DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Information Systems

Analyzing Cloud Reachability via Wireless Networks on Global Scale

The Khang Dang

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Information Systems

Analyzing Cloud Reachability via Wireless Networks on Global Scale

Globale Analyse der Erreichbarkeit der Cloud in drahtlosen Netzwerken

Author:	The Khang Dang
Supervisor:	Prof. Dr. Jörg Ott
Advisor:	Dr. Nitinder Mohan
Submission Date:	15.05.2021

I confirm that this bachelor's thesis in information systems is my own work and I have documented all sources and material used.

Munich, 15.05.2021 The Khang Dang

Acknowledgements

I would like to thank my supervisor Prof. Dr.-Ing. Jörg Ott for allowing me to write this thesis at his chair. Furthermore, I want to thank my advisor Dr. Nitinder Mohan for his support and help in understanding this topic as well as his invaluable feedback. I also thank Speedchecker, especially Janusz Jezowicz, for providing us access to their measurement platform.

Abstract

Cloud computing has seen a continuous growth in importance, offering immense computing power without the need for the physical processing hardware and infrastructure at the end-user location. This enables emerging technologies which rely on this concept ranging from AR and cloud gaming to autonomous vehicles. Latency and reliability of the connection plays an important role in realizing these technologies, therefore, cloud providers continue to invest massively into expansion of their network and deployment of data centers to offer better global reachability. In this thesis we investigate user-to-cloud connectivity using the Speedchecker platform to run extensive ping and traceroute measurements towards 195 data centers from 10 different cloud providers. Because Speedchecker operates a large number of wirelessly connected probes across the globe which are exclusively placed on the last-mile, this accurately reflects real end-user experience. It also shows the current state of wireless cloud connectivity which is relevant for many of the aforementioned technologies. Our results show that users in a majority of the countries can reach a data center within 100 ms while mainly countries in Africa, South America and Western Asia display higher latency influenced by the sparse cloud data center deployment in these regions. Our findings indicate that outside of geographical distance, the last-mile and especially the wireless connection introduces a significant portion of the overall latency. We also see that the private backbone built by some of the major cloud providers allow them to control most of the route that traffic takes by peering with many different ISPs which offers noticable improvements in latency and consistency for longer distances.

Kurzfassung

Die Bedeutung von Cloud Computing wächst kontinuierlich weiter, da diese Technologie immense Rechenleistung bietet ohne die physische Infrastruktur am Ort des Endnutzers zu benötigen. Dies ermöglicht die Realisierung von neuen Technologien die vom Cloud Computing abhängen, von AR über Cloud Gaming bis hin zum autonomen Fahren. Hierbei spielt die Latenz und Verlässlichkeit der Internetverbindung eine wichtige Rolle, weswegen Cloud Provider weiterhin massiv in den Ausbau ihres Netzwerkes und die Verteilung ihrer Datenzentren investieren. In dieser Arbeit untersuchen wir die Verbindung vom Nutzer zur Cloud mit Hilfe der Speedchecker Plattform, über die wir Ping und Traceroute Messungen zu 195 Datenzentern von 10 Cloud Providern ausführen. Da Speedchecker zahlreiche, über die ganze Welt verteilte „Probes“ mit drahtloser Internetverbindung betreibt die sich alle auf der „Last Mile“ befinden, erlaubt dies einen akkuraten Einblick in die aktuelle Situation von drahtlosen Verbindungen aus der Endnutzerperspektive, die für oben genannten Technologien äußerst relevant sind. Unsere Ergebnisse zeigen auf, dass Nutzer in einem Großteil der Länder ein Datenzentrum in unter 100 ms erreichen können, wobei einige Länder in Afrika, Südafrika und Westasien höhere Latenzen aufweisen, die durch die spärliche Verteilung von Datenzentren in diesen Regionen bedingt sind. Die Ergebnisse deuten zusätzlich darauf hin, dass neben der geografischen Distanz auch die „Last Mile“ einen großen Einfluss auf die Latenz hat, insbesondere bedingt durch die drahtlose Verbindung. Wir merken außerdem, dass einige Cloud Provider mit privaten Netzwerken einen großen Teil des Pfads den die Daten nehmen kontrollieren. Dies geschieht mit Hilfe von Peering mit vielen verschiedenen Internetanbietern und führt bei längeren Distanzen zu spürbaren Verbesserungen in der Latenz und Konsistenz der Messungen.

Contents

1. Introduction

The number of internet users and the amount of connected devices keeps growing year by year. According to Cisco [1], half of the globally connected devices will be connected in a Machine-To-Machine (M2M) connection by 2023, of which connected home applications will have the largest share, while connected cars are expected to be the application with the largest relative growth. Cloud computing and connectivity play an important role in this growth since many of these applications rely on a fast and reliable wireless connection to the cloud. Using the cloud offers users many advantages in regards to availability, cost and reliability because of the vast computing resources that are available without having to pay a large upfront cost to build the necessary infrastructure, as well as the ability to pay for computing resources as needed, even for short-term demand spikes [2, 3]. On the other hand, ensuring a low latency (and in many cases wireless) connection to the cloud at all times for applications like autonomous vehicles or cloud gaming is one of the big challenges in this area.

1.1. Research Questions

The purpose of this thesis is to show an overview of the global cloud reachability from wireless networks. To do this, we will focus on two key aspects:

RQ1: What is the state of global cloud access latency for the end-user? Because latency is a critical component in enabling and guaranteeing the safety of many next-generation technologies as mentioned above, it is important to investigate the global situation of cloud access latency from an end-user perspective. This is especially the case for access latency from wirelessly connected devices because this is the primary way that many of these devices and applications connect to the Internet. We are interested in finding out the differences between different regions and how far current cloud access latency can support these new technologies currently.

RQ2: What factors influence the quality of the connection and latency? Because there are a lot of regional differences in latency, we want to further investigate what factors might affect the overall access latency. Providers like Amazon, Google, and

Microsoft continue to invest into the expansion of their private networks to enable shorter paths and increase the quality of experience for their cloud services. We want to see whether this improves latency for the end-user compared to other cloud providers who use mainly the public Internet infrastructure to connect to users. There is another factor that ties right into this: peering. Operating such a large-scale network allows for more different ISPs and other networks to directly connect with the cloud provider network and circumvent having to rely on intermediary networks to transport traffic which might influence the observed latency. Analyzing these and other factors allows us to see where improvements can be made and where current bottlenecks regarding the latency are present.

1.2. Contribution

We conducted a cloud reachability analysis for ten globally distributed cloud providers of differing sizes, with different network infrastructures and different geographical presence. We looked at the latency observed from thousands of wireless probes from Speedchecker [4] that reflect true end-user experience and compare the situation between different areas. We also provide some reference latency thresholds so the current state of global cloud connectivity can be better understood and to see which kinds of applications the current latency allows. Additionally, some light is also shed on different factors which may affect the overall latency as well as the different degrees to which they affect it. We see that geographical distance to a destination, last-mile latency, the wireless connectivity and the different interconnectivity between cloud providers and other networks all impact observed latency to a varying degree.

1.3. Outline

We first give an overview of some key terminology as well as past work and recent developments in related topics in chapter 2. Afterward, we describe our approach in gathering and storing the data in chapter 3, which includes an overview of the used platform and the specific details of the measurements. These results are then presented in chapter 4, showing some of the relevant factors in user connectivity such as the importance of the geographical location as well as the impact of the wireless connection. Finally, in chapter 5 we summarize our results and reach our conclusion.

2. Background

In this chapter, we explain some of the required background knowledge and terminology that is widely used and look at different aspects of cloud connectivity and past work on this topic that we aim to further research and expand upon.

2.1. ISP Tiers

The Internet consists of many interconnected networks called autonomous systems (ASes) which are managed by different organizations. The ASes which primarily transport the traffic over the Internet are controlled by Internet Service Providers (ISPs) [5]. These ISPs are classified into three different tiers. Tier 1 providers are considered the backbone of the Internet since they allow for global connectivity and can reach and transport traffic to every location in the world through their own network as well as through settlement-free peering interconnections with other Tier 1 ISPs. Tier

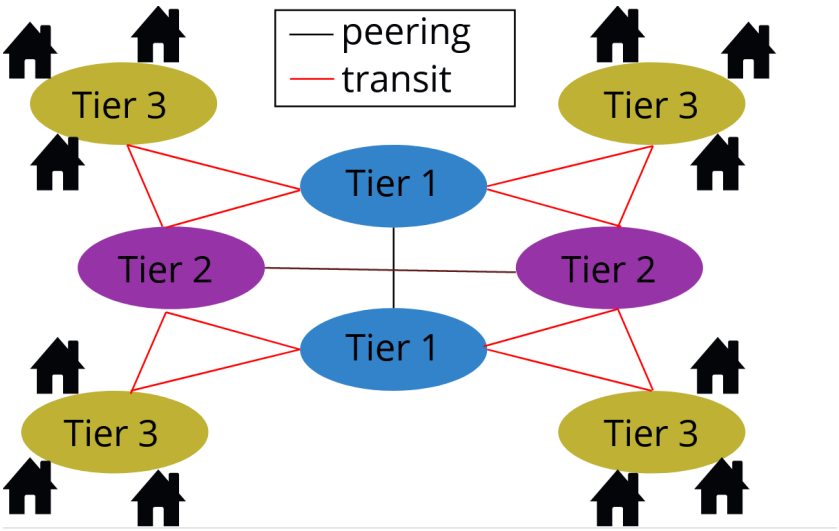


Figure 2.1.: Overview of ISP Tiers with home users connecting to Tier 3 ISPs based on the model from [5]

2 ISPs rely on a combination of purchasing transit from Tier 1 providers and peering with other Tier 2 ISPs to reach the destination network, while Tier 3 ISPs are mainly the ones delivering Internet access to consumers by strictly purchasing transit from other providers. This hierarchy is also illustrated in Fig. 2.1 showing the different relationships between the ISP Tiers and where the end-user is located in this model. Further in-depth information regarding this model can be found on the ThousandEyes website [5].

2.2. Differences in Cloud Providers

With the growing importance of cloud computing, the competition in this market grows as well which allows more choices regarding which cloud provider to use. This also makes choosing the best cloud provider for a specific use and location harder, which is why an increasing amount of reports and articles have been released over the past few years. One notable report is the Cloud Performance Benchmark from ThousandEyes [6], which is a yearly report measuring and comparing network performance for different cloud providers. In 2019 they compared five top cloud providers: Amazon Web Services(AWS), Microsoft Azure, Google Cloud Platform (GCP), Alibaba Cloud, and IBM Cloud. They gathered over 320 million data points from 98 different vantage points connected to Tier 2 and Tier 3 ISPs, as well as cloud backbone networks, and offer insights into many interesting location- and provider-specific issues and improvements that have been achieved. For example, they found out that users from Europe accessing compute engine workloads in GCP's Mumbai data center continue to experience a much higher latency compared to other cloud providers in this region since there is a lack of direct connectivity between Europe and India which leads to the traffic traveling halfway around the world to reach the destination. The difference in reliance on the public Internet between the different providers is also a point which is explored in the report and lead to the findings, that while Alibaba and Amazon force the traffic from the end-user through the public Internet, with it only entering their private network very close to the target region, Microsoft and Google absorb traffic into their own backbone as close to the user as possible and IBM having a hybrid approach depending on the region. In addition, the report also goes over many other points, e.g., the difference in using the private cloud backbone compared to the public Internet for AWS and comparing the performance of different ISPs in North America.

Another report of importance is the annually released Magic Quadrant by Gartner [7]. The 2020 Magic Quadrant for Cloud Infrastructure and Platform Services offers insight into the strengths and weaknesses of seven different cloud providers, including Tencent Cloud and Oracle in addition to the five vendors which were analyzed in the

ThousandEyes report. They are evaluated them based on different criteria summarized under the two categories "Ability to Execute", which includes criteria like the depth of feature sets, the success of the business, the value for money they offer, and the customer experience and "Completeness of Vision", which is comprised of subcategories like the understanding of the market, marketing and sales strategies, and innovation. Based on these criteria, AWS emerged as the leader, followed by Microsoft Azure and GCP, all three having a good offering for most, if not all, use cases, while the other providers are recommended for specific use cases, e.g., Alibaba Cloud for cloud infrastructure within China and Southeast Asia.

