Surrounded by the Clouds

A Comprehensive Cloud Reachability Study

Lorenzo Corneo[♯] Maximilian Eder[♮] Nitinder Mohan[♮] Aleksandr Zavodovski[♯] Suzan Bayhan[♭]
Walter Wong[♭♭] Per Gunningberg[♯] Jussi Kangasharju[♭♭] Jörg Ott[♮]

[♯]Uppsala University [♮]Technical University of Munich [♭]University of Twente [♭♭]University of Helsinki

ABSTRACT

In the early days of cloud computing, datacenters were sparsely deployed at distant locations far from end-users with high end-to-end communication latency. However, today's cloud datacenters have become more geographically spread, the bandwidth of the networks keeps increasing, pushing the end-users latency down. In this paper, we provide a comprehensive cloud reachability study as we perform extensive global client-to-cloud latency measurements towards 189 datacenters from all major cloud providers. We leverage the well-known measurement platform RIPE Atlas, involving up to 8500 probes deployed in heterogeneous environments, e.g., home and offices. Our goal is to evaluate the suitability of modern cloud environments for various current and predicted applications. We achieve this by comparing our latency measurements against known human perception thresholds and are able to draw inferences on the suitability of current clouds for novel applications, such as augmented reality. Our results indicate that the current cloud coverage can easily support several latency-critical applications, like cloud gaming, for the majority of the world's population.

CCS CONCEPTS

• **Networks** → **Network measurement**; *Public Internet*.

KEYWORDS

Cloud reachability; Internet measurements

ACM Reference Format:
Lorenzo Corneo, Maximilian Eder, Nitinder Mohan, Aleksandr Zavodovski, Suzan Bayhan, Walter Wong, Per Gunningberg, Jussi Kangasharju, and Jörg Ott. 2021. Surrounded by the Clouds: A Comprehensive Cloud Reachability Study. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3442381.3449854

1 INTRODUCTION

During the last decade, cloud computing has gained a central role in many networked services over the Internet. According to Gartner [13], the cloud services market grossed $242.7 billion in 2019 and is expected to grow by 6.3% in 2020. This computing paradigm became so popular due to its ability to provide seemingly unlimited storage and computational capabilities through its highly efficient and optimized hardware infrastructure. In the early days of cloud computing, commodity equipment could not compare to datacenters' powerful hardware as it would have required considerable purchasing expenses. Cloud computing provided a way to reduce complex computation times dramatically. Additionally, the storage functionality allowed users to synchronize personal data over multiple devices. Cloud computing, through its appealing and flexible pricing models (e.g., pay-as-you-go [12, 15] and transient virtual machines [32]), relieves businesses, institutions, and individuals from equipment investments for storage and computation.

Since 2009, cloud computing has been challenged by the advent of edge computing, a new computing paradigm that has become very popular and well received by both industry [4] and academia [24, 30]. In fact, the research community started questioning the general applicability of cloud computing with respect to emerging enabling technologies and novel applications, such as augmented reality, industrial Internet of Things, etc. The primary motivating assumption within the edge computing community is rather long end-to-end cloud access latency due to limited and sparse deployment of datacenters across the globe. As a result, next-generation networked applications cannot meet their latency requirements while operating over the cloud infrastructure.

However, since 2009, several trends in networking and IT have drastically changed the reach of cloud computing. Cloud providers have expanded their geographical coverage by extensively establishing cloud regions in different parts of the globe while preserving their key success enabler – economies-of-scale. For example, Amazon's cloud network has expanded from 3 to 16 countries, with 22 newly built datacenters, over the last decade. Consequently, cloud providers can now support computationally complex tasks, such as voice assistance services like Siri or Cortana, without noticeable delays. Moreover, a number of recent latency-critical applications, e.g., cloud-based gaming [14, 22], backed by major cloud providers, are already available in the market. Such offerings largely rely on the throughput of the underlying networks, which continue to show steady growth.

We believe that, driven by the enthusiasm for newer computing paradigms, both practitioners and researchers of edge computing may have missed the significant efforts of cloud providers to become more and more pervasive towards the end-users. Interestingly, very little attention has been paid to quantify the reach and (consequently) applicability of current cloud infrastructure for latency-critical applications. Related works on this subject are either too dated [20] or focus on a single cloud provider [19]. In this work, we fill this chasm by expanding on our previous work [23] and present a comprehensive global cloud reachability study – aimed to estimate the cloud access latency and the path length between

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449854

end-users and the datacenters. The key contributions we make in this paper are as follows:

(1) We conduct a large-scale measurement study over the span of 12 months, analyzing the reachability of ten major cloud networks – totaling 189 datacenters deployed in 28 countries. We use more than 8000 RIPE Atlas probes deployed in 184 countries to periodically measure user-to-cloud latency (`ping`) and path length (`traceroute`) over ICMP and TCP. Our collected dataset reaches almost 60 GB in size and is publicly available at [9].
(2) We analyze the spread of the clouds globally and identify their suitability for deploying latency-critical applications in the cloud. We do this by comparing the obtained latency distributions against three well-known timing thresholds, namely, human reaction time, human perceivable latency, and motion-to-photon. Throughout the paper, we compare our results with these timing thresholds to provide an application-centric perspective. Further, we take a closer look at the cloud reachability in the US and in Asia: two regions with different datacenters coverage as well as different networking infrastructures.
(3) We conduct a thorough user-to-cloud path analysis to showcase the extent of cloud pervasiveness over the Internet. Our results show that cloud providers that deploy their private wide area network (WAN) exhibit a high level of cloud pervasiveness. Conversely, cloud providers that depend on public Internet have a low level of pervasiveness. Furthermore, while we find latency differences between private and public WANs to be comparable, providers on the public WAN deliver higher latency variation, when compared to providers with their own network infrastructure.
(4) Supported by our large scale dataset, we also present a plausible road map for future cloud deployment strategies. Our findings show that cloud deployment in continents such as North America, Europe, and Oceania would bring little benefit to the end-user latency because of the existing high density of datacenter deployments. In contrast, Asia, South America, and Africa can benefit greatly from increased cloud deployment and show the largest potential in latency gains.

The remainder of this paper is organized as follows. Section 2 describes the three latency concepts we use to analyze the performance of cloud in meeting the latency requirements. Section 3 introduces our measurement methodology, while Section 4 presents our findings on cloud reachability via our measurements. It is worth noting that this work and the dataset we collected does not raise any ethical issues. In Section 5, we discuss the implications of our study as well as its limitations, while we provide the related work in Section 6. Section 7 concludes our paper.

2 BACKGROUND

In this section, we describe three well-known timing thresholds that we use to quantify the level of cloud reachability across the world. We match these thresholds to the requirements of current and future networked applications that demand strict latency requirements for operation. As a result, we study cloud reachability from the perspective of understanding if current cloud infrastructure is a feasible option for supporting near-future applications.

(1) Motion-to-Photon (MTP) is the delay between user input and its effect to be reflected on a display screen. MTP is guided by the human vestibular system, which requires sensory inputs and interactions to be in complete sync, failure of which results in motion sickness and dizziness. Maintaining latency below MTP, i.e., $\lesssim$ **20 ms**, is key for immersive applications, such as AR/VR, 360° streaming, etc. [21]. Of this, $\approx$ 13 ms can be taken up by the display technology due to refresh rate, pixel switching, etc., which leaves a budget of $\approx$ 7 ms for computing and rendering (including RTT to compute server) [7].

