Worldwide Measurements Using the Power of Ripe Atlas

Jasper von der Heidt, Jan Oesterle, Zhihang Wei, Paul Wiessner

I. Introduction

Network latencies have become a critical issue in the modern world of technology. As the internet has grown more ubiquitous, the demand for fast and reliable communication has increased, making network latency a critical metric for measuring the performance of networks.

Network latency has become a critical issue in the modern world of technology as the demand for fast and reliable communication has increased. The history of network latencies and bandwidth is rich, with early networks like ARPANET built to enable military and academic institutions to communicate. However, as the internet became more commercialized and the amount of data transmitted increased, latency became a more significant concern. Today, low latency is essential for a wide range of applications, including online gaming, video conferencing, and financial trading, among others. New technologies like 5G networks and edge computing have increased the demand for even lower latencies. Low latency enables computational offloading and has numerous benefits, such as reducing energy consumption and enabling complex applications to run on less powerful devices. While latency is critical for quality of experience, it also offers potential applications for businesses and consumers.

Reduced network latency allows expensive computations to be offloaded to nearby servers, which is made possible by edge computing, fog computing, and peering strategies. Edge computing moves processing tasks to the edge of the network, closer to where data is generated, reducing the distance data has to travel. Fog computing offloads processing tasks to a distributed network of edge devices. Peering allows networks to connect directly, reducing the number of hops between the server and client, resulting in lower latency and faster content delivery.

Low latency is crucial for AI technologies that require significant computational power. The ability to offload processing tasks to more powerful computing resources has been essential in making AI accessible to various applications. Deploying AI on low-power edge devices like smartphones and wearables enables AI to be used in real-time applications that require low latency and can operate even in disconnected environments.

However, achieving low latency is challenging, and the last mile bottleneck is one of the most significant obstacles. The last mile is the physical connection between the user and the internet service provider and can cause latency due to factors like network congestion, signal interference, and geographical distance. Understanding the latencies in the internet and the impact of the last mile bottleneck is essential to improving latency behavior and estimating requirements for low latency applications.

Analyzing the connectivity of wireless vs. wired networks globally and the impact of distance between the probe and cloud data center to end-to-end latency, we aim to examine the performance of different access technologies and compare the state of different last-mile access technologies in different continents. By doing so, we hope to contribute to the ongoing efforts to improve internet latency and bandwidth and enable the development of low latency applications that can revolutionize various industries.

Latency of internet connections varies significantly globally due to political, economic, and technological factors. Political decisions or lack of investment in broadband infrastructure can slow down internet speeds and increase latency in some countries, such as Germany, where broadband expansion has been delayed due to political decisions. In parts of Africa and South America, limited monetary incentives for internet service providers to deploy large numbers of servers can result in slower internet speeds and higher latency. Technological factors, such as the type of last-mile access technology used in a particular region, can also impact latency. Analyzing these factors can help find solutions to improve internet connectivity and reduce latency worldwide.

II. Research Goals

- understand the diversity of mobile and fixed Internet access across the globe
- understand impact of last-mile access
- find and examine routing paths from ISP clients to cloud providers
- evaluate the potential for offloading computations to compute devices closer to the users
- analyze operational difference/similarities between LEO satellite connections and traditional WiFi or mobile connectivity.

III. Experiment Design

A. Probe Selection

Choosing the right probes for network measurement is crucial to obtaining accurate and representative data. Probes should be distributed over all continents and countries, over cities and rural areas as well as over multiple network types and Internet Service Providers (ISPs) to ensure that a comprehensive and representative view of internet performance can be obtained. Our probe selection is based on the probe information data as of $2023-02-07$. Having this information

	cellular	home wifi	satellite	ethernet
Tag (obl.)	–	*home*	*starlink*	*home*
Tag (any)	*3g*, *4g*, *lte*, *5g*, *mobile*	*wi-fi*, *wireless*, *system-wifi*, *fixed-wireless*, *wireless-isp*	–	*fibre*, *cable*, *ftth*
Tag (exclude)	*dsl*, *vdsl*, *vdsl2*, *adsl*	–	–	–
Filter	–	–	$ASN14593$	distance (min)
# Probes	168	122	63	290