Arnold et al. [8] closely investigated the performance differences between using the public Internet and using the private network to connect to the cloud providers through measurements to two large cloud providers, Amazon and Google. The reason for using these two providers is their extensive private backbone and their offering of routing through that private network in their premium plans (Google offers a premium tier, Amazon has AWS Global Accelerator), which allowed for direct performance comparison. Arnold et al. concluded through their extensive measurements that in most cases routes using the private network had better or at least indistinguishable performance compared to routes over the public Internet. They also noted that it is not this straightforward for all cases and that the performance also depends on many other factors like geographic location and network connectivity as well, which vary individually.

Lastly, Eder [9] used measurements through RIPE Atlas [10] to assess cloud reachability globally. Our thesis is mainly inspired by their work and seeks to use their data as a comparison to our own measurements and expands upon it with further investigation into further aspects that are partially enabled through our use of a different platform. Their results have also been integrated into and published in [11] and we will simply refer to this data set as the RIPE Atlas data set.

2.3. Internet Flattening and AS Relationships

In addition to the changes in the specific cloud providers and their network expansion, the landscape of the Internet as a whole is also changing drastically over time. While the Internet originally had a hierarchical structure where traffic was mainly routed through a small group of interconnected networks (i.e. Tier 1 ISPs) through a transit interconnection, a type of interconnection where a network purchases connectivity services from another ISP [12], in recent years, many networks have started to bypass this traditional structure in a process described as Internet flattening. Instead, peering, a form of interconnection where two networks have an agreement to exchange their own

and their customer's traffic, has gained more popularity since it allows for more control over routing [13] and increased performance compared to transit interconnections [14]. This is facilitated through the rise of public peering in Internet exchange points (IXP) [15], facilities where networks can physically connect to a switching fabric which then provides the possibility to connect to every other AS connected to the fabric [12], as well as an increasing number of direct or private peering interconnections through colocation facilities [13, 16].

In parallel, cloud providers have invested massively into improving and expanding their infrastructure, with Google investing $30 billion over three years into their network which by some accounts deliver 25 percent of worldwide traffic [17]. This further supports the shift away from transit and instead towards an increasing reliance on cloud providers and their networks. For example, Chiu et al. [18] have shown that a majority of Google's paths go directly from their network into their client's network, especially for clients who produce a high volume of traffic, while Arnold et al. [13] show that major cloud providers, namely Amazon, Google, IBM, and Microsoft, can reach over 76% of the Internet while bypassing Tier 1 and Tier 2 ISPs. Because of this, large online companies like Apple, Spotify, Netflix, Lyft, and Snap are also moving towards cloud providers, spending hundreds of millions on their services [19]. Lastly, some cloud providers have also started to offer a new interconnection service called virtual private interconnection (VPI) [20, 21, 22], which enables customers to purchase a single port in a given colocation facility [23], establishing a dedicated connection towards the provider and bypassing the public internet.

The relationships between ASes, especially in this rapidly changing Internet landscape, are thus a widely discussed topic. Many attempts have been made to uncover the AS-level topology of the Internet, but due to the confidential nature of these relationships and the different types of relationships, this is an area that is still not fully explored. CAIDA provides AS relationship datasets which were inferred from BGP through different methods [24, 25] as well as their own traceroutes [26]. Others have focused on the view from specific cloud providers [13, 23] or within colocation facilities [16] since public BGP data has a good coverage of Tier 1 and Tier 2 ISPs, but a limited view into smaller ASes at the edge of the network [13, 27, 23], but the Internet topology is still an area that has not been fully revealed yet.

2.4. Last-mile Latency

With this change in Internet routes and the shortening of the paths, the last-mile latency, which generally denotes the latency between the home network and the providing ISP, becomes an increasing contributor to overall latency. This is because cloud providers

and other content providers do not have any influence on this section of the path since it is mostly controlled by regional Tier 3 ISPs who provide Internet access to the user. Therefore, this has become a more relevant topic for recent research. Sundaresan et al. [28] investigated page load times for several popular sites in 2013 and the impact of downstream throughput and latency on these times. They found that last-mile latency is a major factor in overall latency and significantly contributes to DNS lookup and page-load times. It especially becomes the main bottleneck in networks where downstream throughput exceeds 16 Mbit/s. Bajpai, Eravuchira, and Schönwälder [29] have used the SamKnows [30] and RIPE Atlas [10] platforms to conduct further investigations into last-mile latency in the EU and US. They first noted that latency within the home network can vary a lot and have a noticeable impact when measuring last-mile latency and thus do not take this into account when measuring the last-mile latency, instead only focusing on the path between the home router and the ISP denoted as *hop1* and *hop2* respectively. They found out that last-mile latency differs depending on the connection type with DSL displaying a median of 16 ms, cable 8 ms, and fiber 4 ms of last-mile latency which is stable across the day. Another consideration in their findings is that last-mile latency differs based on the location with the US east coast showing a much higher last-mile latency than the west coast and confirmed that latency for DSL connections is influenced by broadband speeds.

Another important part to look at here is the latency that a wireless connection introduces, in particular cellular connections, where the entire last-mile is made up of the wireless connection compared to only the section until the router for home users. This was investigated by Schulz et al. in [31] where they measured current 4G network latency via ICMP pings from a conventional Android smartphone. They discovered that when targeting `www.google.de`, the latency to reach the gateway of the core network was at a minimum 39 ms, while the time to receive the reply from Google only took an additional 5 ms, effectively making the cellular network responsible for 90% of the overall measured latency. The observed total latency is then further influenced by the time of the measurement and as such, the load of the LTE cell that one is connected to. This shows, that for wireless connections the distance towards the destination and the number of devices nearby and connecting to the same LTE cell are the main limiting factors.

However, research into this particular topic is still relatively sparse and most of the research has been conducted via probes with wired connections and is limited in terms of geographical distribution of these probes. This is the reason why we want to look into this from a global perspective focusing on wireless connections.

2.5. Measurement Platforms

Lastly, we want to give a short overview of some of the well-established measurement platforms that have been used for many different research purposes and compare Speedchecker, the platform we used for our measurements, to them.

RIPE Atlas RIPE Atlas [10] is a well-known and widely used measurement platform that has a network of over 12,000 probes which are distributed globally through mainly hardware-based probes that are connected via Ethernet [32], which are hosted by volunteers, including private individuals and large corporations alike. These probes can be used to schedule a range of different measurements including but not limited to: ping, traceroute, DNS and HTTP tests [33].

Measurement Lab (M-Lab) M-Lab [34] is a platform that has servers deployed across the globe which allow end-users to run tests towards these servers. Since it collects measurements from end-users, it offers a unique view into the end-to-end connections of real users [35]. They also allow researchers to deploy their measurement tools on their servers, therefore making these accessible for a wider audience. This allows for a wide range of different tests including a networks diagnostics test for TCP connections, reverse traceroutes and performance tests for specific applications like YouTube and Spotify [36]. All of the measurements are saved and published publicly, creating one of the largest Internet performance data sets. This enables interested individuals to use an extensive data set for analysis without having to design a measurement tool and collect data on their own.

Speedchecker Speedchecker is the platform that we used for our measurements. It is a platform that hosts hundreds of thousands of probes placed on the last-mile [4] which allows for results that accurately reflect the end-user experience. A majority of their probes are hosted on Android devices that have a wireless Internet connection. We will go into further detail into the platform in chapter 3.1, but the main advantages are the number of probes and especially the wireless connectivity compared to RIPE Atlas, and the customization options while running measurements from a global network of probes compared to M-Lab. Since we are interested in the measurement metrics from a wireless perspective, Speedchecker offered us the best combination of features to gain deeper insights into these.

3. Methodology

In this chapter, we will illustrate the resources that were used to gather and store the data which will be used for further analysis afterward.

3.1. Platform

The platform we used for our measurements is Speedchecker [4]. They operate a global measurement network spanning over 170 countries and allow for tests like pings, traceroutes and HTTP GET requests among others. Their probes are exclusively placed on the last-mile which enables them to reflect the true end-user experience. There are three types of probes in use by Speedchecker, PC, Router, and Android, with Android probes being the majority, constituting around 89% of the total of over 470,000 probes observed throughout the measurement period, of which at least 29,000 probes were available at any given time. The Android probes are also exclusively using wireless connections which allows for a great comparison to the RIPE Atlas data where probes are mostly connected via Ethernet.

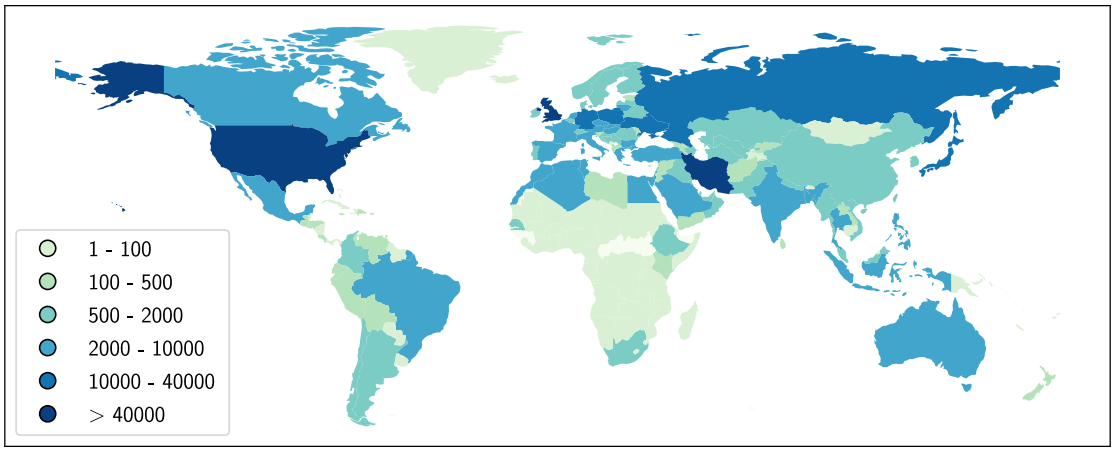


Figure 3.1.: Distribution of Speedchecker probes observed throughout the measurements

The distribution of the probes can be seen in Fig. 3.1. This shows us that the probe coverage in North America and Europe (especially Central and Western Europe) is exceptional, with most of the countries showing at least 2000 local probes and even up to over 40,000 for Great Britain, Iran and the United States. On the other hand, coverage in other continents is not optimal, especially in Africa, where a large number of countries only have a negligible number of probes. Nevertheless, there still are some countries in these continents that have a reasonable number of probes and allow us to gain insight into these specific areas.

3.2. Measurement Endpoints

Since the aim of our analysis is to get a comprehensive overview of cloud reachability and performance, we try to take different aspects like the geographical location of data centers and the size and type of the network backbone of different providers into account. Accordingly, we chose a range of different cloud providers, ranging from large, well-established ones (Amazon, Google, Microsoft) which operate a massive private network where traffic can get routed through, to smaller providers which use the public internet for traffic like Vultr and Linode. We also included Alibaba Cloud since they operate mainly in Asia, with a special focus on China. This culminated into a total of ten different cloud providers and 195 different data centers. The exact distribution of the data centers can be seen in Table 3.1 with their locations illustrated in Fig. 3.2. Similarly to the probe distribution of Speedchecker, the distribution of data centers is also varying widely depending on the continent. Most of them are located in

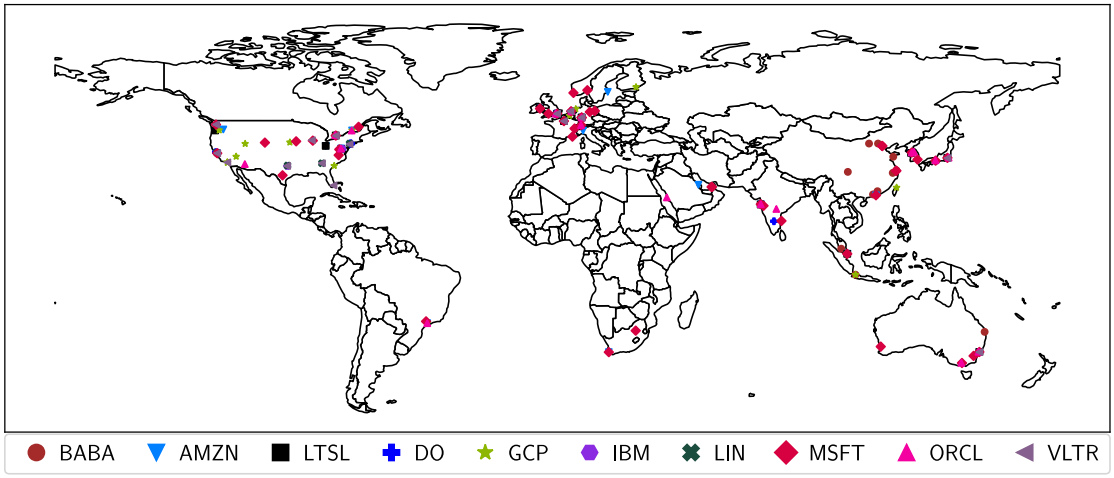


Figure 3.2.: Location of targeted data centers

Provider	Datacenters per continent						Total
	AF	AS	EU	NA	OC	SA	
Alibaba (BABA)	-	16	2	2	1	-	**21**
Amazon EC2 (AMZN)	1	6	6	6	1	1	**21**
Amazon Lightsail (LTSL)	-	4	4	4	1	-	**13**
Digital Ocean (DO)	-	1	4	6	-	-	**11**
Google (GCP)	-	8	6	10	1	1	**26**
IBM (IBM)	-	1	6	6	-	-	**13**
Linode (LIN)	-	3	2	5	1	-	**11**
Microsoft (MSFT)	2	15	14	10	4	1	**46**
Oracle (ORCL)	-	7	4	4	2	1	**18**
Vultr (VLTR)	-	1	4	9	1	-	**15**
Total	**3**	**62**	**52**	**62**	**12**	**4**	**195**

Table 3.1.: Data center distribution of different cloud providers

Asia, Europe, or North America, while there are only three in Africa and four in South America.