(2) Human Perceivable Latency (HPL) threshold is reached if the delay between user input and visual feedback becomes large enough to be detected by the human eye [27]. HPL threshold plays a key role in the QoE of applications where the user interaction with the system is fully or semi-passive, e.g., video streaming (stuttering), cloud gaming (input lags), etc. HPL is roughly estimated to be **100 ms**.

(3) Human Reaction Time (HRT) is the delay between the presentation of a stimulus and the associated motor response by a human. While HRT is highly dependent on the individual (and can be improved by training), its value is reported to be $\approx$ **250 ms** [37]. Latencies for applications that require active human engagement, such as remote surgery, teleoperated vehicles, etc., must operate within HRT bounds.

3 MEASUREMENT METHODOLOGY

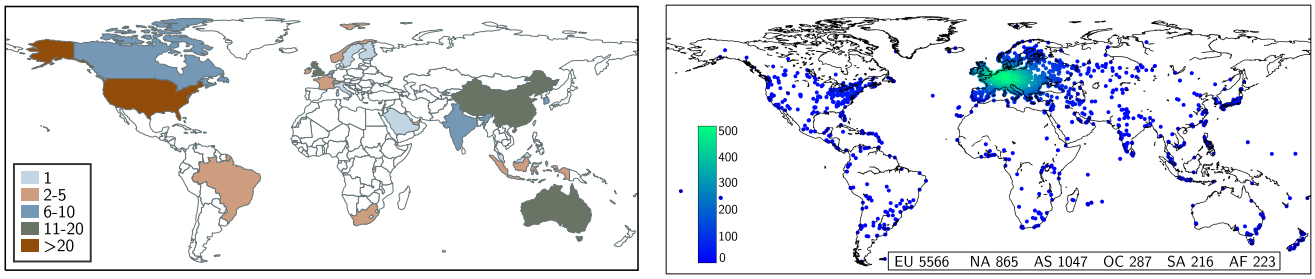
In this section, we describe our methodology for measuring cloud reachability across the globe. We begin by introducing our selection criteria for targeted datacenters and vantage points, followed by a description of our experiments.

3.1 End-Points Selection

We chose datacenters from *nine* different cloud providers as endpoints, namely, Amazon, Google, Microsoft Azure, IBM, Oracle, Alibaba, Digital Ocean, Linode, and Vultr. For Amazon, we chose both its EC2 and Lightsail offerings. The chosen operators are widely used, well-established, and provide global coverage with a distinct infrastructure, that is, their backbones could be either private or public. For every cloud provider, we retrieved the host name of a public virtual machine hosted by CloudHarmony [8]. In total, our dataset includes 189 cloud region end-points as targets, the geo-distribution of which is shown in Figure 1a. Moreover,

Table 1: Global density of cloud provider endpoints, and their backbone network infrastructure type used in our measurements.

	Datacenters per continent						Backbone
	EU	NA	SA	AS	AF	OC	N/W
Amazon EC2 (AMZN)	6	6	1	6	1	1	Private
Google (GCP)	6	10	1	8	-	1	Private
Microsoft (MSFT)	12	10	1	11	2	4	Private
Digital Ocean (DO)	4	6	-	1	-	-	Semi
Alibaba (BABA)	2	2	-	16	-	1	Semi
Vultr (VLTR)	4	9	-	1	-	1	Public
Linode (LIN)	2	5	-	3	-	1	Public
Amazon Lightsail (LTSL)	4	4	-	4	-	1	Private
Oracle (ORCL)	4	4	1	7	-	2	Private
IBM (IBM)	6	6	-	1	-	-	Semi
Total	**50**	**62**	**4**	**58**	**3**	**12**	



(a) Distribution of datacenters operated by nine major cloud providers (refer to Table 1 for per-provider distribution).

(b) Distribution of 8000+ RIPE Atlas probes used in our measurements.

Figure 1: Global coverage of our measurement setup. Cloud datacenters in (a) represent our endpoints, and RIPE Atlas probes in (b) are the vantage points for our measurements.

Table 1 shows the distribution of the datacenters in our dataset by cloud provider and deployed continent.

Besides global coverage, cloud performance is also heavily influenced by the network between users and datacenters, and between datacenters. Some providers, e.g., Linode, have set up their datacenters as independent "islands" and largely rely on the *public Internet* for inter-datacenter connectivity. On the other hand, providers such as Amazon have set up *private*, large-bandwidth, low latency network backbones to interconnect all their datacenters [31]. Additionally, several cloud providers also sign agreements with major ISPs operating globally to enable direct peering between the ISP gateway and their private point-of-presence (PoP) [2]. This allows end-users to avoid the public Internet completely while transiting to cloud network. For instance, a recent study shows that Google peers with more than 5700 ASes around the globe, and the number has been increasing consistently every month [6]. Table 1 also lists whether a cloud provider has a fully-private (*Private*), private within a continent (*Semi*), or a public Internet based (*Public*) network backbone.

3.2 Vantage Points Selection

Our vantage points are probes from the RIPE Atlas platform [33], which is a *de-facto* standard for conducting measurements within the network research community. RIPE Atlas is a global Internet measurement network, especially used for reachability, connectivity, and performance studies. The platform includes thousands of small hardware probes[1] connected to the Internet all over the globe. Users can perform active network measurements (`ping` or `traceroute`, etc.) using these probes to end-points of their choice. Atlas probes are installed in heterogeneous network environments, such as core, access, or home networks, allowing us to analyze the reachability of cloud datacenters globally. Despite Atlas's dense deployment, many of the probes are hosted by cloud and network providers – allowing them to monitor their network reachability from outside their infrastructure [3]. Since these probes do not reflect the user connectivity and have the potential to add bias to our measurements, we manually filter out all such probes from our measurements using their user-defined tags [29] (e.g., `datacentre`, `us-east*`, `us-west*`, `gcp`, and `aws`, etc.). This left us with more than 8000 probes distributed in 184 countries across the globe. Figure 1b shows the geo-distribution of the probes used in our experiments. The majority of the selected probes are located in Europe and North America (33.5% and 26.5%), which allowed us to exhaustively analyze the performance of the bulk of datacenter deployment on the same continents.

[1]RIPE Atlas now also integrates software probes but they were not yet available at the time this study was carried out.

3.3 Experiments

Our objective was to analyze *two* key aspects of cloud reachability: (i) *user-to-cloud latency* and (ii) *path lengths*. Both our experiments ran in parallel from *September 2019* to *September 2020*, resulting in an almost 60 GB dataset. Our collected data is publicly available at [9].

(i) Latency Estimation. We estimate end-to-end latencies between users and cloud datacenters via `ping` measurements. We configured the Atlas probes to `ping` all available datacenters within the same continent every 3 hours throughout the measurement period. For probes in continents with low datacenter density, e.g., Africa and South America, we also included `ping` latencies to datacenters in adjacent continents, i.e., Europe and North America, respectively. We augment the latencies from ICMP-based pings by those from TCP traceroute, see (ii).

