/data.html.

[14] GeoNames. http://www.geonames.org.

[15] Announcing Google Distributed Cloud Edge and Hosted | Google Cloud Blog. https://cloud.google.com/blog/topics/hybrid-cloud/announcing-google-distributed-cloud-edge-and-hosted, .

[16] Geeking with Greg: Marissa Mayer at Web 2.0. http://glinden.blogspot.com/2006/11/marissa-mayer-at-web-20.html, .

[17] Expanding Internet Access Improves Health Outcomes. https://www.govtech.com/network/expanding-internet-access-improves-health-outcomes.html.

[18] Lambda@Edge. https://aws.amazon.com/lambda/edge/.

[19] AWS Local Zones and AWS Outposts, choosing the right technology for your edge workload. https://aws.amazon.com/blogs/compute/aws-local-zones-and-aws-outposts-choosing-the-right-technology-for-your-edge-workload/, .

[20] Couchbase Reduces Latency by 80% for Its Distributed Database Solutions Using AWS Local Zones. https://aws.amazon.com/solutions/case-studies/couchbase-case-study/, .

[21] Local Zone Locations - Amazon Web Services. https://aws.amazon.com/about-aws/global-infrastructure/localzones/locations, .

[22] Introducing AWS Local Zone in Los Angeles, CA. https://aws.amazon.com/about-aws/whats-new/2019/12/introducing-aws-local-zone-in-los-angeles-ca/, .

[23] Top Poorest Cities in US. https://www.neoch.org/top-poorest-cities-in-us.

[24] Megaport | Cloud Interconnectivity Simplified. https://www.megaport.com.

[25] Netflix Empowers Remote Artistry with Low-Latency Workstations Using AWS Local Zones. https://aws.amazon.com/solutions/case-studies/netflix-aws-local-zones-case-study/.

[26] New AWS Local Zones deliver lower latency in Chicago and New York City. https://aws.amazon.com/blogs/industries/new-aws-local-zones-deliver-lower-latency-in-chicago-and-new-york-city/.

[27] A study of lights at night suggests dictators lie about economic growth. https://www.economist.com/graphic-detail/2022/09/29/a-study-of-lights-at-night-suggests-dictators-lie-about-economic-growth.

[28] RIPE Atlas. https://atlas.ripe.net.

[29] Sustainable Development Goal 9: Investing in ICT access and quality education to promote lasting peace. https://www.un.org/sustainabledevelopment/blog/2017/06/sustainable-development-goal-9-investing-in-ict-access-and-quality-education-to-promote-lasting-peace/.

[30] SpeedChecker. https://www.speedchecker.com.

[31] Starlink's Latency Will Become Fit for Competitive Online Gaming, Musk Says. https://www.pcmag.com/news/starlinks-latency-will-become-fit-for-competitive-online-gaming-musk-says.

[32] TIGER/Line Shapefiles. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2020.html.

[33] Esri Updated Demographics. https://doc.arcgis.com/en/esri-demographics/latest/regional-data/updated-demographics.htm.

[34] VIIRS Nighttime Light. https://eogdata.mines.edu/products/vnl/.

[35] T. Ahsen, F. R. Dogar, and A. L. Gardony. Exploring the impact of network impairments on remote collaborative augmented reality applications. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

[36] T. Ahsen, Z. Y. Lim, A. L. Gardony, H. A. Taylor, J. P. d. Ruiter, and F. Dogar. The effects of network outages on user experience in augmented reality based remote collaboration-an empirical study. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27, 2021.

[37] A. Ali, A. A. Raza, and I. A. Qazi. Validated digital literacy measures for populations with low levels of internet experiences. *Development Engineering*, 8:100107, 2023. ISSN 2352-7285. doi: https://doi.org/10.1016/j.deveng.2023.100107. URL https://www.sciencedirect.com/science/article/pii/S2352728523000015.

[38] G. Ananthanarayanan, V. Bahl, L. Cox, A. Crown, S. Nogbahi, and Y. Shu. Video analytics - killer app for edge computing. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '19, page 695–696, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366618. doi: 10.1145/3307334.3328589. URL https://doi.org/10.1145/3307334.3328589.