For the specific endpoints within the data centers, we used CloudHarmony [37], a platform that offers vantage points for network measurements through virtual machines located in the data centers of 93 different providers, including the ones that we used in our measurements.

3.3. Data Collection

Speedchecker provides access to their measurement network through their API [38] which operates based on a two-request data query model consisting of a start and a get method. In the start method, the test parameters can be specified that control what test to run and how to run it. These include parameters like the ping type, the probes that should be used based on for example the country, AS number (ASN) or the platform, the destination, and which information about the probe gets displayed among other things. This all gets sent as an HTTP request towards the specific API endpoint and returns a test ID which can be used to gather the results through the corresponding get method.

3.3.1. Measurement Implementation & Scheduling

We ran two different types of measurements throughout the measurement period, which lasted from October 2020 to April 2021, TCP pings and ICMP traceroutes. For the

pings, we ran five pings for each probe, while we used three measurements for each of the hops in the traceroutes. Alongside these, we collected information about available probes, at the beginning only once a day before starting with the other measurements, later on increasing it to six times spread evenly throughout the day in 4-hour intervals. However, because we only save a probe when its Speedchecker probe ID is not yet in the database, this happens under the assumption that the probe ID does not change, which may have introduced a small number of artifacts into our later analysis. To automate all of this, we created a python package in which we implement the different steps of the data collection and processing. We conducted the measurements by cycling through the continents, going through all of the countries within a continent in which we observed at least 100 unique probes, running a test towards each of the data centers on the same continent as the current country. We set this restriction since the number of measurements was limited through a quota limit and only one destination can be specified per API call. For the last month we then fully switched the measurements over to intercontinental measurements from Africa towards Europe and North America as well as from South America to North America to get further information since the number of data centers in these continents is comparatively low. As to not overload and slow down the API while still making full use of the quota limit, the script sent out requests at a rate of 1 request per minute. To get the result, which is returned in JSON format, it then checked the state of the measurement every minute through an HTTP GET request, only saving it when the measurement is fully finished.

Overall, this lead us to a total of over 3.8 million data points for pings and over 7 million for traceroutes with measurements within Europe contributing to over 50% of total measurements. Following this is Asia with around 20% and North America with around 10%, while intracontinental and intercontinental measurements in other continents are comparatively low in numbers which is due to the geographical distribution of the probes and the way that we schedule measurements.

3.4. Data Processing

After a measurement has finished, the JSON result is then parsed and further processed.

Before storing the gathered data, we first filtered out erroneous data and enriched it through other sources afterward. In the case of the available probes, the only information to add was the continent the probe is located in, which was gathered from a python dictionary that had been parsed from a JSON dataset of John Snow Labs [39]. Regarding the ping measurements, the only modifications that needed to happen were the filtering of unfinished pings through their returned status codes and the removal of pings where no probe id was shown. After that we also checked whether

five measurements were present for each of the pings, defaulting to -1 where necessary to represent missing measurements.

For the traceroutes, some more modifications were necessary. We first filtered out traceroutes that had not finished or which had no associated probe id, similar to the pings. In the remaining traceroutes, we then went through the hops that were shown in the JSON, ignoring hops where the IP address was not present at all and after a few weeks also started to exclude hops that were either in the private IP address space or in the link-local address space [40]. For all other hops, we got the three round-trip times (RTT), again defaulting to -1 for missing measurements. Afterward, the data of each hop was enriched in several steps, starting by attempting to look up the number of the AS to which the IP belonged to. This was done through either the CAIDA Internet eXchange Point (IXP) dataset [41] in case the IP address was located within an IXP or through the use of the pyasn python package [42]. In both cases, the ASN was derived from an IP-prefix lookup. For the CAIDA data set, this was done through a lookup after parsing the relevant file into a python dictionary, while the pyasn package uses IPASN data files which are generated through an included script by converting BGP archives from Routeviews or similar sources. Following this, we then used the ASN to get the name and type of the organization managing this AS through the help of PeeringDB [43] as well as the organization name as found in the CAIDA data. Lastly, we also added whether the IP belonged to an IXP as well as geographical information which was either derived from the CAIDA dataset or through the GeoIP lookup API [44]. All of the used datasets were downloaded in October 2020 and not further updated in order to keep the data, e.g., the organization names, consistent.

3.5. Database

After all measurements had been processed and modified they were then stored into an SQLite database, which allowed us to query the results as necessary for our analysis. The database structure was kept similar to the database used to store the RIPE Atlas data in [11] as much as possible in order to be able to use the data from both datasets for our analysis without having to use two completely different approaches.

The database can be split into three major components: the pings, the traceroutes, and the meta-information needed for both of the other parts. A general overview of the main tables can be seen in Fig. 3.3 which is briefly explained in this chapter; the detailed explanation of the table columns and their values can be found in Appendix A.

The ping data was stored into a single `Pings` table, where we stored the ID of the probe that was used which can be used to get the specific details about it from the

`Probes` table, the URL of the targeted data center with the underlying information in table `Datacenters` as well as timestamp and the measured values.

The traceroute data was split into different tables to organize it. This included a `Traceroutes` table containing an overview of the issued traceroutes with information like timestamp and destination similar to the `Pings` table and a `Hops` table containing the measurement data for each specific hop, namely IP address, hop number, the measured latency values and the ID of the traceroute the Hop belongs to. In addition, the `NodeInfo` table contains the enriched data regarding the IP addresses which were encountered as Hops including the PeeringDB and CAIDA data as well as some geographical information about the node.

Lastly, there are the already mentioned `Datacenters` and `Probes` which contain the supplementary data, as well as two tables that are exclusive to our database and were used for further private investigations. We stored information about the time when specific probes were observed into the `ProbeActivity` tabel while the `Relationships` table was created based on the CAIDA AS relationship data [26] and stores data about relationships between different ASes which have been inferred by CAIDA.

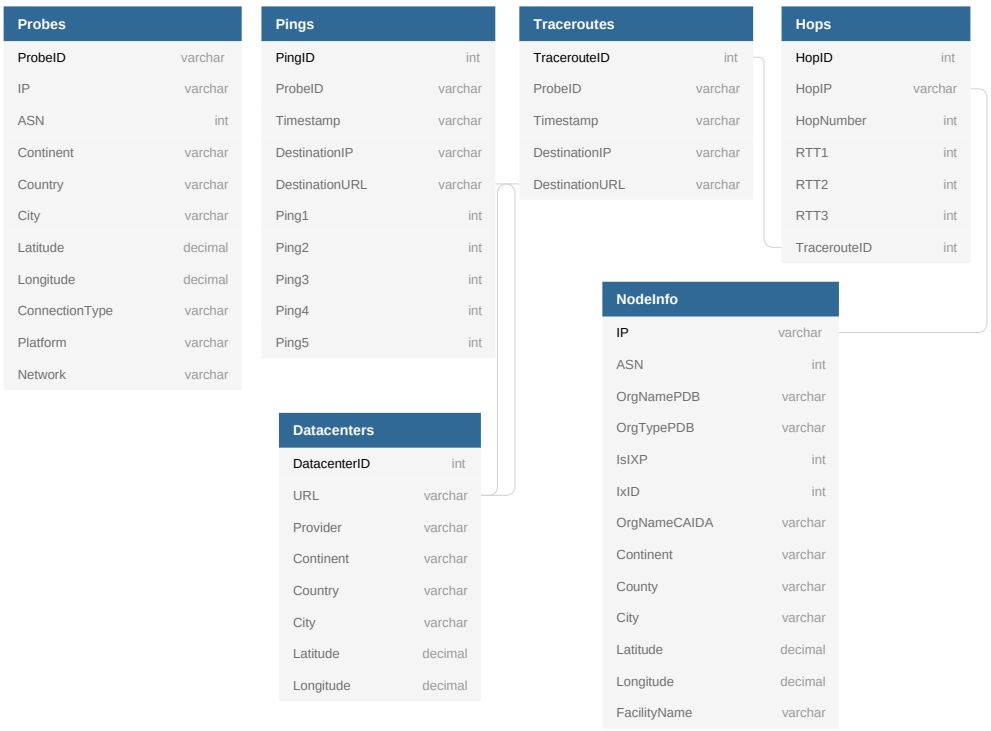


Figure 3.3.: Overview of the database tables

4. Analysis

In this chapter we will analyze the results from our measurements in order to fulfill our objectives and answer our questions. We will also compare our results to the results from the RIPE Atlas data for applicable aspects.

4.1. Clarifications

Before we start with the actual analysis, some clarifications have to be made. First, Amazon Lightsail data centers do not respond to ICMP measurements which means that there is no traceroute data for these. Secondly, due to a faulty address, we do not have sufficient measurements for measurements towards Google's South America data center. As such, results for these specific cases have been excluded from the analysis. There are other data centers that did not respond consistently as well (especially to traceroutes) but in most cases, there are enough data centers in these regions that do deliver results.

Thirdly, we noticed that for Amazon specifically, almost all of their internal routing cannot be uncovered with our methodology detailed in chapter 3.4. This is the case because either their internal routers did not respond to the ICMP packets issued during the traceroute or we could not identify the accompanying ASN for a specific IP address. To alleviate this issue slightly, which is especially impactful when looking at the pervasiveness in the traceroutes in a later section, we corrected some of the entries in the `NodeInfo` table in our database by rechecking whether the IPs where no ASN could be determined belong to Amazon with the help of the published IP address ranges from Amazon themselves [45] and set the ASN accordingly. This was done after our measurements had finished in April, so there are still some inaccuracies due to the delay between the measurements and the correction.

Lastly, unless explicitly stated, the measurements that are analyzed will only include intracontinental measurements from our Speedchecker data and when we mention the RIPE Atlas data, we refer to the data used by Corneo et al. in [11].

4.2. Global Cloud Access Latency

The first aspect which we want to investigate is how good user latency towards the cloud is globally. For this, we only take TCP ping measurements towards the closest data center (in terms of lowest average latency, not actual geographical distance) into consideration for each of the probes to show the best case results. In this context we will also use three common latency thresholds that were used in [11] in order to put these latencies into perspective:

- **Motion-To-Photon (MTP)** at around 20 ms, which is important for latency-critical applications like Augmented Reality (AR) and Virtual Reality (VR) to avoid motion sickness and dizziness.
- **Perceivable latency (PL)** at around 100 ms when the delay between user input and visual feedback becomes noticeable for humans which is relevant for applications like video streaming and cloud gaming as well as for autonomous driving [46].
- **Human Reaction Time (HRT)** which is the delay between a stimulus and the according response and set at around 250 ms. This is important for applications like remote surgery where active human engagement is necessary.