(ii) Path Length Estimation. We estimate the end-to-end distance (as hop count) between users and cloud datacenters via `traceroute` measurements (§4.2) repeated on a daily basis. In addition to ICMP-based traceroutes, we also launched `TCP traceroute` and record per-hop latency. Unlike ICMP, our TCP measurements are guaranteed to be end-to-end and provide us with an accurate representation of connection latencies encountered by real applications operating in the cloud. The latency in the last-hop of TCP `traceroute` characterizes probe-to-VM RTT, which we use to augment our latency measurements from `ping`. Unlike our latency measurement setup, we configured Atlas probes to record `traceroutes` towards *all* datacenter endpoints. As a result, we were able to identify many unique paths from users to cloud in our resulting dataset - specifically more than 450,000 in the US, 8345 in South America, nearly 3 million in Europe, over 630,000 in Asia, and 6880 in Africa. We removed any unresponsive hops and private IP addresses in our

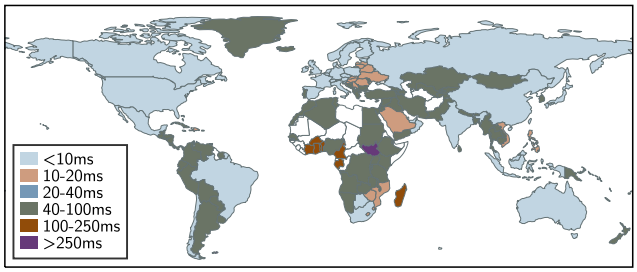


Figure 2: Minimum latency to datacenters observed across the globe.

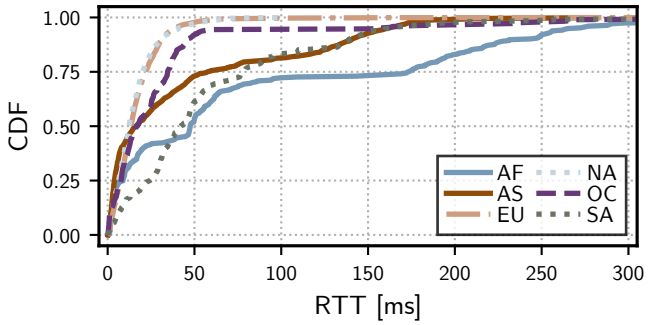


Figure 3: Distribution of minimum RTT by all probes to the nearest datacenter grouped by continent.

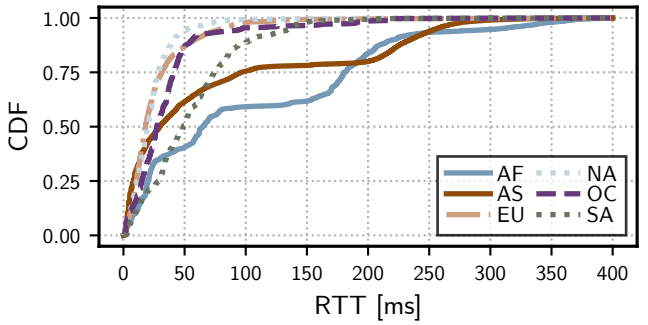


Figure 4: Distribution of all RTT values recorded from all Atlas probes in our dataset to the closest datacenter.

processing phase since they depend only on the internal LAN configuration and are not part of the public Internet.

(iii) Network composition. To further quantify the footprint of cloud providers and identify several organizations that operate on a user's path to the cloud, we first map the Autonomous System Number (ASN) with IP address of every hop recorded in our `traceroute` measurement using PyASN [16]. Further, we query PeeringDB [26] dataset, which provides us with the name, location, and network type of organizations operating on the path.

Experiment configuration. In order to ensure that our analysis is statistically significant, we calculate the minimum measurement sample size required for each country. We define the required confidence interval for the measurement as $n = \frac{z^2 \times \hat{p}(1-\hat{p})}{\epsilon^2}$, where z is the *z-score*, $\hat{p}$ is the population proportion, n is the target sample size, and ϵ is the margin of error. Therefore, for an interval of confidence of 95% and an error of $\epsilon = 2\%$, we collect at least 2400 measurements per country. Furthermore, while comparing the end-to-end latencies from our ICMP and TCP measurements, we found ICMP values to be consistently larger than TCP. This was the case in Asia, Europe, Oceania and South America. On the other hand, TCP exhibited larger distribution in Africa an North America, even though the median RTT is comparable with ICMP. We believe this happens because ICMP packets are often treated as low priority by cloud organization's firewall and can be treated differently than regular application packets (like HTTP), which use TCP as the underlying protocol. Therefore, while our TCP measurements closely mimic application connectivity latencies, our ICMP-based measurements represent the worst-case connectivity between user and cloud. A more extensive comparison of the differences between TCP and ICMP is left for future study.

4 MEASUREMENTS ANALYSIS

In this section, we offer a two-fold analysis mainly addressed to estimate the pervasiveness of cloud datacenters, from the point of view of access latency and access path length.

4.1 Cloud Access Latency

The Potential. We begin by showcasing *the least possible latency for a user to access the closest datacenter across the globe.* We extract the minimum `ping` latency observed by the best-performing probe for every country to any cloud datacenter. Figure 2 shows the distribution of latency per country as a heatmap. The results show that 45 countries can access a cloud datacenter with RTTs less than 10ms, and 21 countries with RTTs between 10 to 20 ms. While our results are "optimistic" – in that they show the minimum latency – they also indicate that the cloud potentially is able to provide latency within the boundaries of MTP to the majority of the world. We will revisit this issue later in the paper.

Our findings become more intuitive when considering the density of datacenters across the globe, as shown in Figure 1a. Countries with access latency less than 10 ms typically have a local datacenter (one or more), which offers very low latency if accessed from a managed (public or otherwise) network. Some countries, such as the US, UK, Japan, and India have more than one datacenter deployed by the same provider. Countries with access latencies less than 20 ms either share borders or have direct fiber connectivity [34] to the country housing a datacenter. Out of the rest, 49 countries have latencies between 20-40 ms and 53 between 40-100 ms. Note that probes in all *but* 16 countries (majorly in Africa) can potentially access a cloud datacenter within HPL bounds (100 ms).