Fig. 1: Probe Filter Criteria

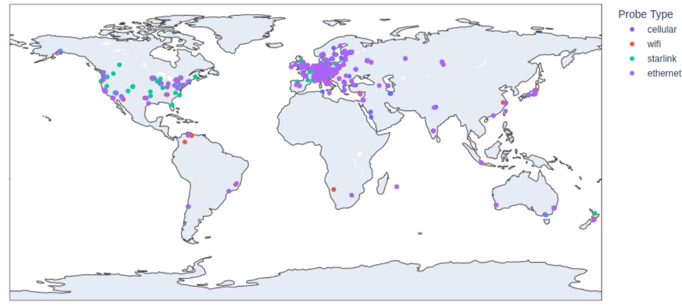


Fig. 2: Map of Probe Selection

we need to set criteria for pre-filter them. We need our selected probes to be active (exclude probes having another status than *connected*). Furthermore, we don't want them to be located in a datacenter (exclude all probes having the tag *datacenter*, *datacentre*, *aws*, *gcp*). The result is a selection which was active at the day of starting the measurements. The last mile of a connection might have a significant effect on the overall latency. We differ between four different last mile alternatives: cellular, home WiFi, Ethernet and satellite. For each of them we define criteria to filter probes. See Table 1 for an overview on filtering criteria.

1) Cellular: Probes, connected via cellular radio must use one of the established mobile internet technologies from 3G to 5g. As there exist devices which are routers receiving internet connection by a cellular module, we explicitly exclude them.

2) Home WiFi: Home WiFis can be of various different kinds. From most common variant of a router connected by a wire to devices by the ISP connected by cellular setting up a WiFi hotspot in e.g. rural areas, where no cable infrastructure is available. However, we require them to be related to a home WiFi, sparing out public WiFi hotspots.

3) Satellite: Internet connection via satellite is mainly represented by Starlink, a low earth orbit (LEO) system. Probes connected to Starlink can be easily identified by Starlink's autonomous system number 14593.

4) Ethernet: For comparison to cellular and home WiFi, we select one probe for each cellular and WiFi probe connected via Ethernet which minimizes geographical distance. Additionally it should be related to a home network.

Using these criteria, we found a total of 643 probes spread across the earth as shown in Figure 2. From the map we can observe big differences in distribution of probes. The major part, almost 70% is located in Europe, rather in its norther part. Second most probes with 18% are located in North America, having more than half of the total satellite probes. On the last place is Africa, hosting only 1% of all probes. This leads to a highly unbalanced distribution of probes between Europe and North America, and the remaining continents. Taking this as basis, we regard these differences in the following analysis.

B. Datacenter Selection

As we need to compare whether there exists different routers on-path for different providers in destination datacenters, and also compare the performance of these different providers, we need to exclude other variables. We tried to choose the locations where all three providers (Amazon, Google and Microsoft) have deployed their datacenters nearby, ideally in the same city. In this way, we found 12 sets of datacenters, each set includes 3 datacenters belongs to each provider. These datacenters locates here:

- Oceania, Australia: Sydney
- South America, Brazil: Sao Paulo, Campinas
- North America, US east: Ashburn, Washington
- North America, US west: San Jose, Los Angeles
- North America, Canada: Toronto, Montreal
- Europe, Germany: Frankfurt am Main
- Europe, UK: London
- Asia, India: Mumbai
- Asia, Hong Kong
- Asia, Singapore
- Asia, Japan: Tokyo
- Asia, Korea: Seoul, Incheon

In the next step we will map these sets of datacenters with probes to create complete routes for ping and traceroute.

C. Route Selection

As the project description has stated, our measurement should be a good representation of global networks. At the same time, in several sub tasks, we need to compare the route and latency from one probe to different locations, especially when the datacenters locate in the same country as the probe, in the neighbor country but in the same continent as the probe, and in the neighbor continent of the probe.

By this guideline, we first try to assign the probes with a set of datacenters in their home country if available, while trying to assign them 2-3 sets of datacenters which are in the same continent, preferably the neighbor country. And finally, assign 1-2 sets of datacenters which locate in neighbor continents. In total we assigned 4 sets of datacenters (12 datacenters) for each probe.

Also, we tried to choose the destination datacenters which a client in the same area of this probe would most likely to be connected. We found a plot online [1]which shows the amount of network connection between different regions in the world. We had the following discoveries and made corresponding decisions on datacenters mapping:

- for Asia, Hong Kong and Singapore hold the most internet traffic, and the whole continent connects with North America (west coast) and Europe the most; we choose two sets of datacenters inside Asia, either in their own country or HongKong / Singapore, and then Frankfurt and US west;
- for Europe, Frankfurt and London hold the most internet traffic, and the whole continent connects with North America (east coast), Middle East and Asia the most; we choose two sets of datacenters inside Europe, Frankfurt and London, then US east and Hong Kong;
- for North America, obviously most traffic flows to east and west coast of US, outside the continent it goes to South America, Europe and Asia; we choose one in US, whichever is nearer, Canada, South America and Frankfurt;
- for South America, most traffic flows to US; we choose the only datacenters in South America, then one in US, then Frankfurt and Hong Kong;
- for Oceania, most traffic flows to US west coast and Asia; we choose the only one in Sydney, then Hong Kong, US west and Frankfurt;
- for Africa, unfortunately there's no datacenter for us to select; we choose Mumbai, Hong Kong, Frankfurt and US east, which could be the nearest on map;

We believe this semi-static mapping could be the best representation of global network traffic, considering the limited datacenters we can choose from.