4.2.1. Intracontinental Cloud Access Latency

The results when looking at intracontinental latency are shown in Fig. 4.1 in a map for all of the countries where we have gathered measurements from. From this figure, we

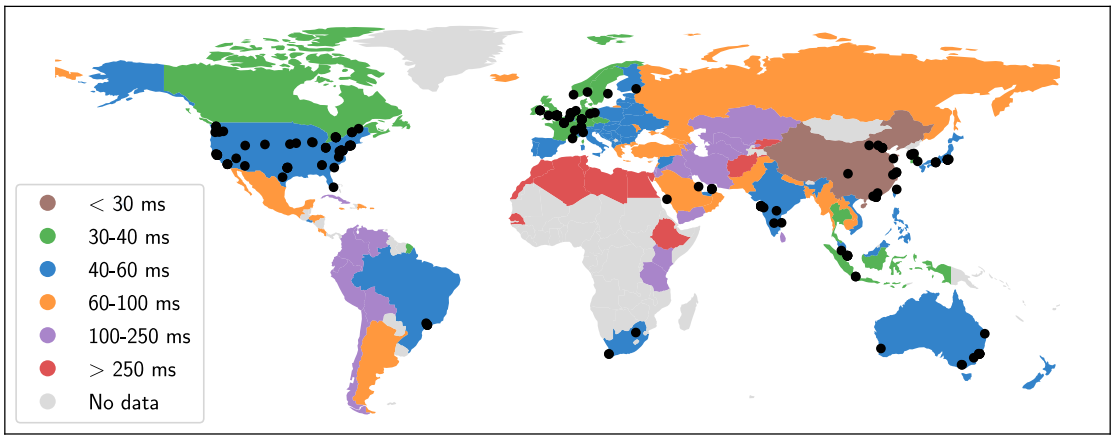
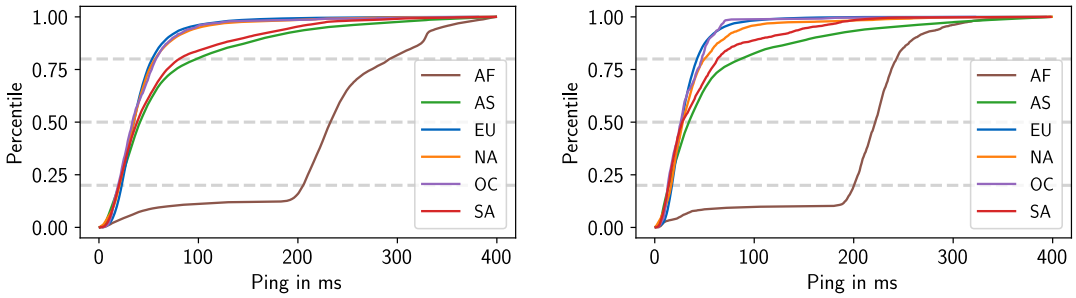


Figure 4.1.: Average of the TCP latencies from all probes in each country towards the closest data center with data center locations as a reference

can see that latency across Europe is very good with users in almost all countries being able to reach a cloud data center within 60 ms. The same can be said for Oceania. For other continents, however, the results are more mixed. In North America for example, users in the US and Canada also experience an average latency below 60 ms, while probes in almost all other countries in the region can reach a cloud data center within 100 ms, the PL threshold. For the other continents, results are even more diverse, ranging from countries where the average connection almost fulfills the MTP threshold to countries where latency is even outside the HRT boundaries. In general, we can see that distance towards data centers (shown as the black dots) plays an important role in user access latency. Users in countries with data center presence or within close proximity experience much better latency (in most cases below 60 ms, but at worst under 100 ms) compared to countries that are far away from data centers which will be more closely investigated in a later section.

We have also plotted the measurements grouped by continent instead into an empirical cumulative distribution function (ECDF) in Fig. 4.2a. We can see that user access latencies from Europe, North America, and Oceania are very similar with South America and Asia following closely behind, while Africa is last by a large margin. These distribution shapes can be explained when we combine this figure with Fig. 4.1 and the number of probes in the specific countries of the continents as well as the data center locations which are also included in the map. Even though there are several countries in North America (or more exactly Middle America) that have worse average latencies than most countries in Europe or Oceania, the distribution when looking at the continent as a whole is roughly the same for all three continents. This is the case because over 90% of the measurements come from probes in North America which are located in either Canada or the US where the measured latency is under 60 ms and as such the distribution is mainly shaped by these measurements. The longer tails of the distributions for South America and Asia exist due to the larger variances between the different countries.

Over half of the countries in South America show an average latency of over 100 ms, while only Brazil has a latency of under 60 ms which is because the only data centers are located in south-eastern Brazil, far away from most other countries in the continent. In Asia, most of the data centers are located in the eastern parts of India with a few of them also being in the Middle East. The correlation between distance to a data center and average latency that we will investigate more closely in a later section, is very visible due to this data center distribution. While all countries in East Asia and around India have an average latency of under 100 ms, most countries in the Middle East and Western Asia have poor latencies above 100 ms or even above 250 ms in some cases. Additionally, most of these countries with higher latency are also developing countries where the whole internet infrastructure is likely not as far



(a) All latencies

(b) Minimum latencies

Figure 4.2.: Comparison of all vs. only the minimum TCP latencies from all probes towards the closest data center grouped by continent for Speedchecker

developed yet. Comparing the distributions for South America and Asia shows that for our data set, South America seems to be ahead, but this is again because over 80% of our measurements originate from probes in Brazil where the data centers are located, while for Asia the number of measurements is more distributed across all of the countries.

Lastly, the ECDF for Africa plateaus between 75 and 200 ms with around 10% of the measurements being able to reach the closest data center within this time frame. After the 200 ms mark, it then starts to rapidly increase. This is also due to the geographical distribution of our probes. Roughly 10% of recorded measurements came from probes in South Africa where the only African data centers are located in. As such, as can be seen in Fig. 4.1, the average latency for South Africa is relatively low at between 40 and 60 ms which represents the first part of the ECDF. All other African countries where we have gathered measurements from have a considerable distance between them and the data centers in South Africa leading to very poor latency averages of over 250 ms for all countries except Kenya. These are represented in the second part of the ECDF after 200 ms.

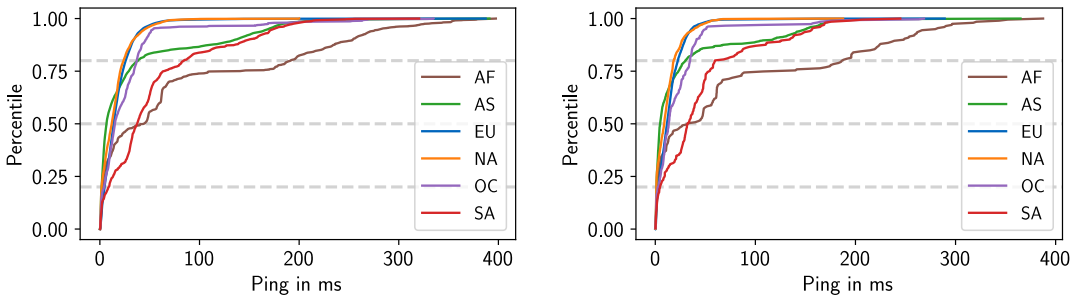
The results we get when we only take the minimum ping to the closest data center, which is the absolute best case latency that was observed throughout the measurements, are then shown in Fig. 4.2b. For most continents, we only observe a slight shift of the curve to the left indicating that the gathered latency is relatively consistent for the majority of probes as the average latency does not differ majorly from the minimum. The small changes that are present may be due to the inconsistency of wireless connections mentioned in chapter 2.4 and general measurement variance. The two continents where we do observe more impactful differences are again Africa and

South America, indicating that the measured latency in these areas fluctuate more than in other regions which may be the result of the infrastructure there not being as well developed leading to measurable differences depending on the time and load that is placed on it.

4.2.2. Comparison with RIPE Atlas

When we compare our Speedchecker measurements to the RIPE Atlas data, which is shown in Fig. 4.3a, we can see that while the distributions are similar for Europe, North America and Oceania, the ECDFs for the other three continents differ considerably. This is caused mainly because of the different probe distributions throughout the continents between the two platforms. The probe distribution for RIPE Atlas probes has been gathered from the probe density map that RIPE Atlas provides [47] as well as our database. In the case of South America, the RIPE Atlas data set contains measurements from probes which are more evenly spread throughout the countries with less than 40% of them being located in Brazil where the data centers are compared to our Speedchecker database where this number is over 80%. This leads to the slower, but more steadily increasing curve for RIPE Atlas. The same holds for the ECDF for Africa, where instead of a large number of probes being focused more in the northern half of Africa, most of the probes of RIPE Atlas are located in or close to South Africa leading to a steeper curve. It is also the same for Asia where RIPE Atlas probes are mainly located in and around countries with data center presence and fewer or zero probes in Western Asia where we record the worst average latency in our Speedchecker measurements.

We can again compare these average latencies against the minimum observed latency



(a) All latencies

(b) Minimum latencies

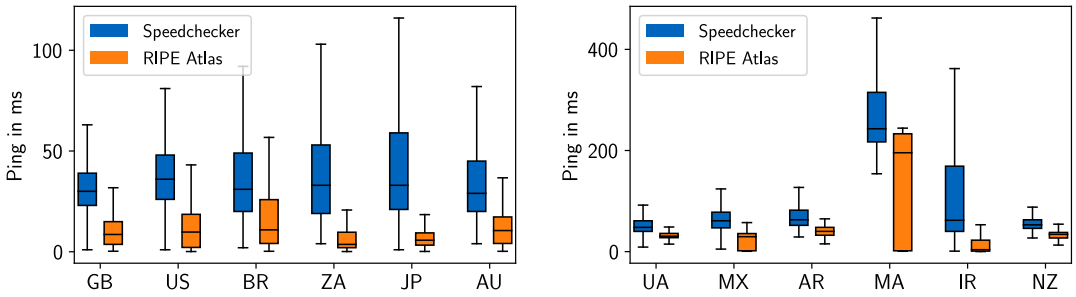
Figure 4.3.: Comparison of all vs. only the minimum TCP latencies from all probes towards the closest data center grouped by continent for RIPE Atlas

in Fig. 4.3b. In this case the differences between the two are even smaller than in our Speedchecker data with no really impactful changes for any of the continents. This might be the case because RIPE Atlas probes are mostly wired probes leading to more consistent results which are not impacted by congestion and latency increase due to wireless interference.

In general, we can see that the latency gathered from the RIPE Atlas probes is much lower than the data from our Speedchecker probes, especially for Europe, North America, and Oceania, where the probe distribution is similar for both platforms and thus more comparable. In these three cases, around 80% of the RIPE Atlas probes can reach a data center within 30 ms, while for Speedchecker this 80% mark is reached only at around 50 ms. We also note that while there are a lot of countries in the RIPE Atlas data set which meet the MTP threshold of 20 ms on average, there are no countries where this is the case for our data from Speedchecker. There are several likely causes of this. First, while we focused on Android-based software probes for our Speedchecker measurements, RIPE Atlas probes are dedicated hardware probes that are connected to a router or switch via Ethernet [32]. Secondly, RIPE Atlas hosts probes in many different environments, e.g., organizations like ISPs, IXPs, and even data centers, which differ from the Speedchecker probes which are placed exclusively on the last-mile [4]. Lastly, the absolute probe number differs considerably between the two platforms, with slightly more than 8,000 in the RIPE Atlas database versus around 110,000 probes which were used for our Speedchecker measurements, leading to more variance in the measurements. While these points mean that RIPE Atlas allows for more measurements in different environments with probes that are more consistently available, we believe that our measurements from Speedchecker more accurately represent the end-user connectivity.

4.2.3. Impact of Geographical Location

We already noticed an increase in latency with further distance towards a data center in chapter 4.2.1, now we want to take a closer look at some example countries to check whether this is consistent across the world and how large the impact actually is because only focusing on the average latencies does not reveal the whole picture. We chose twelve countries, two from each of the continents, where one country has at least one data center within its borders and another country that does not. The results are shown in Fig. 4.4 for countries with and without local data centers respectively. For this, we also only take measurements towards the closest data center for each of the probes into account to get an accurate representation of the impact that geographical proximity has on latency. For the first group of countries with in-country data centers, we can see that even though there are differences in the consistency of the latency across the countries,



(a) Countries with data centers

(b) Countries without data centers

Figure 4.4.: Comparison of TCP latency towards the closest data centers between countries with differing data center presence

the median for all of the countries are very similar for both our Speedchecker and the RIPE Atlas measurement results, around 30 ms for the first and 5–10 ms for the latter. Curiously, we also see that while South Africa and Japan display the highest amount of inconsistency among the countries for the Speedchecker results, it is exactly the opposite for the RIPE Atlas data. This may simply be the result of the number of probes and their locations for the specific countries since RIPE Atlas hosts a smaller amount of probes and also probes in different environments like data centers and organizations in addition to home user probes.