Figure 3 depicts the smallest latency distribution experienced by every probe to any datacenter, grouped by continents. Note that while the measurements are largely from probes deployed in home networks, they also include probes which may not have a stable Internet connection. Despite this, the results look quite in favor of the cloud and support the findings in Figure 2. Around 80% of probes in Europe and North America, which is around 50% of the total

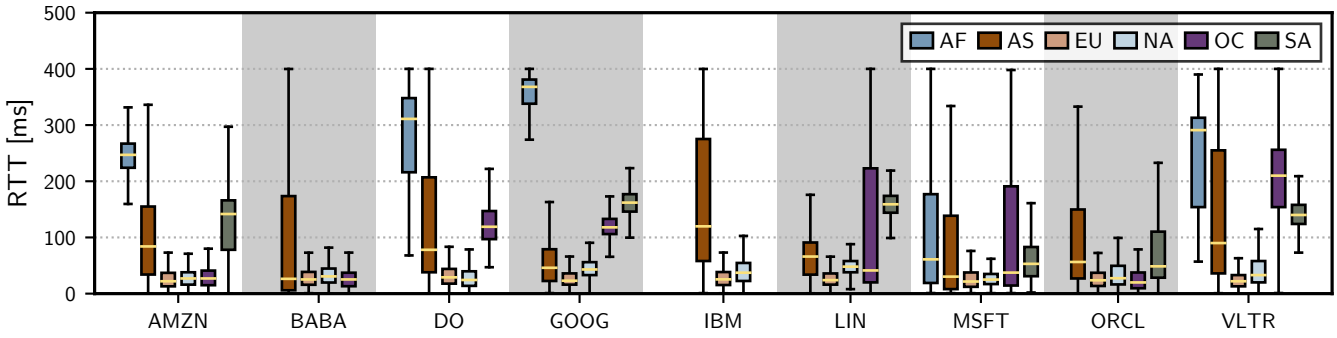


Figure 5: Cloud access latency to the closest cloud datacenter of every provider in different continents.

number of probes used in our experiments, can access a datacenter within 20 ms. Probes in Oceania follow a similar performance pattern as almost all of them can access the cloud within 50 ms RTT. Surprisingly, despite the low density of available datacenters and substandard network deployment, 75% of probes in Africa and Latin America achieve less than 100 ms access latency to the cloud, thereby meeting HPL requirements. Almost all, but a few probes in Africa, can reach the cloud within HRT threshold (i.e., 250 ms).

Takeaway — 168 countries out of a total of 184 in our dataset can support applications bounded by human perception. All probes (excluding long tails) in North America, South America, Europe, and Oceania can reach the cloud within 90 ms. Moreover, slightly more than 75% of the probes in Asia and Africa satisfy the HPL threshold.

The Reality. Till now, our latency analysis focused on the best-case scenarios to illustrate the potential reach of the cloud. We now turn to our entire latency dataset to showcase what the reality of cloud reachability is today. Figure 4 shows a comprehensive view of our latency dataset. We show the distribution of *all* latencies observed by probes to their nearest datacenter throughout our measurements duration of 12 months.

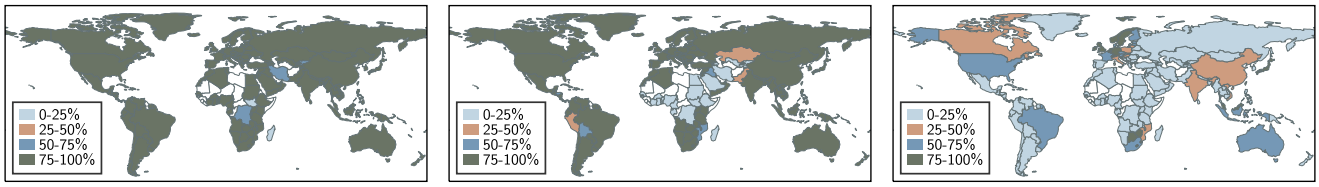
It is evident that probes in North America, Europe, and Oceania exhibit excellent cloud reachability. *More than 75% of the total probes* in all three continents have RTT to the cloud within the HPL. A closer look reveals that the top 25% of connected probes in North America and Europe can reach the cloud within the bounds of MTP latency, indicating that the cloud can support emerging applications such as AR/VR and autonomous vehicles. The reason for this exceptional performance, as made evident from the previous section, is the concentrated efforts of cloud providers to deploy datacenters throughout the US and central Europe. Additionally, thanks to the vast number of ISPs operating in these two continents, the majority of users can consistently connect to the cloud via high-bandwidth, low-latency fiber connections. However, note the long tail of latency distribution for Europe, which is largely missing from North America. The cause is the absence of a local datacenter or high latency to connect to the one located in a neighbouring country. The result is in line with our initial assessment of Figure 2, where the bulk of countries exhibiting high access latencies did not have a datacenter in close proximity.

We now focus on the remaining continents, i.e., South America, Asia, and Africa. Cloud reachability from within these continents is quite poor as only a fraction of probes are able to satisfy the 100 ms HPL threshold. Probes in Asia show very diverse latencies primarily due to scattered datacenter deployment favoring certain countries, like China and India. Unsurprisingly, the worst performance is in Africa as the continent is severely under-served, both in terms of cloud presence (only three operating datacenter in South Africa) and reliable network infrastructure [5].

Takeaway — North America, Europe, and Oceania easily satisfy the HPL, and almost 25% even support MTP latency. On the other hand, Asia, South America, and Africa show considerably longer latencies to the cloud due to a lack of extensive cloud and network infrastructure deployment.

Wide Area Network Latency Differences. We now assess the impact of network backbone infrastructure on cloud reachability performance. As previously summarized in § 3.1, many cloud providers deploy extensive private wide area network (WAN) to interconnect their datacenters that provide clients fast-track paths to services hosted in their infrastructure. Table 1 enlists the network backbone type used by different cloud providers targeted in our measurements. Figure 5 shows the distribution of the latencies achieved by Atlas probes, at continental granularity, towards the closest datacenter of every cloud provider. Since the aim of this work is not to provide a benchmark study comparing the performance of different cloud operators across the globe, we do not probe all cloud regions in this analysis. Instead, we only draw results from those regions which were found closest (in latency) to our vantage points.

From the figure, it is evident that the availability of private network backbone in continents with extensive network deployment, like North America and Europe, does not seem to have much impact on cloud reachability. In fact, we find that all cloud providers exhibit similar latency distributions in these regions – accentuated more towards providers relying on the public Internet. In Oceania, Amazon EC2, Alibaba Cloud, and Oracle achieve the least latency results while Microsoft Azure and Linode perform similarly but with higher variance. We justify their superior performance to their extensive deployment in the continent. Within Asia, almost all providers perform similarly, and we do not observe any significant



(a) Probes RTT within HRT. (b) Probes RTT within HPL. (c) Probes RTT within MTP.

Figure 6: Global coverage of our measurement with respect to the three timing thresholds defined in §2

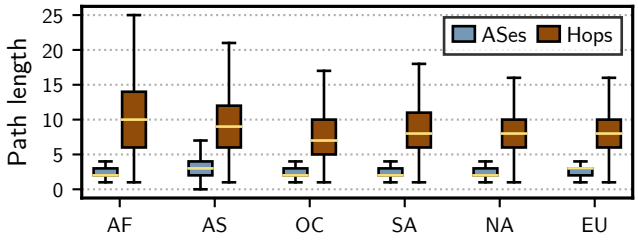


Figure 7: Path length to the closest cloud.