[39] T. Arnold, E. Gürmeriçliler, G. Essig, A. Gupta, M. Calder, V. Giotsas, and E. Katz-Bassett. (how much) does a private wan improve cloud performance? In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, pages 79–88, 2020. doi: 10.1109/INFOCOM41043.2020.9155428.

[40] T. Arnold, J. He, W. Jiang, M. Calder, I. Cunha, V. Giotsas, and E. Katz-Bassett. Cloud provider connectivity in the flat internet. In *Proceedings of the ACM Internet Measurement Conference*, IMC '20, page 230–246, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450381383. doi: 10.1145/3419394.3423613. URL https://doi.org/10.1145/3419394.3423613.

[41] Y. Asada, J. Hurley, O. F. Norheim, and M. Johri. A three-stage approach to measuring health inequalities and inequities. *International Journal for Equity in Health*, 13(1):98, Nov 2014. ISSN 1475-9276. doi: 10.1186/s12939-014-0098-y. URL https://doi.org/10.1186/s12939-014-0098-y.

[42] V. Bajpai, S. J. Eravuchira, and J. Schönwälder. Lessons learned from using the ripe atlas platform for measurement research. *SIGCOMM Comput. Commun. Rev.*, 45(3):35–42, jul 2015. ISSN 0146-4833. doi: 10.1145/2805789.2805796. URL https://doi.org/10.1145/2805789.2805796.

[43] D. Bhattacherjee, W. Aqeel, G. Laughlin, B. M. Maggs, and A. Singla. A bird's eye view of the world's fastest networks. In *Proceedings of the ACM Internet Measurement Conference*, IMC '20, page 521–527, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450381383. doi: 10.1145/3419394.3423620. URL https://doi.org/10.1145/3419394.3423620.

[44] D. Bhattacherjee, S. Kassing, M. Licciardello, and A. Singla. In-orbit computing: An outlandish thought experiment? In *Proceedings of the 19th ACM Workshop on Hot Topics in Networks*, HotNets '20, page 197–204, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450381451. doi: 10.1145/3422604.3425937. URL https://doi.org/10.1145/3422604.3425937.

[45] M. Calder, A. Flavel, E. Katz-Bassett, R. Mahajan, and J. Padhye. Analyzing the performance of an anycast cdn. IMC '15, page 531–537, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450338486. doi: 10.1145/2815675.2815717. URL https://doi.org/10.1145/2815675.2815717.

[46] B. Charyyev, E. Arslan, and M. H. Gunes. Latency comparison of cloud datacenters and edge servers. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, pages 1–6, 2020. doi: 10.1109/GLOBECOM42002.2020.9322406.

[47] Y. Chavehpour, A. Rashidian, A. Woldemichael, and A. Takian. Inequality in geographical distribution of hospitals and hospital beds in densely populated metropolitan cities of iran. *BMC Health Services Research*, 19, 08 2019. doi: 10.1186/s12913-019-4443-0.

[48] J. Chavula, A. Phokeer, and E. Calandro. Performance barriers to cloud services in africa's public sector: A latency perspective. In G. Mendy, S. Ouya, I. Dioum, and O. Thiaré, editors, *e-Infrastructure and e-Services for Developing Countries*, pages 152–163, Cham, 2019. Springer International Publishing. ISBN 978-3-030-16042-5.

[49] L. Chiou and C. Tucker. Social distancing, internet access and inequality. Working Paper 26982, National Bureau of Economic Research, April 2020. URL http://www.nber.org/papers/w26982.

[50] Chris Wedel. Starlink internet is going from rural savior to unreliable luxury. https://www.xda-developers.com/starlink-internet-rural-savior-unreliable-luxury/.

[51] L. Corneo, M. Eder, N. Mohan, A. Zavodovski, S. Bayhan, W. Wong, P. Gunningberg, J. Kangasharju, and J. Ott. Surrounded by the clouds: A comprehensive cloud reachability study. In *Proceedings of the Web Conference 2021*, WWW '21, page 295–304, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449854. URL https://doi.org/10.1145/3442381.3449854.