When we then look at the other countries without a data center in Fig. 4.4b, we immediately witness an increase in overall latency. Most of these countries display a latency of between 50–70 ms in their median for Speedchecker and around 30 ms for RIPE Atlas. This is already a substantial increase of over 100% compared to the counterparts in the same continent seen in Fig. 4.4a, but this is exacerbated for Morocco since the geographical distance for this specific case is several times larger than for the other countries. Aside from that we again see a similar relative order comparing the Speedchecker and RIPE Atlas except for Iran where the range of observed latency is particularly large for the Speedchecker measurements.

These results show that geographical distance is and will also likely always be one of the largest factors in overall latency because of the physical limitations. This is also a relevant factor for the following section.

4.2.4. Intercontinental Cloud Access Latency

Because data center coverage in Africa and South America is very sparse and their locations are clustered into only one area, intercontinental connectivity becomes a

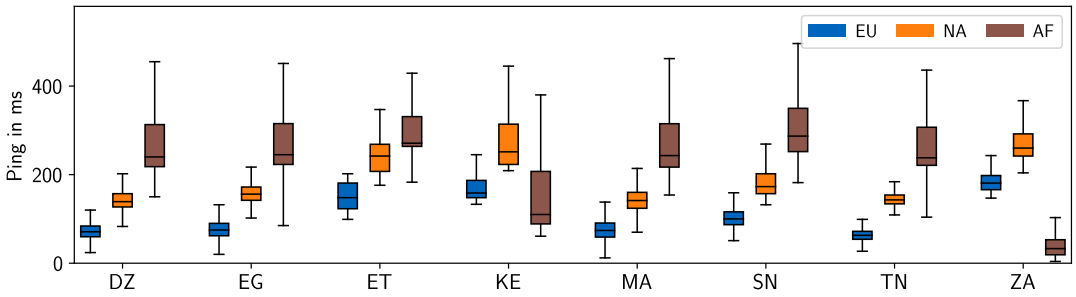


Figure 4.5.: TCP latency from probes in Africa towards the closest data center in different continents

relevant topic when looking for the best-case latency. Therefore, we will take the intercontinental latency towards the closest data center into consideration for these two cases.

For Africa, we look at both the intercontinental latency towards Europe and North America in comparison to the intracontinental measurements shown in Fig. 4.5. The first point to note is that latency towards data centers in North America is consistently higher than the latency towards European data centers for all of the countries we look at, with a difference of 50–100 ms, which is reasonable because of the closer proximity for the latter. The impact of the geographical distance can also be seen when looking at the countries individually: Countries located at the northern coast of Africa, i.e., Algeria (DZ), Egypt (EG), Morocco (MA), and Tunisia (TN) display the lowest intercontinental latency at around 80 ms when targeting Europe while simultaneously showing the highest intracontinental latency towards the South Africa data centers. This means that while most probes in these countries do not reach the HRT threshold when targeting data centers within the continent, latency below the PL threshold can be achieved via connections to the closest European data center for a majority of them. Following this are the countries which are located further to the south, Ethiopia (ET), Kenya (KE), and Senegal (SN) which display a slightly higher intercontinental latency at around 100–150 ms which is still lower than the intracontinental one except for Kenya. Lastly, South Africa (ZA) has the highest measured intercontinental latency, while the intracontinental latency is the lowest by far due to the presence of data centers within the country.

Because we have already seen that the distance plays a large role in the overall latency twice, we only look at the intercontinental measurements from South America towards North America in Fig. 4.6. We can see the same general trend in this case where countries in the northern parts of South America, i.e. Colombia (CO), Ecuador

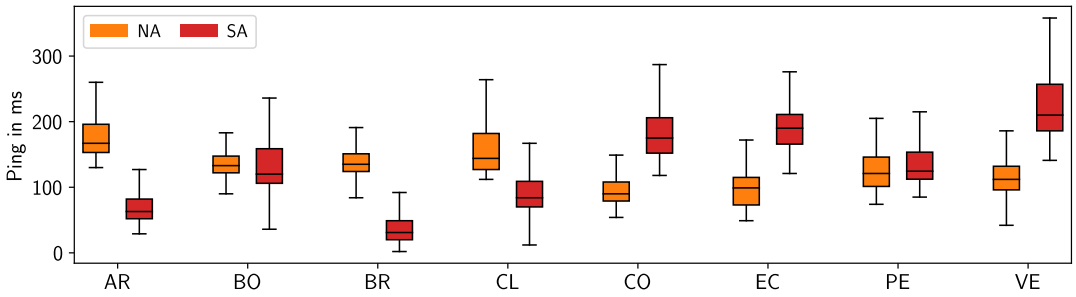


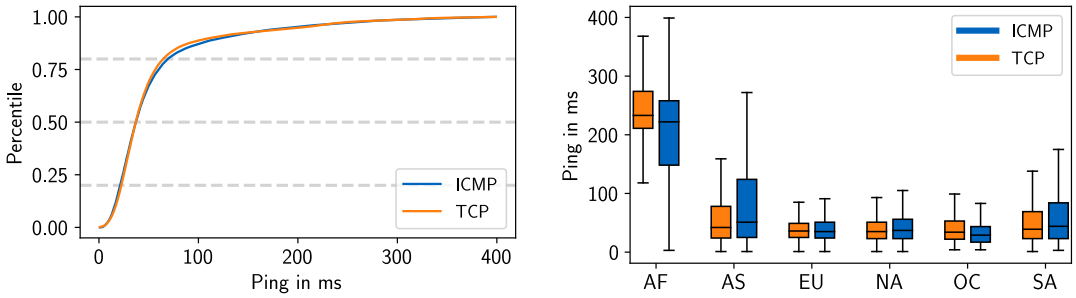
Figure 4.6.: TCP latency from probes in South America towards the closest data center in North America

(EC), and Venezuela (VE) show the lowest latency when targeting North American data centers while displaying the highest intracontinental latency among all of these countries. Following these are Bolivia (BO) and Peru (PE) which are farther to the south and closer to the South American data centers reflecting accordingly in the latencies. Lastly are Brazil (BR), which has in-country data centers, as well as Argentina (AR) and Chile (CL) which are located at the southern tip of the continent where the intracontinental latency is much lower than its intercontinental counterpart.

For both these continents, we can see that the latency is directly proportional to the geographical distance, which further supports our observations from the previous section. But additionally, other factors like the general Internet infrastructure also likely play an important role in the latency for these continents specifically because we can see that for some countries like Bolivia, Peru, and Ethiopia, the absolute distance is not necessarily lower for the intercontinental measurements while the latency is at least as good.

4.2.5. ICMP vs. TCP Latency

We also take a look at the latency gathered through the TCP pings and compare them to the last-hop RTT measured in the ICMP traceroutes. The results are plotted grouped by continent in Fig. 4.7b and combined in Fig. 4.7a. In the combined plot it can be seen that the TCP latencies measured are at least as low as the ICMP latencies and sometimes a little bit lower. This roughly matches the findings from the RIPE Atlas data [9]. However, this difference is very small and could be skewed because of inaccuracies or errors, especially for the ICMP measurements which are known to be less accurate and responsive than TCP-based measurements [48]. The results of these errors can partly be observed in Fig. 4.7b. While the TCP latency is lower



(a) Combined global latency

(b) Grouped by continent

Figure 4.7.: Comparison between ICMP and TCP latency measurements

and has a smaller variance than the ICMP latency for Asia, Europe, North America, and South America, the results for Africa and Oceania are different. The difference in measurements in Africa is likely due to the difference in the number of measurements for the two protocols. There are seven times as many measurements for TCP pings compared to ICMP traceroutes for Africa and while measurements from South Africa only comprise roughly 10% of the total TCP ping measurements, in the traceroutes this percentage is around 30 which skews the latency in favor of ICMP because of the closer proximity of the probes to the data centers. For Oceania, there are only around two and a half times as many measurements for TCP vs. ICMP, but the latency difference between these two is also smaller which might be caused by the different sample sizes.

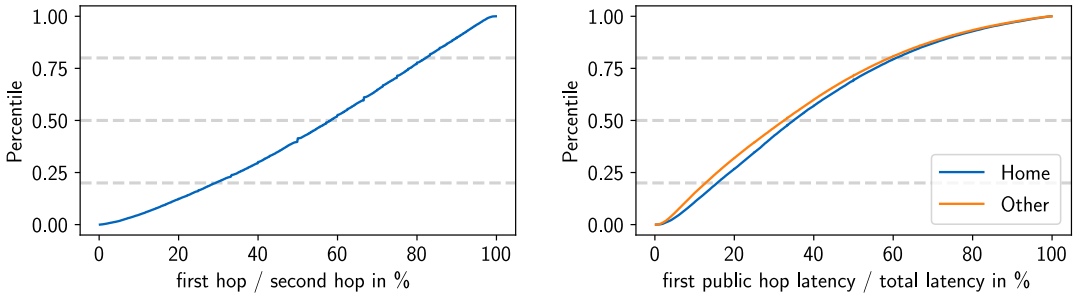
We therefore conclude, that TCP-based measurements are likely more accurate and reflective of real-world latency than ICMP-based ones, since they are not as prone to errors and non-responses and because TCP is actually used for data transmission and by network applications, while ICMP is used mainly for sending error messages and operational information.

4.3. Impact of the Last-mile and Wireless Connectivity

Since last-mile latency is known to be a large, if not the largest, factor regarding overall latency, we want to use our data to investigate this aspect as well as the impact of the wireless connectivity on total latency. To do this, we take our traceroute measurements and compare the latency of the first hop to the overall latency, which is obtained by looking at the latency of the last hop. To further differentiate our measurements into measurements from probes connected wirelessly in a home network vs. probes using another connection type, we first check whether the first hop is within the private IP

address space. If this is the case, we then check if the second hop is a public IP address and classify the measurement as a measurement from a home network only if both conditions apply. Since we decided to not record private IP addresses after a few weeks of our measurements, the number of measurements obtained belonging to this group is smaller than the second group. Otherwise, if the first hop is a public IP address, we treat it separately. For this second group, we only take the measurements into account where this public IP address was recorded on the hop with hop number one to filter out measurements from home networks where the private IP address was not recorded. However, because Speedchecker does not provide specific connection types for our Android probes, this group can contain measurements from cellular devices as well as possibly a few artifacts from probes using a VPN and other sources which we cannot identify.

We first take a look at the home network probes globally and how much latency the hop with the private IP address, which represents the wireless connection to the router, contributes to the overall last-mile latency obtained from the second hop, which represents the connection between the home router and the ISP. This is shown in Fig. 4.8a, where we can see that over half of the measurements show the wireless connection being responsible for over 50% of the total last-mile latency. Additionally, the distribution is very linear, showing that the latency introduced through the wireless connection varies massively, most likely due to the specific location and how much congestion or general traffic exists because of other wireless devices present nearby. In Fig. 4.8b we then plot the total last-mile latency from the home probes compared to the latency obtained from the second group labeled as "Other". We see, that surprisingly, both of the distributions look very similar to each other, with the last-mile latency contributing to up to 60% of total latency for 80% of the measurements with the median



(a) Impact of wireless connectivity

(b) Total last-mile latency

Figure 4.8.: Last-mile latency based on traceroute measurements

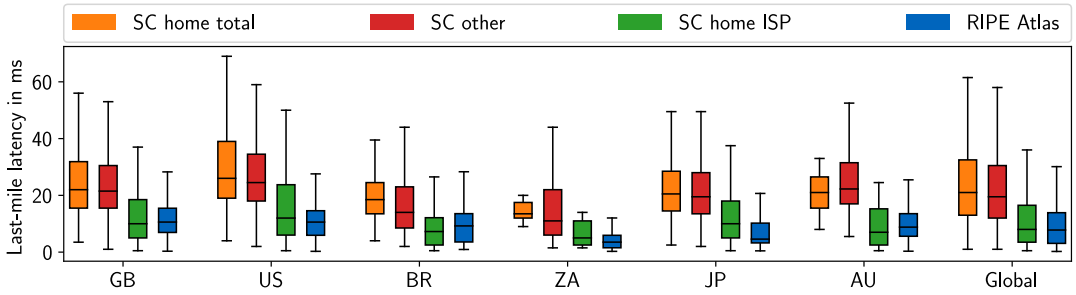


Figure 4.9.: Comparison of absolute latency in different countries between Speedchecker and RIPE Atlas

at around 35% of total latency. In terms of total latency, this means an overall latency of around 20 ms at the median and 35 ms at the 80% mark.