D. Measurement Techniques

The RIPE Atlas platform provides a variety of powerful tools for conducting internet measurements on a global scale. In our study, we focused on two of these tools: ping and (Paris) traceroute.

Ping is a widely used tool for measuring the round-trip-time (RTT) between two endpoints in a network. By sending small packets from a source probe to a destination datacenter, it measures the time how long it takes to travel to the destination and back. By doing this three times for each scheduled ping, we can obtain an accurate estimation of the RTT.

In addition to ping, generating solely information about RTT, we use (paris) traceroute. Traceroute is a tool providing more information about the network path between a probe and a datacenter. I.e. it aims to provide a full list of hosts traversed from source to destination. Paris traceroute [2] is a modified version of traceroute. It's designed to provide a better understanding of the true network infrastructure by introducing slight deviations in the header. As a result, packets may be forwarded differently by load balancers and different routes of underlying network structure is revealed.

By combining measurements of ping and paris traceroute we are able to obtain detailed understanding of the network path and temporal influence on network performance.

IV. Analysis

A. Latency Differences Between Different Access Technologies

In order to compare the impact of different access technologies on the latency in the internet, it is a necessity to create similar conditions. With the goal of approximating such conditions, we only directly compare results where the probes are in the same country and ping the same data center. While this ensures intra-comparability, this allows also analyze the impact different geographical regions have on the latency. In figure 4 one can see the results of these measurements. An important point to consider is, that due to the heavily biased distributions of probes strongly favoring Europe, followed by North America, Asia, Oceania and South America, the significance of the measurements for the latter regions is limited. Nevertheless, it allows for a basic impression of the potential differences. The results of our measurements clearly show, throughout all regions, starlink has the highest negative impact on the latency. While the impact is relatively low in Europe and Asia, relatively similar to latency caused by cellular access, the results form North America show that there, Starlink causes an around four times higher latency than WiFi and Ethernet. On the other hand, the results from Ethernet and WiFi based pings show the best performance in most cases. The latency in general has shown to be lower overall in Europe and Northern America compared to the other continents. On both these continents, regardless of the chosen access technology, the latency rarely passes the mark of 100ms, while our measurements show that especially in South America and Asia, latencies around 150ms are the standard.

Aside form the latency differences between different access technologies, we also further investigated wether there exists correlation between latency and time of day. For these measurements, we again used the same choice of probes to ensure a certain degree of comparability of our measurements. The results can be seen in figure 5. These results reinforce our initial observations with regards to the tendency of Starlink performing worse than , and Ethernet and WiFi performing above average. Aside form allowing the comparison of general latency, analyzing changes in RTT over time allow for a deeper insight in the stability and reliability of the different access technologies. Across the different measurements, it is observable that the latency fluctuation of Ethernet is overall vastly lower than the other technologies, followed again by WiFi. On the other hand, starlink seems to have an edge over cellular. In terms of periodicity of the latency fluctuation, disregarding the outliers, the data hints at a cycle of higher latency in the middle of the day and on average lower latencies at night.

To further understand why such latency fluctuations can happen, we took a closer look at latencies based on the network provider. Here, we took the the all our measurements into consideration. While overall the results are relatively similar, according to our measurements, data-centers operated

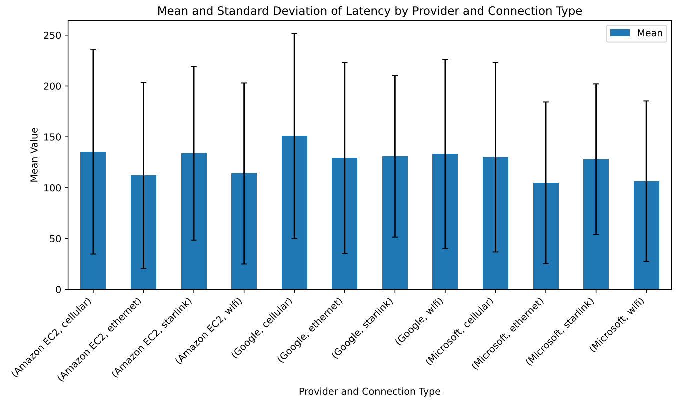


Fig. 3: Mean Worldwide Latency and Standard-deviation for Provider-Access-Technology-Pairs

by Microsoft operate slightly better than ones from Amazon or Google, regardless of the access technology.