[52] L. Corneo, N. Mohan, A. Zavodovski, W. Wong, C. Rohner, P. Gunningberg, and J. Kangasharju. (how much) can edge computing change network latency? In *2021 IFIP Networking Conference (IFIP Networking)*, pages 1–9, 2021. doi: 10.23919/IFIPNetworking52078.2021.9472847.

[53] D. Crankshaw, X. Wang, G. Zhou, M. J. Franklin, J. E. Gonzalez, and I. Stoica. Clipper: A Low-Latency online prediction serving system. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, pages 613–627, Boston, MA, Mar. 2017. USENIX Association. ISBN 978-1-931971-37-9. URL https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/crankshaw.

[54] T. K. Dang, N. Mohan, L. Corneo, A. Zavodovski, J. Ott, and J. Kangasharju. Cloudy with a chance of short rtts: Analyzing cloud connectivity in the internet. In *Proceedings of the 21st ACM Internet Measurement Conference*, IMC '21, page 62–79, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450391290. doi: 10.1145/3487552.3487854. URL https://doi.org/10.1145/3487552.3487854.

[55] F. R. Dogar. Towards slack-aware networking. *ACM SIGCOMM Computer Communication Review*, 48(2):24–30, 2018.

[56] F. R. Dogar and P. Steenkiste. Architecting for edge diversity: Supporting rich services over an unbundled transport. In *Proceedings of the 8th international conference on Emerging networking experiments and technologies (CoNEXT)*, pages 13–24, 2012.

[57] F. R. Dogar, P. Steenkiste, and K. Papagiannaki. Catnap: exploiting high bandwidth wireless interfaces to save energy for mobile devices. In *Proc. ACM Mobisys*, pages 107–122, 2010.

[58] F. R. Dogar, I. A. Qazi, A. R. Tariq, G. Murtaza, A. Ahmad, and N. Stocking. Missit: Using missed calls for free, extremely low bit-rate communication in developing regions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–12, 2020.

[59] R. Durairajan and P. Barford. A techno-economic approach for broadband deployment in underserved areas. *SIGCOMM Comput. Commun. Rev.*, 47(2):13–18, may 2017. ISSN 0146-4833. doi: 10.1145/3089262.3089265. URL https://doi.org/10.1145/3089262.3089265.

[60] R. Fontugne, A. Shah, and K. Cho. Persistent last-mile congestion: Not so uncommon. In *Proceedings of the ACM Internet Measurement Conference*, IMC '20, page 420–427, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450381383. doi: 10.1145/3419394.3423648. URL https://doi.org/10.1145/3419394.3423648.

[61] C. for International Earth Science Information Network CIESIN Columbia University. Gridded population of the world, version 4 (gpwv4): Population density, revision 11, 20230522 2018. URL https://doi.org/10.7927/H49C6VHW.

[62] A. Formoso, J. Chavula, A. Phokeer, A. Sathiaseelan, and G. Tyson. Deep diving into africa's inter-country latencies. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 2231–2239, 2018. doi: 10.1109/INFOCOM.2018.8486024.

[63] H. Galperin, T. V. Le, and K. Wyatt. Who gets access to fast broadband? evidence from los angeles county. *Government Information Quarterly*, 38(3):101594, 2021. ISSN 0740-624X. doi: https://doi.org/10.1016/j.giq.2021.101594. URL https://www.sciencedirect.com/science/article/pii/S0740624X21000307.

[64] D. Han, A. Anand, F. Dogar, B. Li, H. Lim, M. Machado, A. Mukundan, W. Wu, A. Akella, D. G. Andersen, et al. Xia: Efficient support for evolvable internetworking. In *Proceedings of NSDI*, 2012.

[65] O. Haq and F. R. Dogar. Leveraging the power of cloud for reliable wide area communication. In *Proceedings of the 14th ACM Workshop on Hot Topics in Networks (Hotnets)*, pages 1–7, 2015.