Lastly, we look at the last-mile latency for different countries in addition to the global results and compare these to results from home probes in the RIPE Atlas database. The results are plotted in Figs. 4.9 and 4.10 with absolute and relative values respectively. For Speedchecker, we plot three different boxes: one for the total last-mile latency of the home probes, one for the total last-mile latency of the "Other" group, and a last one for the latency that exists between the home router and the ISP, which is obtained through subtraction of the latency of the first hop from the second hop. Since there is no indication for the connection type in the RIPE Atlas data, we assume all of the measurements to be from dedicated hardware probes that are connected via Ethernet and therefore plot only one combined result. When looking at the absolute values, we see that they are very comparable throughout the different countries, with the median being around 20–25 ms for both of the Speedchecker groups, with similar distributions which we have already seen in Fig. 4.8b, and close to 10ms for the RIPE Atlas measurements. We can also see, that the latency from the two different platforms is very similar when we remove the wireless part of the measurements, only keeping the wired portion to the ISP from our Speedchecker data. This shows again that the wireless connectivity does add quite a lot of extra latency — around 15 ms for the globally combined results. Additionally, this confirms, that a majority, if not all of the measurements from the RIPE Atlas data set, are most likely obtained from probes that have a wired connection.

When we then look at the relative latency, i.e., how much the last-mile latency contributes to overall latency in Fig. 4.10, there is much more variance between the countries. While for the US and Japan this percentage value is relatively low, the other countries show a much higher value. There are several factors that could lead to these

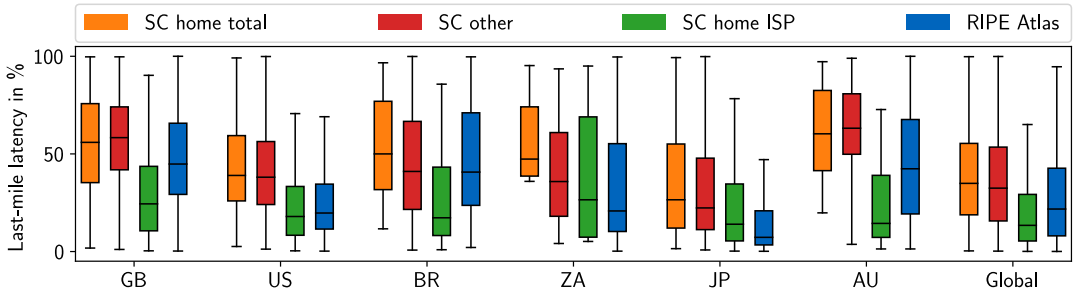


Figure 4.10.: Comparison of relative latency in different countries between Speedchecker and RIPE Atlas

discrepancies. First, since we include all intracontinental measurements for the last-mile analysis, there could be differences that arise from the geographical factors. For example, the US is a relatively large country with many data centers located throughout the country, so there are measurements from one side of the country towards the other side included, where the last-mile latency might not contribute as much of the overall latency as it would for other measurements where probes are located closer to the data centers. Similar things can be said for Japan, where measurements go towards distant data centers all over Asia and even towards the Middle East. This, inversely, could also be the reason why our measurements from Great Britain show such a high percentage contribution since a majority of the European data centers are in Western and Central Europe (Fig. 3.2), in much closer proximity compared to the US. Secondly, the connection between homes and ISPs might be a contributing factor for the other countries like Brazil, South Africa, and Australia, where especially in rural regions, away from large cities, the infrastructure is relatively poor, which may lead to higher last-mile latency.

Overall, the results are mostly consistent between both the Speedchecker and RIPE Atlas data sets and we see that the last-mile connection contributes a significant portion of latency on the Internet and that the wireless connectivity, no matter whether it is from home networks or through mobile connections, also has a large impact on that. The last-mile is also the section of the path is also the one where the cloud providers and other large networks have the least amount of control over, which is why it is difficult or even impossible for them to improve this section.

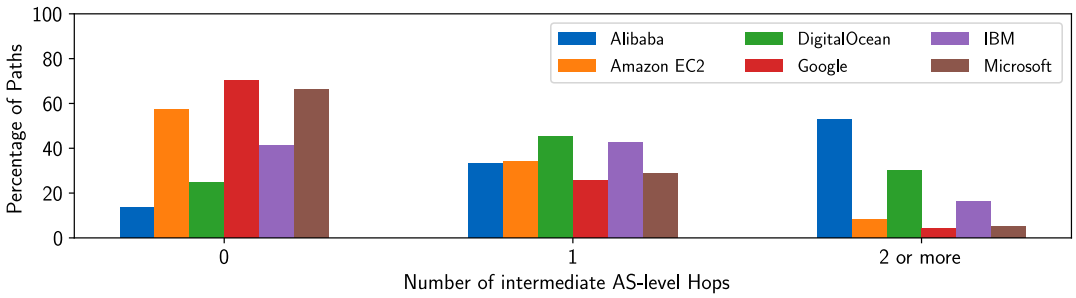


Figure 4.11.: Number of distinct intermediate ASes encountered on the way to data centers

4.4. Path to the Cloud

In the next sections, we investigate the Internet topology, looking at network interconnectivity and cloud provider presence in traceroutes to determine the current state of the Internet flattening process from a cloud provider perspective.

The first aspect we focus on is the difference in the cloud providers regarding the path to their data centers. To start with, we look at the number of intermediate ASes, i.e., the number of distinct ASes in a traceroute excluding the source AS of the probe and destination AS of the cloud provider, to see how many networks have to be traversed to reach a data center. From Fig. 4.11 we can see that for the three largest cloud providers, Amazon, Google, and Microsoft, a majority of the paths do not traverse any intermediate networks, but instead directly enter the cloud provider network from the source ISP network. An additional 20 to 30% of the paths have a single intermediate AS while there are only less than 10% of paths that have to traverse several ASes. Following behind this group of cloud providers is IBM, where around 40% of the traceroutes do not traverse any intermediate ASes and the same amount traverse a single intermediate AS. Lastly, there is the third group of cloud providers containing Alibaba and DigitalOcean as well as the other cloud providers we have included in our measurements, Oracle, Linode, and Vultr. The latter three follow a very similar distribution to Alibaba and are not drawn for the sake of clarity. In this last group, a majority of the paths traverse at least one intermediate network. But from these five providers, DigitalOcean seems to have a larger amount of direct paths and paths with a single intermediate AS compared to the other four. These results are also in line with the kind of network these cloud providers operate, with the largest three having an extensive private network on a global scale, IBM following a hybrid approach with a private backbone in some regions, while the other cloud providers mainly rely on the

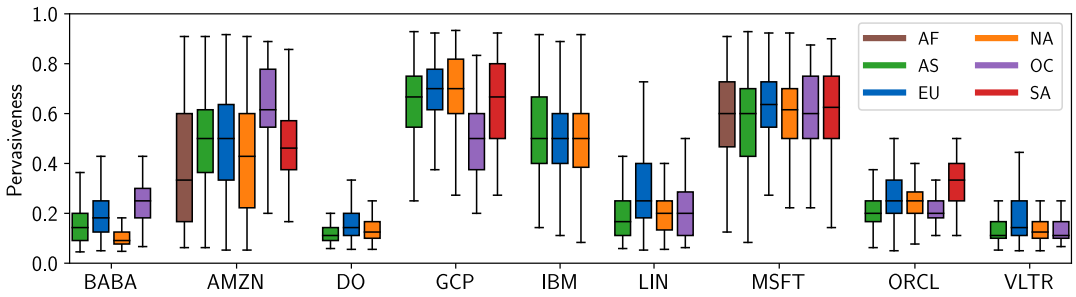


Figure 4.12.: Pervasiveness of cloud providers based on responsive hops

public Internet infrastructure to relay the traffic.

Next, we can also take a closer look at the paths at a router level instead of the AS-level. On the router level, we can determine how much of the hops in a traceroute belong to a cloud provider, which we refer to as pervasiveness, through two approaches. In the first, which is depicted in Fig. 4.12, we only take the responsive hops into account when calculating the pervasiveness on the paths. In this figure, we can see that the pervasiveness of the cloud providers follows a similar trend to the AS-level hop distribution. Google and Microsoft come out on top, with the medians in almost all regions exceeding the 60% mark, meaning that more than half of the hops traversed in the traceroutes of these two cloud providers belong to their own network. Following closely behind are Amazon and IBM, where the medians for most regions are roughly at 50%. The only exception for these first four cloud providers are in the Oceania region, where Google does not seem to have as high of a pervasiveness as in the other continents where their data centers are present, while it is the opposite for Amazon, with the Oceania region showing a higher pervasiveness. The other five cloud providers who do not operate their own private network infrastructure come out at the bottom in this aspect as well, with the medians in most regions showing a pervasiveness of 20% at most. Still, all of them show a higher pervasiveness in Europe compared to the other continents except Alibaba and Oracle who show a better pervasiveness in Oceania and South America respectively, but in these regions (as well as Africa) we have far fewer probes and measurements than in the other three regions as well as a smaller spread over different countries for Oceania specifically, which may lead to some inaccuracies.

In a second approach, which is shown in the boxplot of Fig. 4.13, we also include unresponsive hops and calculate the number of hops belonging to a cloud provider by subtracting the number of the first hop which belongs to the cloud provider from the total number of hops. The result is then divided by the total number of hops to reach the pervasiveness. The changes in the pervasiveness can then be characterized as

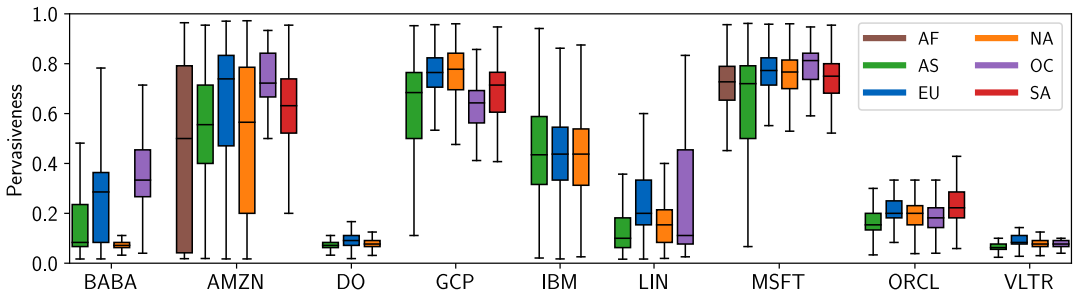


Figure 4.13.: Pervasiveness of cloud providers including unresponsive hops

follows: If the pervasiveness increases compared to the result obtained by the previous method with the responsive hops it means that there were more unidentified hops after entering the cloud provider network than before it and vice versa for reduced pervasiveness. These changes, which we can see in Fig. 4.13, also follow a similar trend within the three groups of cloud providers based on their network backbone seen previously. For the cloud providers operating a private network the pervasiveness generally increases with this second approach, while there are barely any changes for IBM and a generally reduced pervasiveness for the group of cloud providers using the public Internet. These changes can mostly be explained through the type of network backbone that the cloud providers operate and also with the help of Fig. 4.14. Cloud providers with a private backbone route a lot of their traffic internally, where not every router is configured to respond to the ICMP requests sent out during the traceroute and many routers are not correctly inferred to be within that AS. This is especially the case for Amazon, where most of the time only the border router as well as the destination data center could be identified as belonging to Amazon with our methodology. This