So overall, we were able to observe that on average, Ethernet is the most reliable and best choice for low latencies. But on the other hand, it has a limited applicability for a vast amount of applications due to its requirement of a wired connection. Thus, developing and improving wireless connectivity is imperative. In this area, we found that WiFi performed best, with on average lower latency and fluctuation. Cellular and Starlink differ depending on the region.

B. Path Analysis

The path a packet follows from source to its destination leads through different intermediate routers on the way. These routers are typically owned by various organizations such as internet service providers (ISP) or content delivery networks (CDN). This raises the question of which routers a packet traverses on the way to its destination and to which organization they belong to.

Each router normally belongs to an autonomous system (AS) which is managed by an organization. The organization can be found out by a lookup for the AS number by IP addresses of intermediate hops. This reconstructs the AS path containing all ASes a packet traversed to its destination. The metric we consider here is the number of hops per AS. Further, we only count hops, for which the IP is resolvable to an AS number. Figure 6 shows a breakdown by datacenter providers.

As we can see, the number of hops spent in an AS heavily depends on the destination. Packets destined for Google spent the most hops in a Google AS, packets for Amazon in a Amazon AS and packets for Microsoft in a Microsoft AS.

Logically, the AS of the receiver appears in the AS path as it has to be traversed to reach the destination. But regarding this, we still have a big difference between the destination AS and other ASes. This bias may go back to the concepts of policy routing and hot potato routing. Each AS on the way wants to get rid of traffic destined for any location outside of the own AS. Consequently it routes them out of the own AS as fast as possible to decrease load on the own network as well as costs. As a result, the number of hops in each intermediate AS is kept as small as possible. As soon as the packet reaches the destination AS it is usually not routed through the AS to the destination which might be a major part of the whole route.

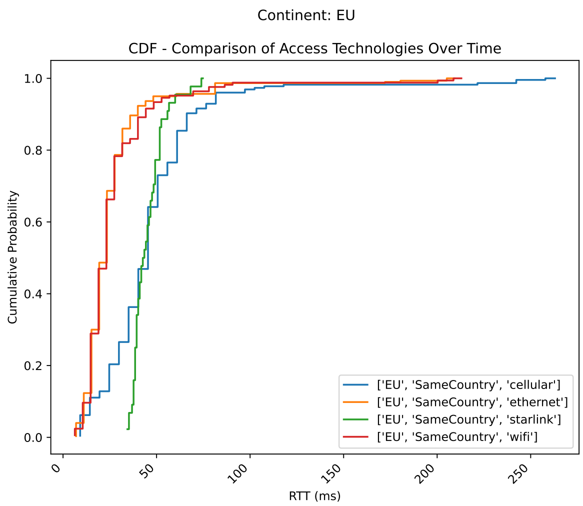
Besides that we can still observe differences among routing paths. The biggest difference of hops in the own and foreign ASes is at Microsoft, mainly using its own. There is no logic explanation for this observation, as Microsoft owns the fewest IP addresses of all three providers. Also, traffic routed to Amazon traverses other ASes for a longer time than Google or Microsoft do.

SpaceX AS has a special role here as it usually routes traffic via its own satellites and drops it down to earth only when it is already close to the destination. Thus the major routing is done, spending the most hops in its own network and only few in others. This might be the reason why it appears among the top 5 ASes.

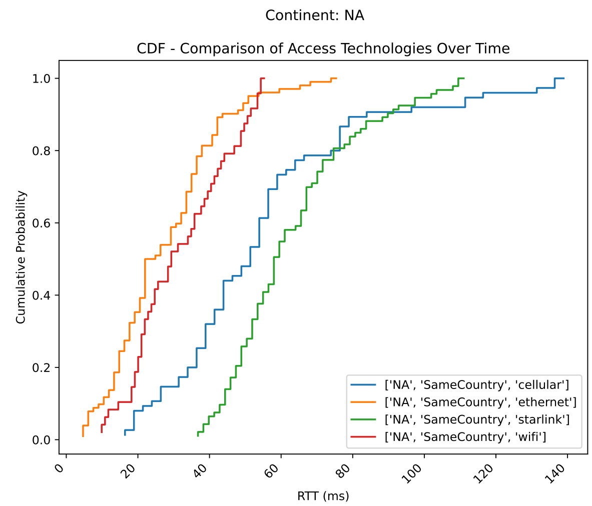
C. Undersea Connection Analysis

A special type of worldwide communication infrastructure represents intercontinental connection. While Europe and Asia are connected overland, Europe and North America are separated by sea. To provide also this type of intercontinental connection, cables are installed on sea floor. Such communication via undersea cables involves the transmission of data packets over long distances. Consequently, the end-to-end latency can underlie significant variations. In contrast to that, starlink satellites route traffic not via undersea cables but through their satellite network. This make it independent from terrestrial infrastructure. Figure 7 compares mean latency of wired connections and satellite between Europe and North America. We spare out wifi and cellular probes here, because the last mile technology is not relevant here.