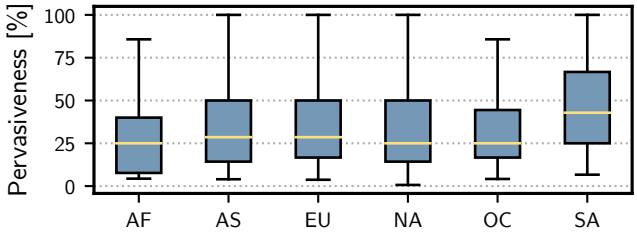


Figure 8: Per-continent cloud pervasiveness.

benefits favoring providers with private WAN and those relying on public Internet. For South America, we probed Amazon EC2, Microsoft, and Oracle since only those have datacenters deployed within the continent. For the rest of the providers, we show latencies from South American probes to their datacenters in North America. We observe that cloud providers with local datacenter deployment perform significantly better than those with infrastructure in the neighbouring continent. A similar trend can also be observed in Africa, where providers with in-land datacenter deployment (specifically Microsoft and Amazon EC2) show much lower latency than their counterparts, which host datacenters in the neighbouring continent of Europe. It is to be noted that we draw our inferences from small-sized ICMP and TCP packets, and the impact of private backbone will be far more significant for elephant flows within the cloud infrastructure.

Takeaway — The impact of private WAN availability on cloud reachability is not as significant as otherwise assumed. In continents with dense network deployment, public Internet delivers almost similar performance compared to a cloud provider that deploys its own private network backbones. Moreover, the availability of datacenters within a continent impacts connectivity far more than the type of interconnecting network infrastructure.

Cloud Application Readiness. We conclude this section by investigating *cloud maturity level* – the state of global cloud connectivity (at country-level granularity) to achieve the timing thresholds discussed in §2, i.e., MTP, HPL, and HRT. Figure 6 illustrates the global RTT distribution from all probes in our dataset, one for each timing threshold. Different color groups denote different percentiles of the distribution. The results suggest that almost every country across the globe can consistently reach the closest cloud datacenter within the boundaries of the HRT. In fact, only two (out of 184) countries in our dataset achieve the HRT less than 25% of the times, and three countries lie between 50 and 75%. For HPL, we observe that the cloud maturity level changes slightly compared to the HRT. The distribution of RTTs degrades only certain countries, mainly clustered in Central Africa, the Middle East, and South America. Specifically, 140 countries achieve RTTs consistently within the boundaries of the HPL, six achieve that only 50 to 75% of the times, another six countries only 25 to 50% and, 16 countries fail to reliably meet the HPL threshold. The distribution changes substantially for the MTP threshold, where only 24 countries can consistently meet the timing deadline (75–100%). Conversely, 125 countries outrightly fail to meet the threshold (0–25% of the sample), while the remaining 25 countries can reach cloud within MTP between 25–75% of the times.

Takeaway — The current cloud infrastructure is able to deliver network latency compliant with both HRT and HPL safely. However, only a small minority of countries reliably meet the MTP threshold suggesting that either cloud deployment or network should be improved.

4.2 Cloud Access Path Length

Distance to the Cloud. We complement our latency analysis in the previous section by investigating path lengths to the cloud. The study allows us to better understand the state of user-to-cloud connectivity from different parts of the globe over the Internet. We exploit our `traceroute` measurements (see §3.3 (ii)) and extract distance from probe to the closest cloud datacenter (in terms of routers), and organizations operating those routers (in terms of ASNs). We derive the latter by mapping the IP addresses of on-path routers to their ASN numbers and further correlating them with distinct organizations using PeeringDB [26]. Figure 7 shows our results, specifically the number of routers and ASNs on a path between a probe and its nearest datacenter in every continent. Our key findings are as follows. End-user paths to the cloud can

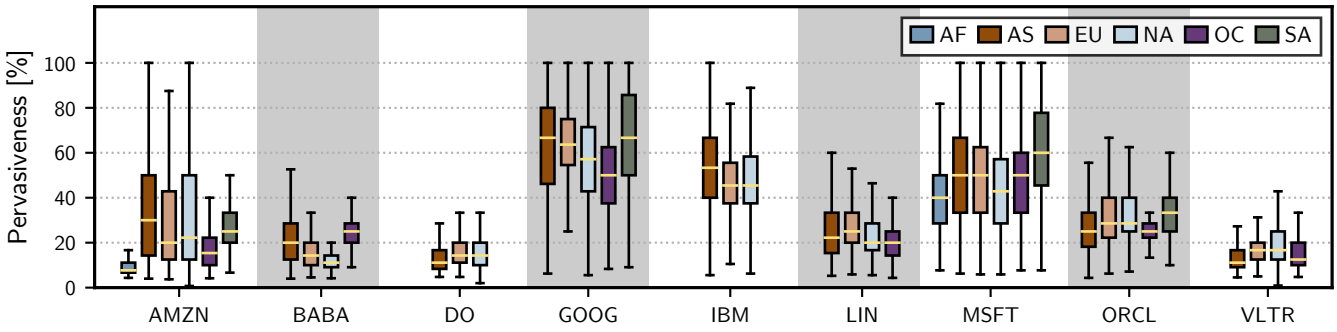


Figure 9: Degree of the pervasiveness of the cloud providers in different continents.

range anywhere between 7–10 hops on average and are shorter in continents with extensive cloud presence (e.g., NA and EU). However, most of these routers belong to a very small set of ASNs; usually, the resulting paths connecting these routers are managed by large network operators as well as cloud providers, and are highly optimized. The largest chasm between the number of routers and ASNs exists in Africa, showing the presence of long but managed paths to the cloud (some even traversing long transatlantic links to connect to datacenters located in NA). Overall, across the globe, a typical user can traverse 3–4 ASNs, on average, before reaching the nearest cloud region.

Takeaway — End-user path to the cloud is still quite long (in terms of the number of hops). However, these long paths are operated by a handful of organizations showcasing a highly managed state of cloud network connectivity.

Pervasiveness of the Cloud. As we are interested in understanding the degree of the pervasiveness of cloud networks, we hereby investigate *how much of the user-to-cloud path is owned by cloud providers*. We define *cloud pervasiveness* as the ratio between the number of routers owned by the cloud providers and the overall path length to the cloud. High pervasiveness indicates that the cloud providers are very close to the end-users and, conversely, a low ratio translates into cloud providers being faraway.

Figure 8 shows the extent of cloud pervasiveness of the closest datacenter for all continents. We can observe that the cloud providers already own 20–40% of the path user-to-cloud on average. It is also quite common for the cloud to own and operate upwards of 50% of the path, often reaching 100% in some cases. This denotes that the first public IP address encountered by the probe is entry to the cloud network. We found this phenomenon to be a common occurrence for probes installed in cities with a colocated datacenter. On the other hand, as noted in Table 1, some cloud providers rely on the public Internet and do not own a private WAN. Consequently, these providers only own the final hop, thus pushing the distribution towards the lower end. To understand this to greater detail, we now investigate the impact of private and public WANs on cloud provider's pervasiveness.

Figure 9 depicts the continental cloud pervasiveness grouped by providers. The figure provides immediate visual feedback for distinguishing providers with private WANs from those relying on the public Internet. Providers using the latter have a level of pervasiveness – constantly below and capped at 50%. On the other hand, Amazon, Google, IBM, and Microsoft exhibit a high degree of cloud pervasiveness by abundantly owning a majority of routers in the paths user-to-cloud. Note that the distributions of Amazon through Africa, Oceania, and South America are skewed as those measurements also target its datacenters in neighboring continents.

Takeaway — Cloud connectivity has become highly pervasive across the globe, with providers installing managed network infrastructure and establishing peering agreements with ISPs in the region. Of these, providers that make use of privately owned networks exhibit a high degree of pervasiveness. Conversely, providers relying on the public Internet have a low level of pervasiveness.