[66] O. Haq, M. Raja, and F. R. Dogar. Measuring and improving the reliability of wide-area cloud paths. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 253–262, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052560. URL https://doi.org/10.1145/3038912.3052560.

[67] O. Haq, C. Doucette, J. W. Byers, and F. R. Dogar. Judicious qos using cloud overlays. In *Proceedings of the 16th International Conference on emerging Networking EXperiments and Technologies (CoNEXT)*, pages 371–385, 2020.

[68] Y. Hauri, D. Bhattacherjee, M. Grossmann, and A. Singla. "internet from space" without inter-satellite links. In *Proceedings of the 19th ACM Workshop on Hot Topics in Networks*, HotNets '20, page 205–211, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450381451. doi: 10.1145/3422604.3425938. URL https://doi.org/10.1145/3422604.3425938.

[69] P. Jain, S. Kumar, S. Wooders, S. G. Patil, J. E. Gonzalez, and I. Stoica. Skyplane: Optimizing transfer cost and throughput using cloud-aware overlays. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23) (To Appear)*, Boston, MA, 2023. USENIX Association. URL https://www.usenix.org/conference/nsdi23/presentation/jain.

[70] Z. Jia and E. Witchel. Nightcore: Efficient and scalable serverless computing for latency-sensitive, interactive microservices. In *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '21, page 152–166, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383172. doi: 10.1145/3445814.3446701. URL https://doi.org/10.1145/3445814.3446701.

[71] E. Jo, D. A. Epstein, H. Jung, and Y.-H. Kim. Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581503. URL https://doi.org/10.1145/3544548.3581503.

[72] M. M. Kassem, A. Raman, D. Perino, and N. Sastry. A browser-side view of starlink connectivity. In *Proceedings of the 22nd ACM Internet Measurement Conference*, IMC '22, page 151–158, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392594. doi: 10.1145/3517745.3561457. URL https://doi.org/10.1145/3517745.3561457.

[73] S. Kassing, D. Bhattacherjee, A. B. Águas, J. E. Saethre, and A. Singla. Exploring the "internet from space" with hypatia. In *Proceedings of the ACM Internet Measurement Conference*, IMC '20, page 214–229, New York, NY, USA, 2020.

Association for Computing Machinery. ISBN 9781450381383. doi: 10.1145/3419394.3423635. URL https://doi.org/10.1145/3419394.3423635.

[74] T. Koch, E. Katz-Bassett, J. Heidemann, M. Calder, C. Ardi, and K. Li. Anycast in context: A tale of two systems. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, SIGCOMM '21, page 398–417, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383837. doi: 10.1145/3452296.3472891. URL https://doi.org/10.1145/3452296.3472891.

[75] K. Mania, B. D. Adelstein, S. R. Ellis, and M. I. Hill. Perceptual sensitivity to head tracking latency in virtual environments with varying degrees of scene complexity. In *Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization*, APGV '04, page 39–47, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581139144. doi: 10.1145/1012551.1012559. URL https://doi-org.ezproxy.library.tufts.edu/10.1145/1012551.1012559.

[76] F. Michel, M. Trevisan, D. Giordano, and O. Bonaventure. A first look at starlink performance. In *Proceedings of the 22nd ACM Internet Measurement Conference*, IMC '22, page 130–136, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392594. doi: 10.1145/3517745.3561416. URL https://doi.org/10.1145/3517745.3561416.

[77] S. Milano, J. A. McGrane, and S. Leonelli. Large language models challenge the future of higher education. *Nature Machine Intelligence*, 5(4):333–334, Apr 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00644-2. URL https://doi.org/10.1038/s42256-023-00644-2.

[78] A. M. Noor, V. A. Alegana, P. W. Gething, A. J. Tatem, and R. W. Snow. Using remotely sensed night-time light as a proxy for poverty in africa. *Population Health Metrics*, 6(1):5, Oct 2008. ISSN 1478-7954. doi: 10.1186/1478-7954-6-5. URL https://doi.org/10.1186/1478-7954-6-5.