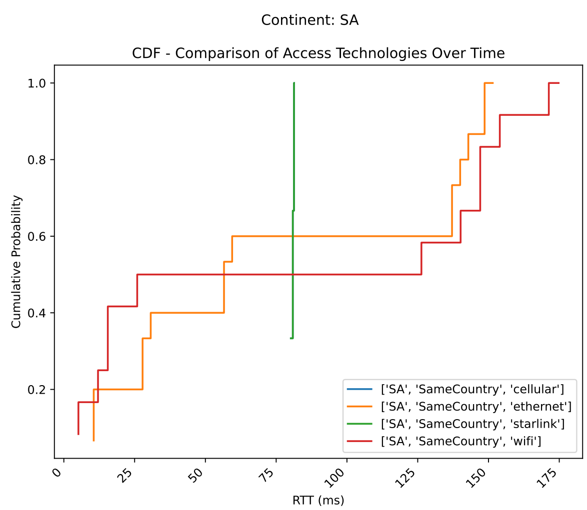
What we can observe from Figure 7 is that first, latency of wired connection is lower than latency of satellite connection. On average, wired latency is at about $108ms$ while satellite needs approximately $127ms$. The second observation we can make is that wired connections have lower jittering compared to satellite connection, except one outlier at the beginning of the measurements. Including this outlier, standard deviation is at $2.84ms$, without at $0.92ms$. Standard deviation of the satellite network is with $2.33ms$ a bit below original standard deviation of wired connection. This is probably due to wired technologies being the more stable alternative. Cables are reliant and are generally less prone to packet loss. As mentioned, the distance is between continents is far, meaning for wired infrastructure, there are many routers to traverse and there might be bottlenecks, including the sea cable itself. This might lead to exceptionally high latency at some times, but in generally being the more stable technology. Satellite instead routes traffic via LEO satellites through the starlink network. Connections between satellites are wireless and thus more prone to transmission failures and packet loss. LEO satellites are not stationary but follow their orbit and are permanently moving. Consequently, the route to the destination can be at some point of time more direct, and at another time taking more detours.



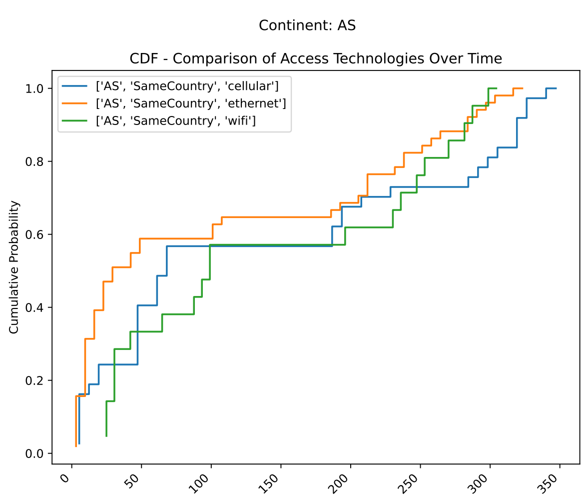
(a) Probes based in the EU



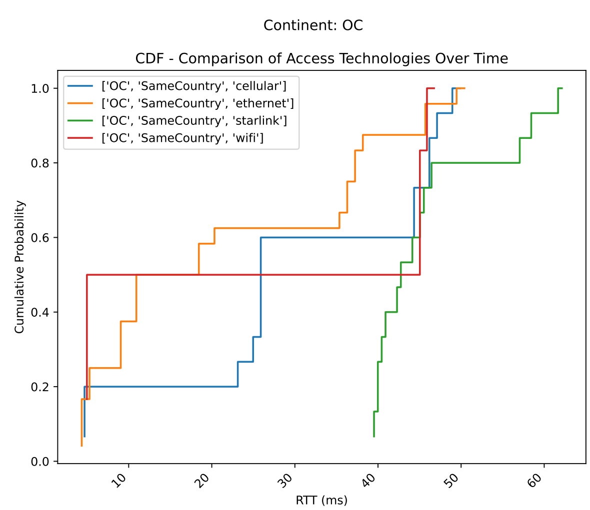
(b) Probes Based in North America



(c) Probes Based in South America



(d) Probes Based in Asia



(e) Probes Based in Oceania

Fig. 4: Probes Pinging on the Same Continent

Continent	Same Country	Neighbor Country	Neighbor Continent
Africa	-	-	100%
Asia	52.56%	100%	85.76%
Europe	74.69%	74.55%	99.3%
North America	86.64%	96.15%	99.77%
Oceania	39.92%	100%	99.84%
South America	-	-	100%

TABLE I: Percentages of pings which average rtt is larger than 30ms

D. Potential Benefits with Latency by Offloading Computation on Path

In this section, we consider the question that, if the wireless RIPE Atlas probes are actually mobile clients running a next-generation application requiring 30 ms end-to-end operational latency, will there be a potential benefit in offloading to an organization on path instead of offloading to the cloud datacenter deployed (i) in the same country as the probe(ii) in the neighboring country in the same continent and (iii) in the neighboring continent.