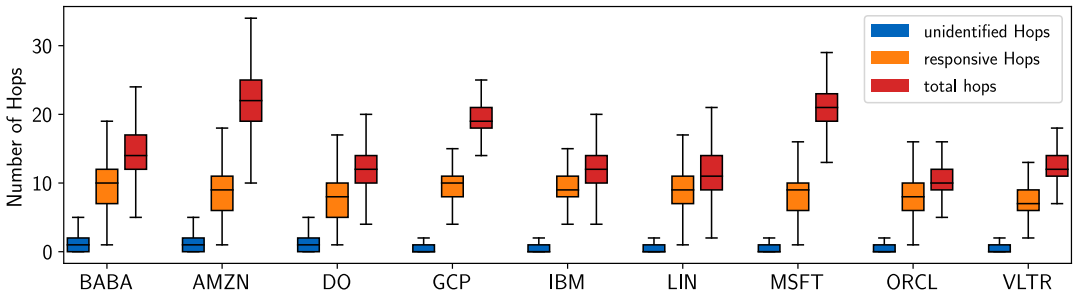


Figure 4.14.: Overview of different hop information

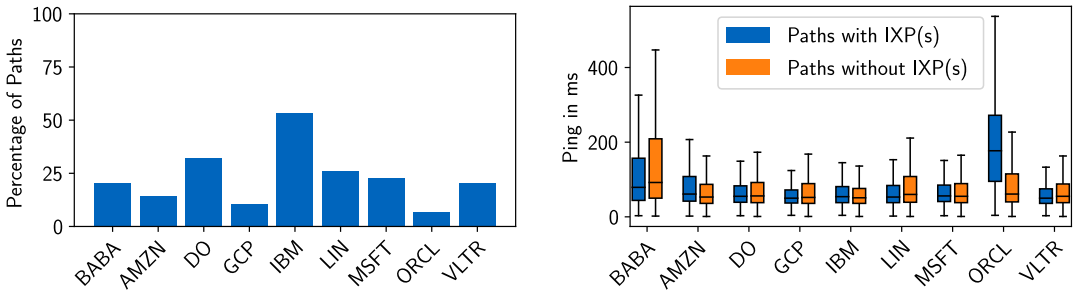
can also be seen in Fig. 4.14, where these three cloud providers have the highest amount of total router hops within the traceroutes while the number of responsive hops is comparable to the ones of all of the other cloud providers. Within this group, Amazon also has more unidentified Hops than Google and Microsoft. Therefore, the change using this second approach is especially large for Amazon, but all of them show a generally higher pervasiveness than before. For IBM, which operates a private network in some regions, there are barely any changes, only in the upper and lower bounds, influenced by paths going over the public Internet and paths going over their own network alike. Lastly, the cloud providers without a private backbone generally show a lower pervasiveness with this second approach, since they own few routers on the paths to their data centers, most of which can be correctly assigned to their respective AS. An exception in this case is Alibaba, which has a similar amount of unidentified hops to Amazon, which leads to a large increase in the upper bounds of the boxes and larger changes compared to the other providers in this group.

While the second approach leads to a likely more accurate result at least for Amazon, we believe that there are too many uncertainties when including unresponsive hops into the results in regards to hop numbers and the actual path taken. Therefore, for further analysis in the later sections, we will follow the first approach, only considering responsive hops, while keeping the results of the second method in mind.

4.5. Cloud Provider Presence in IXPs

Since IXPs are one of the driving factors of the Internet flattening process, we next want to look at the presence of the different cloud providers in them based on our traceroute results.

The occurrence of IXPs on the path to the cloud providers is shown in Fig. 4.15a. We can see that IXPs are traversed in roughly 20% of the traceroutes for most of the cloud providers with notable outliers being IBM, with more than 50% of the paths going through at least one IXP, as well as Google and Oracle on the other side, where less than 10% of the traceroutes contain an IXP. For the group of large cloud providers, Amazon, Google, and Microsoft, the traceroutes containing IXPs are mostly limited to measurements originating from Eastern Europe targeting the data centers in Central and Western Europe. For IBM we can generally observe that most of the measurements that originate from a probe in the same country as the target data center do not traverse an IXP, while cross-country paths do, but even then it is not as clear as for other cloud providers. The rest of the cloud providers show a similar trend where IXPs are traversed for a majority of the measurements within Europe, while measurements in other regions are mostly without IXPs on the path.



(a) Frequency of IXP occurence in traceroutes (b) Impact of IXP occurence on the latency

Figure 4.15.: Frequency and effect of IXPs in traceroutes

Since the IXPs are used for public peering interconnections between different ASes, these results can also be explained through the different network infrastructures that the cloud providers operate. The cloud providers operating a global private network do not have to rely on public peering in most areas, instead, establishing direct, private peering agreements with other networks, allowing for faster absorption of traffic into their own network and as such, more control over the routing. IBM with their hybrid approach, do have localized private networks and can rely on private peering for shorter paths in mainly Europe and North America, but need to use IXPs for the longer paths in these regions, as well as possibly relying on transit interconnections in other continents. And since a majority of our measurements are based on European probes, this leads to this higher occurrence rate of IXPs in the recorded traceroutes. For the smaller cloud providers relying on the public Internet infrastructure, they are just not present in as many IXPs as the larger ones and instead need to rely on peering or transit interconnections with Tier 1 ISPs to connect to end-users.

Lastly, we also take a short look at the latencies observed in the traceroutes grouped by whether an IXP was traversed or not. From Fig. 4.15b we can see that there are small differences for most of the cloud providers with notable differences observed for Alibaba and Oracle. These differences mainly occur because we take the global measurements into account, which leads to an uneven representation of continents within the two groups. For example, for Alibaba, where traceroutes with IXPs display a lower latency, this occurs because the measurements with IXP occurrence are mainly gathered from shorter paths in Europe as well as paths in Asia to data centers located outside of China, while the measurements without IXPs are mainly seen in traceroutes from all over Asia towards China. For Oracle, this is exactly the opposite, where almost all of the traceroutes with IXP occurrence are observed in Asia, particularly

in measurements targeting India data centers, while the other group contains mostly measurements in North America and Europe, which leads to a lower observed latency for the latter. When looking at localized measurements focusing on specific areas however, there is hardly any observable difference between paths traversing IXP(s) and paths without.

4.6. Peering

In this last section, we look deeper into the interconnections between different ASes for two sample countries. We specifically try to separate the interconnections between ISPs and cloud providers into two main groups: direct peering, where the traffic goes directly from the source AS of the probe into the cloud provider AS, and other paths, where intermediate ASes are traversed. To do this, we first gather the relevant hop information like ASN and whether the router is within an IXP and group them based on the AS-level path. We then use this data to look at the most common connection type occurring between source ISP and cloud provider based on the number of intermediary networks which are traversed and save these results for further use after removing paths that were observed too rarely. The specific cut-off point for the latter is an arbitrary number that we choose based on the country and the number of measurements and different source networks that were observed.

4.6.1. Case Study: Germany

The first country we focus on is Germany, which is located in Central Europe and offers a good number of probes and measurements. The types of connections are shown in Fig. 4.16 accompanied by the number of measurements that are present in our database for the specific ISP and cloud provider pair. In the left plot, the color shows how often this connection type occurs with darker color representing a higher percentage. We can see that traffic from all ISPs go through direct peering paths for Amazon, Google, and Microsoft. The only exception for this is between AS 16097 and Microsoft, which we believe to be an artifact stemming from our probe collection methodology since the measurements for this case all come from one single probe and the second hop observed in the traceroute already belongs to a different AS, namely the Deutsche Telekom. Most of these direct paths are also observed at a very high frequency which is why we believe these to be accurate. IBM also has direct peering interconnections with a lot of the ISPs in this country and as have already seen in the previous section, IBM seems to have more interconnections at public IXPs compared to all of the other cloud providers. Lastly, the other cloud providers have a very low amount of direct

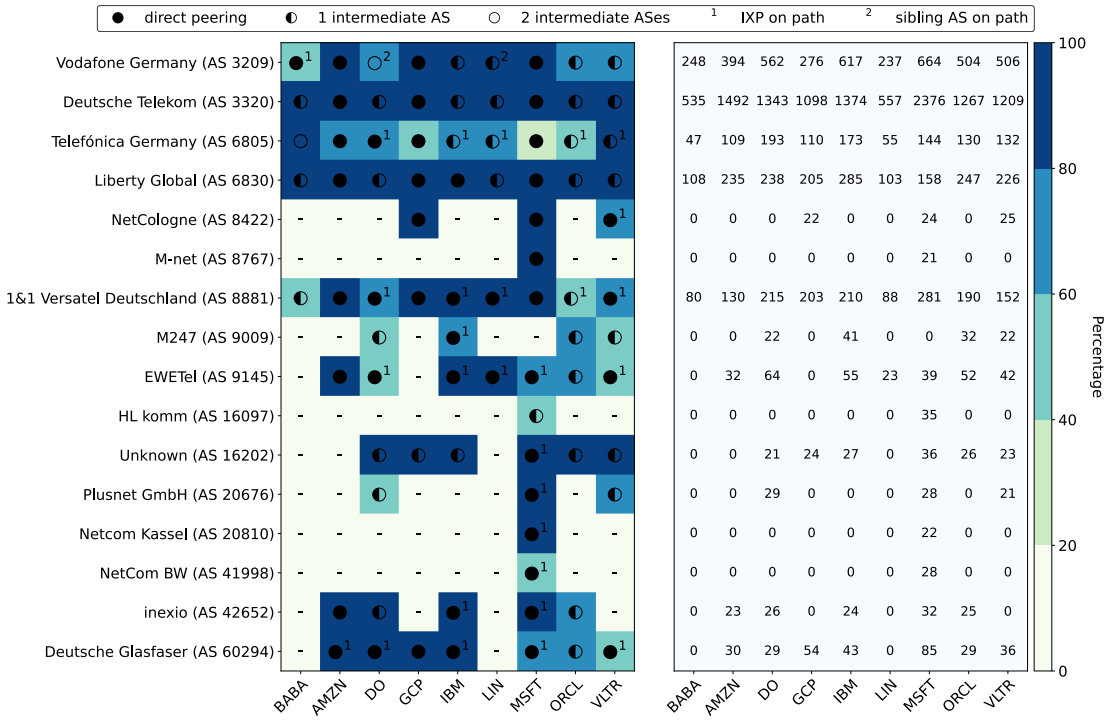


Figure 4.16.: Connection type between ISPs and cloud providers in Germany including number of observations

peering interconnections with the local ISPs, all of which are happening at an IXP. They instead mostly rely on Tier 1 ISPs to transport the traffic.

When looking at the pervasiveness differences in Fig. 4.17, the results are as expected,

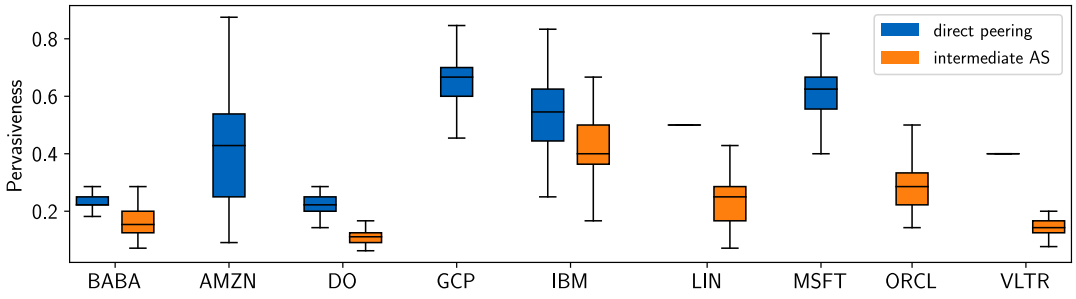


Figure 4.17.: Pervasiveness comparison between the two types of connections

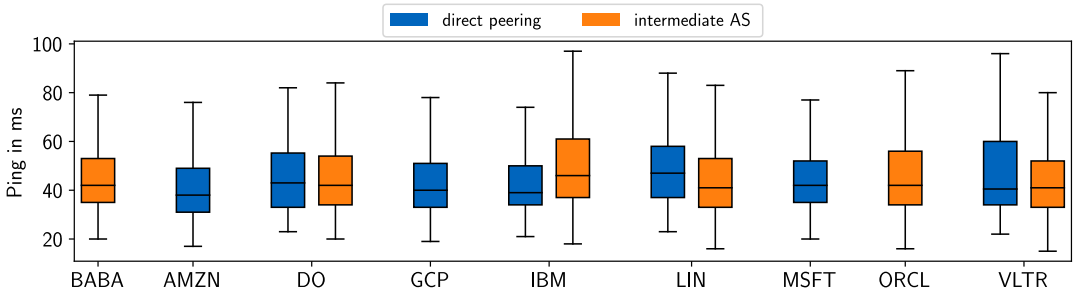


Figure 4.18.: Comparison of latency between direct and non-direct paths from Germany to Great Britain

with direct peering paths showing a higher percentage of traveled hops within the cloud provider network than for other non-direct paths. The percentage values and the relative order between the cloud providers are very similar to the one observed in Fig. 4.12 with a significant increase in the direct paths of the smaller cloud providers.