[79] O'Donnell, Owen; van Doorslaer, Eddy; Wagstaff, Adam; Lindelow, Magnus. Analyzing Health Equity Using Household Survey Data : A Guide to Techniques and Their Implementation. 2008. URL https://openknowledge.worldbank.org/handle/10986/6896.

[80] A. Padmanabhan, N. Agarwal, A. Iyer, G. Ananthanarayanan, Y. Shu, N. Karianakis, G. H. Xu, and R. Netravali. Gemel: Model merging for Memory-Efficient, Real-Time video analytics at the edge. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 973–994, Boston, MA, Apr. 2023. USENIX Association. ISBN 978-1-939133-33-5. URL https://www.usenix.org/conference/nsdi23/presentation/padmanabhan.

[81] U. Paul, J. Liu, D. Farias-llerenas, V. Adarsh, A. Gupta, and E. Belding. Characterizing internet access and quality inequities in california m-lab measurements. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*, COMPASS '22, page 257–265, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393478. doi: 10.1145/3530190.3534813. URL https://doi.org/10.1145/3530190.3534813.

[82] D. Perdices, G. Perna, M. Trevisan, D. Giordano, and M. Mellia. When satellite is all you have: Watching the internet from 550 ms. In *Proceedings of the 22nd ACM Internet Measurement Conference*, IMC '22, page 137–150, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392594. doi: 10.1145/3517745.3561432. URL https://doi.org/10.1145/3517745.3561432.

[83] X. S. Pérez-Sindín, T.-H. K. Chen, and A. V. Prishchepov. Are night-time lights a good proxy of economic activity in rural areas in middle and low-income countries? examining the empirical evidence from colombia. *Remote Sensing Applications: Society and Environment*, 24:100647, 2021. ISSN 2352-9385. doi: https://doi.org/10.1016/j.rsase.2021.100647. URL https://www.sciencedirect.com/science/article/pii/S235293852100183X.

[84] N. H. Rotman, Y. Ben-Itzhak, A. Bergman, I. Cidon, I. Golikov, A. Markuze, and E. Zohar. Cloudcast: Characterizing public clouds connectivity. *CoRR*, abs/2201.06989, 2022. URL https://arxiv.org/abs/2201.06989.

[85] M. Satyanarayanan. How we created edge computing. *Nature Electronics*, 2(1):42–42, Jan 2019. ISSN 2520-1131. doi: 10.1038/s41928-018-0194-x. URL https://doi.org/10.1038/s41928-018-0194-x.

[86] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies. The case for vm-based cloudlets in mobile computing. *IEEE pervasive Computing*, 8(4):14–23, 2009.

[87] R. Sen, S. Ahmad, A. Phokeer, Z. A. Farooq, I. A. Qazi, D. Choffnes, and K. P. Gummadi. Inside the walled garden: Deconstructing facebook's free basics program. *ACM SIGCOMM Computer Communication Review*, 47(5):12–24, 2017.

[88] A. Singla, B. Chandrasekaran, P. B. Godfrey, and B. Maggs. The internet at the speed of light. In *Proceedings of the 13th ACM Workshop on Hot Topics in Networks*, HotNets-XIII, page 1–7, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450332569. doi: 10.1145/2670518.2673876. URL https://doi.org/10.1145/2670518.2673876.

[89] U. Speidel. What the tropical pacific wants from starlink for christmas: Will leo networks finally bridge the digital divide to remote islands? In *Proceedings of the 16th Asian Internet Engineering Conference*, AINTEC '21, page 34–40, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450391849. doi: 10.1145/3497777.3498548. URL https://doi-org.ezproxy.library.tufts.edu/10.1145/3497777.3498548.

[90] T. Tambe, C. Hooper, L. Pentecost, T. Jia, E.-Y. Yang, M. Donato, V. Sanh, P. Whatmough, A. M. Rush, D. Brooks, and G.-Y. Wei. Edgebert: Sentence-level energy optimizations for latency-aware multi-task nlp inference. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO '21, page 830–844, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385572. doi: 10.1145/3466752.3480095. URL

https://doi.org/10.1145/3466752.3480095.

[91] M. F. Thompson. Income Inequality. *Indiana Business Review*, 87(3), 2012.