From the pings we can extract 'rtt', round-trip-time, which is the travel time of a packet from probe to datacenter and then back to the probe. Now it's crucial to discuss what is 'end-to-end operational latency' in this case, whether this latency only includes one way delay or round trip delay. We believe that when considering computation offloading, which meaning that the mobile client triggers a request to conduct certain computation on the server's side then computation result transmitted back to this mobile client, round trip latency matters here. Furthermore, end-to-end latency should also consider processing time on server's side. However as it is very difficult to estimate and this is obviously beyond the scope of this experiment, we will not consider it.

There does exist several cases that only one way delay should be consider. Take disaster warning system as an example, the centralized decision desk need to broadcast to all mobile user about the disaster and what actions needed, this information is only transmitted one way. But this does not fall within the definition of 'computation offloading', we will not consider this situation in analysis.

We labeled our ping results with these three different scenarios: from probe to datacenter which is in the same country as the probe, in the neighboring country of the probe, and in the neighboring continent of the probe. At this point we only considered probes which are connected by 'cellular'. We grouped the pings by continent as Figure 8, Figure 9 and Figure 10, and calculates how many percentages of them are higher than 30ms with Table I

We can assume that by offloading computation to an organization on path, the end-to-end latency will definitely be lower than 30ms. In this case, when the datacenter is deployed in the same country, less than half of clients in Oceania, half of clients in Asia, three thirds of clients in Europe and a majority 86% of clients in North America will benefits from this offloading. When the datacenter is farther away, deployed in the neighbor country but in the same continent, it's interesting to see that, almost the same percentage of clients as the last scenario can benefit from the offloading. While in other continents almost all clients will benefits from this. In the case which the datacenter is deployed in the neighbor continent, all clients around the world need the offloading to reach 30ms end-to-end latency.

To conclude, from our measurement we can see that, there's great potential benefits with offloading computation to an organization on path from our measurement. However, considering the actual density of datacenters are higher than those we can choose in the provided list, latency to datacenters in the real world could be much lower. To evaluate the benefits more precisely, further measurements should be done by finding and choosing datacenters at a finer geographic scale.