4.3 Cloud Access Case Studies

Case Study A: The United States of America. We now investigate the extent of cloud reachability by users within the United States of America. We find the US as a good object of study since it covers a large geographical area, has a large population, and has remained the focal point for major cloud providers – as reflected by the dense cloud presence within the country (Table 1). We selected the most populated regions in the US using the US Primary Statistical Areas (PSA) [10, 36]. The federal government has defined 100 PSAs, which collectively house more than 80% of the total US population. We further selected 93 PSAs (7 PSAs did not have any functioning RIPE Atlas probe) and collected up to 25 probes within a radius of 125 km from the center of PSA location. Overall, we selected 701 probes, each performing multiple `ping` measurements towards 15 datacenters belonging to all cloud providers within the US. Figure 10 shows the results.

We show the minimum, median, and 95th percentile of latency observed in every PSA. The distribution is weighted according to the population of each PSA; visually, this translates to a higher vertical step in the CDF, for a larger population. The median distribution shows that almost the entire US population has median access latency below 75 ms – well within the human-perceivable latency threshold. The differences, however, show up for the 95th percentile distribution of PSAs latency as it includes probes installed in imperfect network conditions. Even in this case, ≈ 60% of

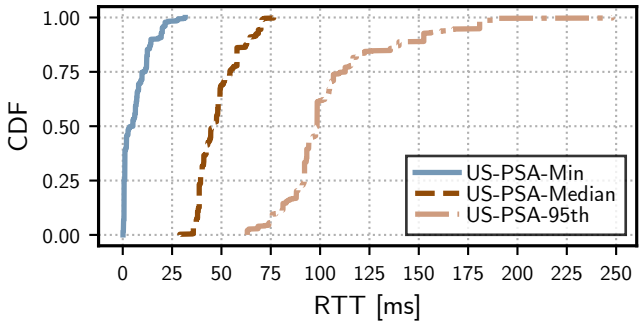


Figure 10: Distribution of RTT in US primary statistical areas weighted by normalized population.

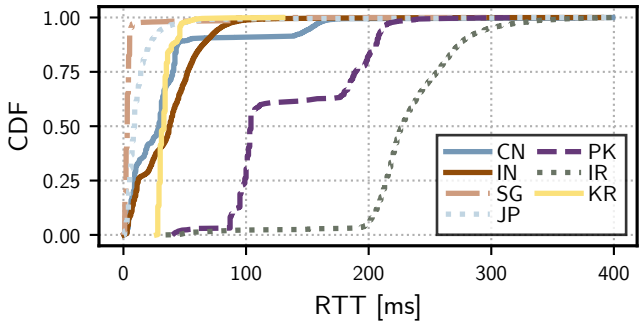


Figure 11: Distribution of RTT in some Asian countries with and without in-country datacenters.

the US population can reach the cloud within the coveted 100 ms threshold.

Case Study B: Asia. We contrast our analysis above by focusing on the state of cloud access in Asia. As previously hinted in § 4.1, latency distributions in Asia are rather skewed, resulting in unequal performance throughout the continent. To further investigate the cause, we carefully select seven Asian countries based on their landmass and physical distance to the cloud. Specifically, we investigate five countries with local datacenter deployments, i.e., China (48 probes), India (108 probes), Singapore (80 probes), Korea (20 probes), and Japan (188 probes); Pakistan (12 probes), which directly shares borders with the country with datacenter (India), and Iran (120 probes), which is farthest from any datacenter in the continent, nearest deployment in UAE and India. Figure 11 shows the results.

It is evident from the figure that countries with locally deployed datacenter can consistently meet the HPL threshold (100 ms). On the other hand, the impact of large geographical distance from the nearest datacenter, becomes evident in countries with no in-land datacenter. For instance, only 40% of samples from Pakistan are below 100 ms, while the rest can only satisfy the HRT threshold (250 ms). Finally, being Iran the geographically farthest from any datacenter, the minimum latency to reach the cloud is ≈200 ms, and almost 30% of samples did not even satisfy the HRT threshold.

Takeaway — The current cloud presence in the US can easily support the bulk of emerging applications, bounded by HPL constraints, for the majority of the population. On the other hand, while cloud reachability in Asia is generally good for countries with local deployment (e.g., China, Korea, Japan, India), it gets significantly worse with increasing geographical distance from the physical location of datacenters. Furthermore, the state of the user's network connectivity does not seem to have much effect on cloud reachability, as evident from consistently high latencies achieved by the majority of probes of Iran.

5 DISCUSSION

Vantage Points Representativeness. It is well known that a vast majority of cloud-hosted applications (e.g., HTTP-based) use TCP as their transport protocol and TCP traffic dominates over other protocols over the Internet [35]. Hence, we believe that our TCP-based measurements closely reflect realistic connection establishment overheads an application would experience while connecting to the cloud. Furthermore, our 12 month data collection absorbs the impact of temporally insignificant changes on the network. On the other hand, our measurements might fall short of accounting for the factors that affect the end-to-end latency of a service, e.g., interactions within the protocol stack or amongst entities on the service delivery route, etc. The most prominent lack in our analysis is the inability to showcase the impact of queuing delays observed by regular application traffic due to significantly smaller footprint of ping packets. Hence, the latencies in this work can be viewed as the minimum end-to-end delay bound an application can observe while connecting to the cloud.

Another limitation of our study stems from our choice of measurement platform. While RIPE Atlas is considered to be a gold-standard within the Internet research community, it is also influenced by several deployment biases that readers should be aware of. Atlas probes are mostly hosted by network enthusiasts and network service providers, which can skew the availability and deployment configurations of the probe. As discussed in § 3.2, while a vast majority of Atlas probes are available in Europe and North America, only a fraction (< 10%) of the probes are deployed in Africa. Furthermore, even within Europe and North America (see Fig.1b), the probes are not uniformly distributed over the continent's geography. As a result, our dataset includes both countries with extremely dense probe availability and those without many options. While we compensate for some of these biases in our post-collection analysis, e.g., by carefully pruning out probes installed in privileged networks, we do not have much control over others.

Measurements Duration. *Would it be possible to obtain similar results with only one month worth of measurements?* We want to stress the fact that this question can be answered only by analyzing a long-term dataset. To provide an answer to the question, we looked at 4 providers: 2 with private WAN and 2 relying on the public Internet. We compared the first four months of measurements, divided into intervals of one month. We found out that one provider with private WAN and one without exhibited very consistent and reliable connectivity. The other two providers also had very similar performance across three months (negligible change) of measurements. However, during one month (October 2019) these providers experienced significant divergence as their latency performances

were rather degraded with respect to the usual. However, we have no means of ascertaining the root cause behind this. Therefore, we conclude that, most likely, we could reach similar conclusions with only 1 month of data, but this strongly depends on the time and interval of the measurements. In other words, without a known baseline, we would not know if a particular time period is "usual" or "unusual".