[92] D. Vilar, M. Freitag, C. Cherry, J. Luo, V. Ratnakar, and G. Foster. Prompting palm for translation: Assessing strategies and performance, 2022.

[93] Y. Wang, K. Chen, H. Tan, and K. Guo. Tabi: An efficient multi-level inference system for large language models. In *Proceedings of the Eighteenth European Conference on Computer Systems*, EuroSys '23, page 233–248, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394871. doi: 10.1145/3552326.3587438. URL https://doi-org.ezproxy.library.tufts.edu/10.1145/3552326.3587438.

[94] M. Xu, Z. Fu, X. Ma, L. Zhang, Y. Li, F. Qian, S. Wang, K. Li, J. Yang, and X. Liu. From cloud to edge: A first look at public edge platforms. In *Proceedings of the 21st ACM Internet Measurement Conference*, IMC '21, page 37–53, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450391290. doi: 10.1145/3487552.3487815. URL https://doi.org/10.1145/3487552.3487815.

[95] J. Zhang, S. Elnikety, S. Zarar, A. Gupta, and S. Garg. Model-Switching: Dealing with fluctuating workloads in Machine-Learning-as-a-Service systems. In *12th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 20)*. USENIX Association, July 2020. URL https://www.usenix.org/conference/hotcloud20/presentation/zhang.

[96] L. Zhang, A. Afanasyev, J. Burke, V. Jacobson, K. Claffy, P. Crowley, C. Papadopoulos, L. Wang, and B. Zhang. Named data networking. *ACM SIGCOMM Computer Communication Review*, 44(3):66–73, 2014.

[97] L. Zou, W. Lu, Y. Liu, H. Cai, X. Chu, D. Ma, D. Shi, Y. Sun, Z. Cheng, S. Gu, S. Wang, and D. Yin. Pre-trained language model-based retrieval and ranking for web search. *ACM Trans. Web*, 17(1), dec 2022. ISSN 1559-1131. doi: 10.1145/3568681. URL https://doi.org/10.1145/3568681.

# A APPENDIX

## A.1 Reduction in p90 and p10 values

In §5.1.2 we showed how edge data centers increase the inequality, $I$. Despite this increase in the relative difference between $p90$ and $p10$, both decrease when considering edge data centers. Table A.1 demonstrates the improvements to both percentiles provided by cloud edge. The minimum distance to a data center drops substantially, with over 2000km reductions at the $p90$ for two continents.

Table 3 presents the $p90$ and $p10$ values from our LEO simulations in §6. In this scenario the higher latencies ($p90$) are the ones most affected by the expansion of cloud edge. This follows from our intuition that the satellite hop will dominate the latency. The data also confirms the lower bound does not change much (due to the Earth to satellite hop) and the RTT can be low enough to support emerging edge applications.

| Continent | Region | | Edge | |
|---|---|---|---|---|
| | p10 | p90 | p10 | p90 |
| Oceania | 41 | 3209 | 10 | 2573 |
| Africa | 1385 | 3914 | 281 | 1807 |
| Europe | 104 | 1842 | 23 | 1274 |
| North America | 219 | 2811 | 23 | 1147 |
| South America | 309 | 3864 | 22 | 1776 |
| Asia | 175 | 1569 | 60 | 871 |

Table 2. Values of $p10$ and $p90$ in kilometers.

| Continent | Region | | Edge | |
|---|---|---|---|---|
| | p10 | p90 | p10 | p90 |
| Oceania | 7.64 | 44.86 | 7.03 | 25.09 |
| Africa | 10.65 | 55.31 | 9.51 | 47.28 |
| Europe | 7.34 | 24.9 | 7.17 | 24.1 |
| North America | 7.84 | 25.55 | 7.73 | 23.49 |
| South America | 8.61 | 72.12 | 8.34 | 36.57 |
| Asia | 9.55 | 60.99 | 9.41 | 60.5 |

Table 3. Values of p10 and p90 RTT in milliseconds from the Starlink simulations used to calculate $I$ for each continent.

Received July 2023; accepted October 2023