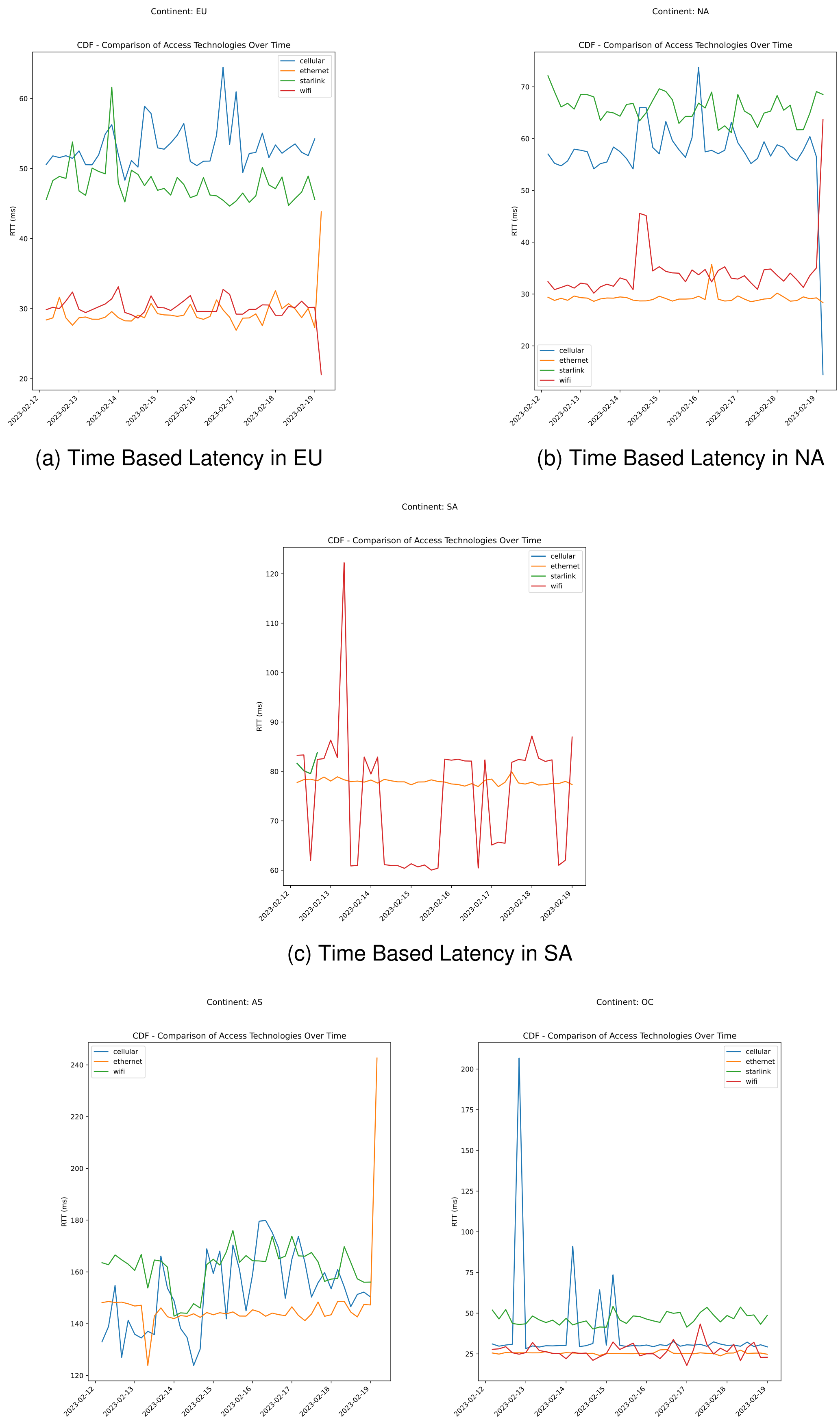
V. Conclusion

At the beginning, we introduced the necessity of lower latency in order to progress technical advancements in both consumer and user applications. Our experiments and the following analysis have shown, that especially Europe and Northern America facilitate the capabilities to allow for latency sensitive applications. Nevertheless, as we have seen, according to our results, this is more suitable in environments with available WiFi or Ethernet connection. Especially the relatively high fluctuations of latencies in Cellular data can hinder the real-world applicability critically time-sensitive applications such as offloading computation for autonomous driving. On the other hand, the results from the measurements from Asia were out of our expectations. A possible explanation for this is, that when thinking of Asia, mostly well developed and densely populated countries such as Japan or Korea come to mind. While Asia is home of such highly developed societies, it is also a continent with unimaginably vast masses of uninhabited or sparsely populated land. We also saw that routing paths vary according to the destination. Policy and hot potato routing cause the packet to be sent to the destination AS as soon as possible and make the most hops there. Comparing wired undersea cables with satellite connection between Europe and North America, we figured out that wired connections are generally more stable but might have serious outliers. Satellite connections are show higher deviations in the mean but no outliers beside that. This is due to the nature of their connections. While wired connections are in general more stable and less prone to packet loss, they might experience bottlenecks or exceptional situations due to the long distance. Satellite technology is less prone to this issue, since the network is more homogeneous. But since it depends on wireless transmission of messages it is more prone to packet loss. At the same time, there exists great beneficial when offloading computation to an organization on path to reach the 30ms end0to-end latency goal.

References

[1] TeleGeography. (2022) Global internet map 2022. [Online]. Available: https://global-internet-map-2022.telegeography.com/

[2] B. Augustin, X. Cuvellier, B. Orgogozo, F. Viger, T. Friedman, M. Latapy, C. Magnien, and R. Teixeira, "Avoiding traceroute anomalies with paris traceroute," in *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, 2006, pp. 153–158.