Where to Proceed from Here. *Can existing cloud infrastructure support the operation of future-forward latency-critical applications?* Our large-scale latency measurements have answered this question affirmatively for the majority of the continents and for the applications requiring a latency bounded by HRT. On the other hand, applications requiring latency below MTP, such as augmented reality, can only be supported, with the current cloud infrastructure, within North America, Europe and Oceania. We also found that private WANs have little-to-no impact in bringing cloud coverage any closer to users. Providers that rely on public Internet achieve almost similar latencies to those with private network backbone in regions without local datacenter deployment. However, while establishing new cloud regions globally may seem like the only viable option to drive down cloud access latencies, we also show that in regions with already high degree of cloud pervasiveness and excellent network connectivity, deploying more datacenters does not bring much benefit (e.g., the USA). Therefore, a key takeaway that existing cloud providers can take from this study could be to prioritize infrastructure expansion in under-provisioned regions, specifically Africa, Asia, and South America.

6 RELATED WORK

To the best of our knowledge, the first significant cloud reachability study dates back to 2010 [20]. However, the substantial evolution of cloud computing and datacenter deployments over the decade since its publication suggests the findings of that paper are worthwhile updating to reflect today's state of the art. More recently, the cloud performance report 2019 from ThousandEyes[2] monitors and compares 95 end-points' performance to the major cloud providers (Amazon, Microsoft, Google, Alibaba, and IBM) for a maximum period of one month during the year 2019. Their measurement methodology consists of collecting network latency and paths from 98 TCP-based vantage points in various countries worldwide. In our study, we target almost the double of endpoints (189), and we use more vantage points (up to 8500) located in 184 different countries. Furthermore, our measurements have been collected for a significantly longer period of time. For these reasons, we believe our study to be more comprehensive and broad. Furthermore, we expand our previous work [23] with more measurements and vantage points, as well as a thorough path characterization study. In particular, our cloud pervasiveness study reports similar findings to [1], where Todd et al. show that cloud providers are already bypassing Tier-1 providers, therefore making the Internet "flatter" (less hierarchical).

Related methodology partly used in our evaluation can be found in [25], [18], and [17]. In [25], the authors measured the performance of the 5G deployment in the USA. In their measurements, they selected three cloud providers (Amazon, Google, and Azure) and evaluated the base RTT without cross-traffic, and download and upload bandwidth and latency times. The results show that the 5G RTT latency has little improvement compared to 4G, and the first-hop accounts for ∼ 27 ms while the remaining latency in the path towards the cloud is similar to our measurements. Therefore, at this stage of 5G deployment, there is still little improvement in the last mile connectivity times, but it is expected to be addressed in the future and achieve the promised sub-millisecond RTT. In [18], the authors compare the performance of ICMP-based ping and traceroute tools to detect cloud service providers' outages. The main issue regarding ICMP measurements is that it can underestimate the availability as it only checks the network-level connectivity and not the application itself. To validate this hypothesis, the authors compare ICMP-traceroute against TCP-traceroute, and the experimental results show that there is disagreement on some measurements (up to 3%) where the ICMP fails while HTTP succeeds, and vice-versa. In [17], where the authors analyzed the inter-continental paths connecting three big cloud providers, namely Amazon, Google, and Azure. They found out that those cloud providers have dedicated paths connecting their data-centers, which increased the network path performance (lower packet loss and latency) compared to regular inter-continental data traversal through different independent ASes. We complement and confirm their analysis by adding the latency information from several Atlas probes to those cloud providers, providing a bigger picture and how close to each cloud provider each country is.

Within the same topic, in [11], the authors study cloud provider outages and dig into the causes of such events by analyzing the connectivity between major cloud service providers, e.g., Google, Amazon, Microsoft, etc. To facilitate the analysis, the authors define a set of metrics based on graph properties and measure the interconnectivity between the ASes of those cloud service providers. In [38], the authors propose a mechanism to verify whether cloud providers are respecting the subscribed SLAs for packets being processed in cloud middleboxes. In [28], the authors performed a large scale study of web page performance, showing the impact of different Web protocols and access media in the performance of page loading and overall user experience.

7 CONCLUSION

We conducted a large-scale cloud reachability study with the aim to evaluate the current state of cloud connectivity globally. In our study, we targeted 189 compute-capable cloud regions of ten major cloud networks from 8500 globally distributed RIPE Atlas probes for a period of 12 months. Through our extensive analysis of network latency, we found that the majority of the world population can access a cloud facility within 100ms – which is a critical threshold for many future-forward networked applications. Furthermore, our analysis of user-to-cloud path lengths revealed that cloud providers relying on private WAN for network interconnections are already very pervasive since the majority of the paths transit through their infrastructure. However, we also found that end-to-end network latency is rarely impacted by underlying network infrastructure as even providers relying on public Internet achieve similar latencies, albeit with higher variability. Our case study analysis showcased the impact of geographical distance to cloud by analysing regions

[2]https://www.thousandeyes.com/research/cloud-performance.

with contrasting datatcenter deployment density – the USA and Asia. Our results revealed that extensive datacenter deployment is key to make cloud access latencies consistently compatible with requirements of next-generation applications, especially for Asia, South America, and Africa.

ACKNOWLEDGEMENT

We would like to acknowledge RIPE Atlas team for providing us access to their platform and supporting our measurements with increased quota limits. This work was supported by the Swedish Foundation for Strategic Research with grant number GMT-14-0032 (Future Factories in the Cloud), the Academy of Finland in the BCDC (314167), AIDA (317086), WMD (313477) projects and Celtic project Piccolo (C2019/2-2).

REFERENCES

[1] Todd Arnold, Jia He, Weifan Jiang, Matt Calder, Italo Cunha, Vasileios Giotsas, and Ethan Katz-Bassett. 2020. Cloud Provider Connectivity in the Flat Internet. In *Proceedings of the ACM Internet Measurement Conference (IMC '20)*. Association for Computing Machinery, New York, NY, USA, 230–246. https://doi.org/10.1145/3419394.3423613

[2] AWS. 2019. AWS Direct Connect Partners. "https://aws.amazon.com/directconnect/partners/". (2019).

[3] Vaibhav Bajpai, Steffie Jacob Eravuchira, and Jürgen Schönwälder. 2015. Lessons learned from using the ripe atlas platform for measurement research. *ACM SIGCOMM Computer Communication Review* 45, 3 (2015), 35–42.

[4] Flavio Bonomi, Rodolfo Milito, Preethi Natarajan, and Jiang Zhu. 2014. Fog computing: A platform for internet of things and analytics. In *Big data and internet of things: A roadmap for smart environments*. Springer, 169–186.

[5] Marshini Chetty, Srikanth Sundaresan, Sachit Muckaden, Nick Feamster, and Enrico Calandro. 2013. Measuring Broadband Performance in South Africa. In *Proceedings of the 4th Annual Symposium on Computing for Development (ACM DEV-4 '13)*. Association for Computing Machinery, New York, NY, USA, Article Article 1, 10 pages. https://doi.org/10.1145/2537052.2537053

[6] Yi-Ching Chiu, Brandon Schlinker, Abhishek Balaji Radhakrishnan, Ethan Katz-Bassett, and Ramesh Govindan. 2015. Are We One Hop Away from a Better Internet?. In *Proceedings of the 2015 Internet Measurement Conference (IMC '15)*. Association for Computing Machinery, New York, NY, USA, 523–529. https://doi.org/10.1145/2815675.2815719

[7] Song-Woo Choi, Siyeong Lee, Min-Woo Seo, and Suk-Ju Kang. 2018. Time sequential motion-to-photon latency measurement system for virtual reality head-mounted displays. *Electronics* 7, 9 (2018), 171.

[8] CloudHarmony. 2020. Transparency for the cloud. "https://cloudharmony.com/". (2020).

[9] Maximilian Eder, Lorenzo Corneo, Nitinder Mohan, Aleksandr Zavodovski, Suzan Bayhan, Walter Wong, Per Gunningberg, Jussi Kangasharju, and Jörg Ott. 2021. Surrounded by the Clouds. (2021). https://doi.org/10.14459/2020mp1593899

[10] Executive Office of the President, Washington, D.C. 20503. 2013. OMB BULLETIN NO. 13-01. https://obamawhitehouse.archives.gov/sites/default/files/omb/bulletins/2013/b13-01.pdf. (28 Feb. 2013).

[11] Benjamin Fabian, Annika Baumann, and Jessika Lackner. 2015. Topological analysis of cloud service connectivity. *Computers & Industrial Engineering* 88 (2015), 151–165. https://doi.org/10.1016/j.cie.2015.06.009

[12] Armando Fox, Rean Griffith, Anthony Joseph, Randy Katz, Andrew Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, et al. 2009. Above the clouds: A berkeley view of cloud computing. *Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS* 28, 13 (2009), 2009.

[13] Gartner. 2020. Gartner Forecasts Worldwide Public Cloud Revenue to Grow 6.3% in 2020. https://www.gartner.com/en/newsroom/press-releases/2020-07-23-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-6point3-percent-in-2020. (2020).

[14] Google. 2019. Stadia. https://stadia.dev/. (2019).

[15] Robert L Grossman. 2009. The case for cloud computing. *IT professional* 11, 2 (2009), 23–27.

[16] Hadi Asghari and Arman Noroozian. 2020. PyASN. "https://pypi.org/project/pyasn/". (2020).

[17] Osama Haq, Mamoon Raja, and Fahad R. Dogar. 2017. Measuring and Improving the Reliability of Wide-Area Cloud Paths. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 253–262. https://doi.org/10.1145/3038912.3052560

[18] Zi Hu, Liang Zhu, Calvin Ardi, Ethan Katz-Bassett, Harsha V. Madhyastha, John Heidemann, and Minlan Yu. 2014. The Need for End-to-End Evaluation of Cloud Availability. In *Passive and Active Measurement*, Michalis Faloutsos and Aleksandar Kuzmanovic (Eds.). Springer International Publishing, Cham, 119–130.

[19] Yuchen Jin, Sundararajan Renganathan, Ganesh Ananthanarayanan, Junchen Jiang, Venkata N. Padmanabhan, Manuel Schroder, Matt Calder, and Arvind Krishnamurthy. 2019. Zooming in on Wide-Area Latencies to a Global Cloud Provider. In *Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM '19)*. Association for Computing Machinery, New York, NY, USA, 104–116. https://doi.org/10.1145/3341302.3342073

[20] Ang Li, Xiaowei Yang, Srikanth Kandula, and Ming Zhang. 2010. CloudCmp: Comparing Public Cloud Providers. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC '10)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/1879141.1879143

[21] Katerina Mania, Bernard D Adelstein, Stephen R Ellis, and Michael I Hill. 2004. 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*. ACM, 39–47.

[22] Microsoft. 2019. Xbox Project xCloud. (2019). https://www.techradar.com/news/project-xcloud-everything-we-know-about-microsofts-cloud-streaming-service.

[23] Nitinder Mohan, Lorenzo Corneo, Aleksandr Zavodovski, Suzan Bayhan, Walter Wong, and Jussi Kangasharju. 2020. Pruning Edge Research with Latency Shears. In *Proceedings of the 19th ACM Workshop on Hot Topics in Networks (HotNets '20)*. Association for Computing Machinery, New York, NY, USA, 182–189. https://doi.org/10.1145/3422604.3425943

[24] Nitinder Mohan and Jussi Kangasharju. 2016. Edge-Fog cloud: A distributed cloud for Internet of Things computations. In *2016 Cloudification of the Internet of Things (CIoT)*. 1–6. https://doi.org/10.1109/CIOT.2016.7872914

[25] Arvind Narayanan, Eman Ramadan, Jason Carpenter, Qingxu Liu, Yu Liu, Feng Qian, and Zhi-Li Zhang. 2020. A First Look at Commercial 5G Performance on Smartphones. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 894–905. https://doi.org/10.1145/3366423.3380169

[26] PeeringDB. 2020. The Interconnection Database. https://www.peeringdb.com/. (2020). accessed on October 16, 2020.

[27] Kjetil Raaen, Ragnhild Eg, and Carsten Griwodz. 2014. Can gamers detect cloud delay?. In *2014 13th Annual Workshop on Network and Systems Support for Games*. IEEE, 1–3.

[28] Mohammad Rajiullah, Andra Lutu, Ali Safari Khatouni, Mah-Rukh Fida, Marco Mellia, Anna Brunstrom, Ozgu Alay, Stefan Alfredsson, and Vincenzo Mancuso. 2019. Web Experience in Mobile Networks: Lessons from Two Million Page Visits. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1532–1543. https://doi.org/10.1145/3308558.3313606

[29] RIPE NCC. 2020. Probe tags. "https://atlas.ripe.net/docs/probe-tags/". (2020).

[30] Mahadev Satyanarayanan, Paramvir Bahl, Ramón Caceres, and Nigel Davies. 2009. The case for vm-based cloudlets in mobile computing. *IEEE pervasive Computing* 4 (2009), 14–23.

[31] Amazon Web Services. 2019. AWS Global Infrastructure Map. "https://aws.amazon.com/about-aws/global-infrastructure/". (2019).

[32] Rahul Singh, Prateek Sharma, David Irwin, Prashant Shenoy, and KK Ramakrishnan. 2014. Here today, gone tomorrow: Exploiting transient servers in datacenters. *IEEE Internet Computing* 18, 4 (2014), 22–29.

[33] RN Staff. 2015. RIPE Atlas: A global internet measurement network. *Internet Protocol Journal* 18, 3 (2015).

[34] TeleGeography. 2019. Submarine Cable Map. "https://www.submarinecablemap.com/". (2019).

[35] M. Trevisan, D. Giordano, I. Drago, M. M. Munafò, and M. Mellia. 2020. Five Years at the Edge: Watching Internet From the ISP Network. *IEEE/ACM Transactions on Networking* 28, 2 (2020), 561–574.

[36] Wikipedia. 2019. Statistical area (United States). https://en.wikipedia.org/wiki/Statistical_area_United_States). (2019).

[37] David L Woods, John M Wyma, E William Yund, Timothy J Herron, and Bruce Reed. 2015. Factors influencing the latency of simple reaction time. *Frontiers in human neuroscience* 9 (2015), 131.

[38] X. Zhang, H. Duan, C. Wang, Q. Li, and J. Wu. 2019. Towards Verifiable Performance Measurement over In-the-Cloud Middleboxes. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*. 1162–1170.