(a) Time Based Latency in EU

(b) Time Based Latency in NA

(c) Time Based Latency in SA

(d) Time Based Latency in AS

(e) Time Based Latency in OC

Fig. 5: Results of Latencies over Time Worldwide

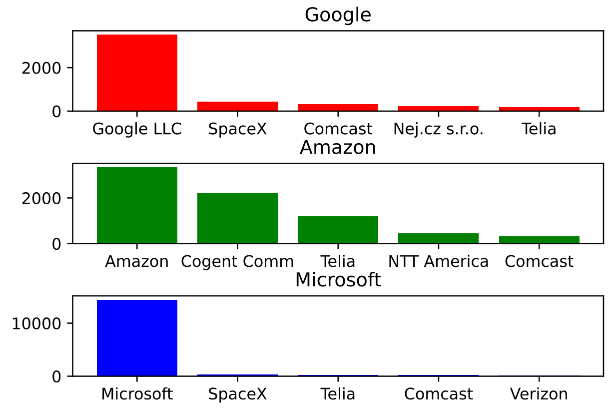


Fig. 6: Number of Hops in Top 5 AS according to Datacenter Providers

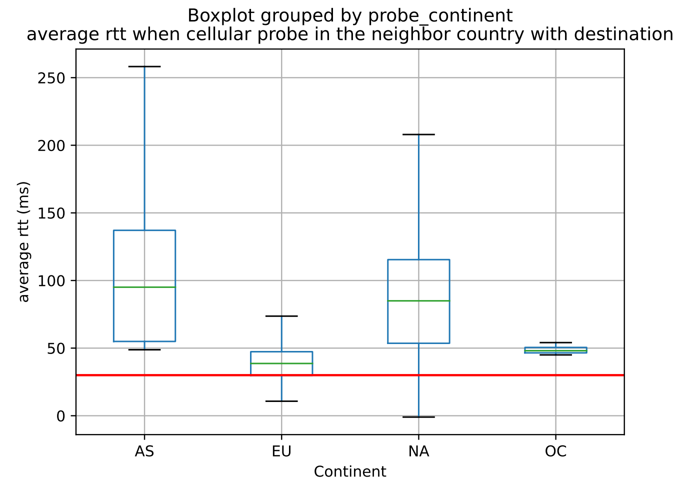


Fig. 9: Average rtt when probes are in the neighbor country of destination

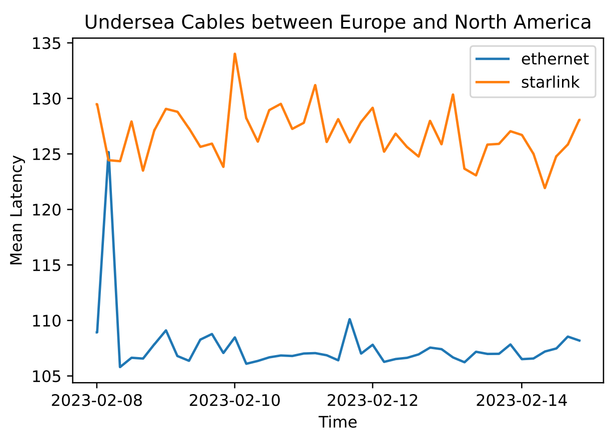


Fig. 7: Undersea Cable between Europe and North America

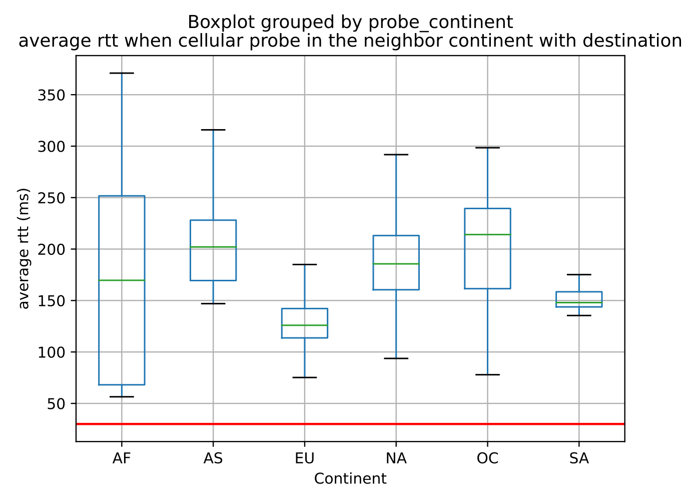


Fig. 10: Average rtt when probes are in the neighbor continent of destination

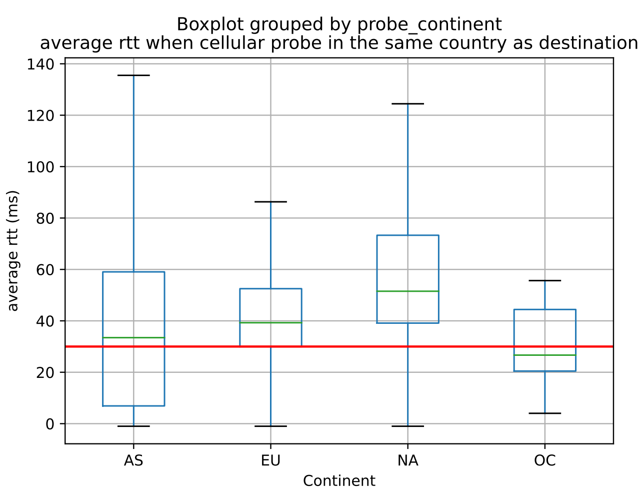


Fig. 8: Average rtt when probes are in the same country as destination