The latency comparison between the two groups is shown in Fig. 4.18, focusing only on measurements towards Great Britain, where every of our cloud providers operates at least one data center. We see that in this case, whether there is a direct path from ISP to cloud provider or not, there are no observable latency differences between them that are outside of a margin of error. This is most likely due to the path being relatively short and between two developed countries where the Internet infrastructure is reasonably well developed, making the geographical distance the largest factor in the latency.

4.6.2. Case Study: Japan

Next, we look at the Asian country where we have the highest amount of measurements. The connection types are shown in Fig. 4.19. For this figure we set the cut-off point a little bit higher than for the previous case to keep the figure legible. Again, we witness almost all of the connections between ISPs and Amazon, Google, and Microsoft are direct paths while IBM has many of their peerings at public IXPs. The other cloud providers rely almost exclusively on intermediate networks to deliver the traffic, showing a lower presence and degree of development here than in Germany.

The pervasiveness is very similar to the previous case, with an observable increase for the direct paths compared to other connections.

For the latency comparison in this case, we look at measurements towards India, the country with the second-highest number of distinct cloud providers following Japan. This is shown in Fig. 4.20, where we see, that the measurements obtained from direct

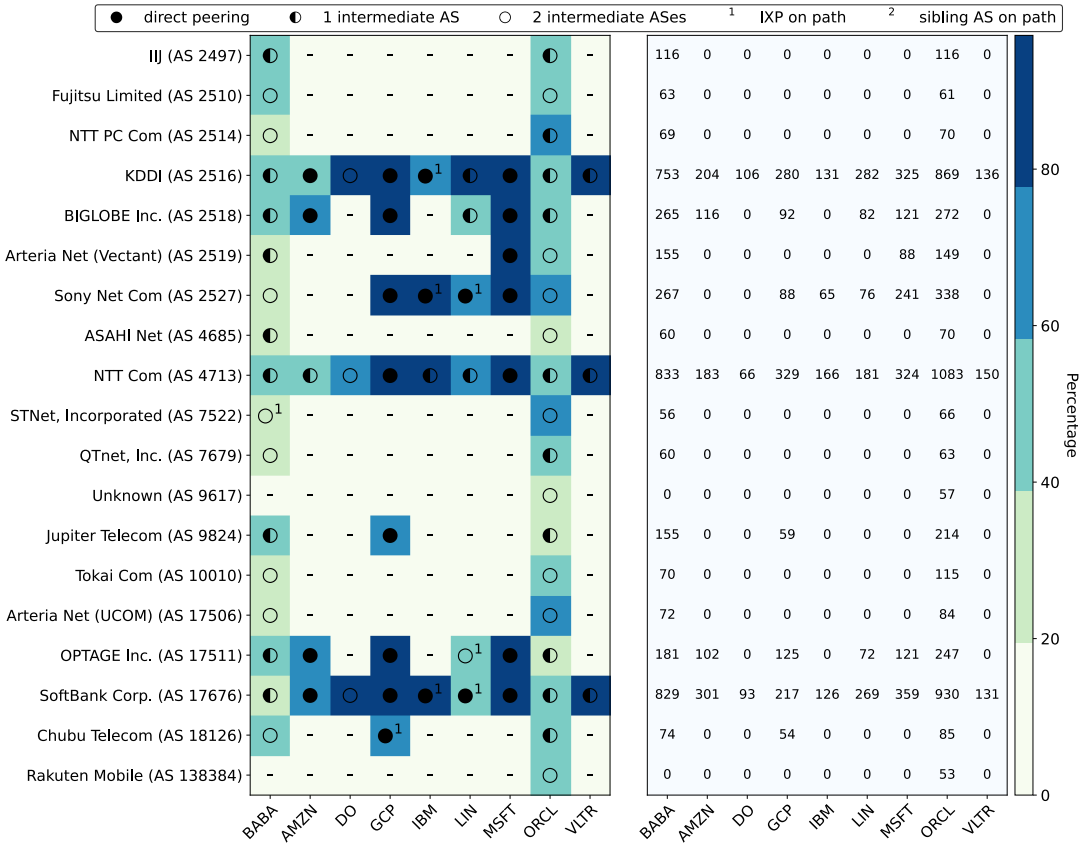


Figure 4.19.: Connection type between ISPs and cloud providers in Japan including number of observations

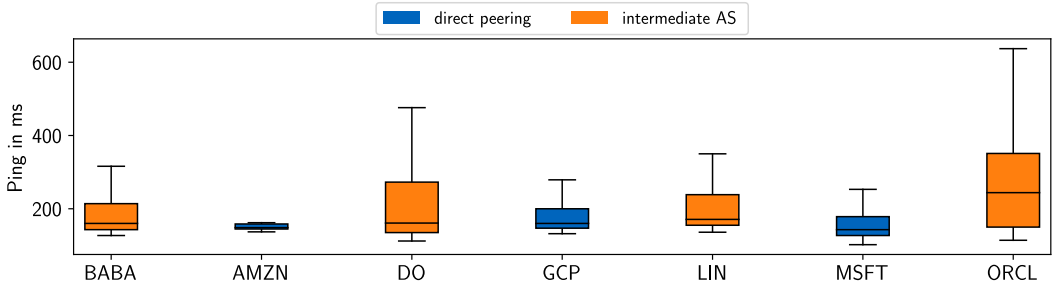


Figure 4.20.: Comparison of latency between direct and non-direct paths from Japan to India

peering connections, which only exist for Amazon, Google and Microsoft in this specific case, display a slightly lower, but more importantly, much more consistent latency.

From the results of these two case studies, we conclude that direct peering does not noticeably impact end-user latency in developed countries with good Internet infrastructure and shorter paths, but offer some significant improvements for paths traveling large geographical distances and most likely also connections in developing countries where the public Internet infrastructure is not as far developed.

5. Conclusion

This thesis aimed to offer insight into cloud reachability from a global perspective for end-users using wireless connections. Based on measurements from the Speedchecker platform and following analysis we saw differing levels of latency based on location and an array of factors influencing these results.

RQ1: What is the state of global cloud access latency for the end-user? We observed that probes in all countries of Europe, North America, and Oceania can reach a data center within PL on average, while results in other continents were more diverse with particularly high intracontinental latency for countries along the northern coast of Africa as well as West Asia where some countries exceed even the HRT boundaries. Users in countries with local data center presence or closer to data centers generally experience consistently lower latency when connecting to the closest data center. This means that for autonomous driving for example, an application where the latency is especially critical to guarantee safety, most of the developed countries fulfill the latency requirements, while wireless connections are still too slow for other more latency-constrained applications.

RQ2: What factors influence the quality of the connection and latency? The aforementioned findings lead us to the conclusion that geographical distance towards a particular destination is one of the most important factors, which is why cloud providers and other content providers continue to expand their data center deployment around the world. Last-mile latency plays another important role for the connectivity, but this is a section where the cloud providers do not have much influence because these sections are mostly controlled by regional providers. We also observed that the wireless connection introduces a non-negligible and unpredictable amount of additional latency in home networks as well as having an even larger impact on cellular connections where the last-mile is entirely made up by the wireless connection. We saw that Amazon, Google, and Microsoft, three large cloud providers who operate an extensive private network, own a majority of the path after the initial ISP, absorbing the traffic early on and close to the user and routing it through their backbone towards their data centers. We observed that this is achieved through direct peering with many different ISPs which has varying impact on actual latency based on the location and actual distance

that data has to travel. For shorter paths in regions where there is a good network infrastructure, there are no noticeable differences, while for long paths direct peering does show an improvement, mainly in the consistency of the latency.

5.1. Limitations

However, there are some limitations in our approach which may lead to slight inaccuracies in the results. Only saving the data for each unique probe ID once works under the assumption that probe IDs are never changing and always uniquely identify a probe. We also relied solely on traceroute measurements to gather information about traffic routes and peering agreements which introduces the possibility of errors in finding the correct path. Lastly, the way we ran our measurements is not very efficient in regards to the quota limit since one API call is used even if there are only a few probes available in the country which lowers the number of measurements gathered within a certain time frame.

5.2. Future Work

To get an even clearer understanding of global cloud reachability, future studies can extend these results through the use of other measurement platforms as well as combining traceroute measurements with BGP and other public data to confirm and enhance our findings in regards to Internet routes and relationships between networks. An investigation into differences between wired and wireless connections on a single platform is also worth considering to gain a deeper understanding of the implications of using a certain connection as well as closely investigating mobile connections specifically because many next-generation technologies like autonomous driving are reliant on this connectivity.

A. Appendix

A.1. Database Tables

Column name	Description
PingID	Unique, autoincrementing ID; primary key
ProbeID	Probe ID string provided by Speedchecker
Timestamp	Timestamp of the measurement in UNIX Epoch time format
DestinationIP	IP of the destination data center
DestinationURL	URL of the destination data center; foreign key on Datacenters table
Ping1–Ping5	Ping measurements in ms

Table A.1.: Ping table

Column name	Description
TracerouteID	Unique, autoincrementing ID; primary key
ProbeID	Probe ID string provided by Speedchecker
Timestamp	Timestamp of the measurement in UNIX Epoch time format
DestinationIP	IP of the destination data center
DestinationURL	URL of the destination data center; foreign key on Datacenters table

Table A.2.: Traceroute table

Column name	Description
HopID	Unique, autoincrementing ID; primary key
HopIP	IP address of the targeted hop
HopNumber	Hop number within the traceroute
RTT1–RTT3	Round-Trip-Time (RTT) of the target hop
TracerouteID	TracerouteID the hop belongs to; foreign key on TracerouteID in Traceroute table

Table A.3.: Hops table

Column name	Description
IP	IP address of the particular node; primary key
ASN	ASN inferred through CAIDA or pyasn; -1 if not found
OrgNamePDB	Organization name of the AS found in PeeringDB; "Unknown" if not found
OrgTypePDB	Accompanying organization type in PeeringDB; "Unknown" if not found
IsIXP	1 if IP address is within an IXP based on CAIDA data, 0 otherwise
IxID	ID of the IXP provided by CAIDA; -1 if node is not within an IXP
OrgNameCAIDA	Organization or IXP name provided by CAIDA; "Unknown" if not found
Continent Country City	Geographical data provided by CAIDA or alternatively geoiplookup; Default to "Unknown" for CAIDA and None for geoiplookup
Latitude Longitude	Latitude and Longitude provider by CAIDA/geoiplookup; None if not found
FacilityName	Name of the facility provided by CAIDA; "No facility" if no facility found, "Not IXP" if node is not within an IXP

Table A.4.: NodeInfo table

Column name	Description
DatacenterID	Unique, autoincrementing ID; primary key
URL	URL of the data center
Provider	Cloud provider the data center belongs to
Continent Country City Latitude Longitude	Geographical data inferred from geoiplookup with manual corrections afterwards

Table A.5.: Datacenters table

Column name	Description
ProbeID	Probe ID provided by Speedchecker; primary key
IP	IP provided for the probe by Speedchecker
ASN	ASN provided by Speedchecker
Continent Country City Latitude Longitude	Geographical data provided by Speedchecker; Default to None if not provided
ConnectionType	Connection type of the probe as provided by Speedchecker
Platform	Platform of the probe provided by Speedchecker (Android, PC or Router)
Network	Network/AS name provided by Speedchecker

Table A.6.: Probes table

Column name	Description
ID	Unique, autoincrementing ID; primary key
ProbeID	ProbeID from Speedchecker
Date	Date of the observation in *yyyy-mm-dd* format
TimeSegment	Number from zero to five to denote the time segment where the probe was observed (starting from zero at 00:30 AM UTC in 4 hour increments)

Table A.7.: ProbeActivity table

Column name	Description
RelationshipID	Unique, autoincrementing ID; primary key
AS1 AS2	ASN of the two ASes that have a relationship provided by CAIDA
Relationship	0 if peering relationship if AS1 is in a provider-to-customer relationship with AS2

Table A.8.: Relationship table

List of Figures

List of Tables

Bibliography

[1] Cisco. *Cisco Annual Internet Report (2018–2023) White Paper*. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html. Accessed: 2021-01-19.

[2] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. "Above the Clouds: A Berkeley View of Cloud Computing". In: *EECS Department, University of California,Berkeley, Tech. Rep. UCB/EECS-2009-28* 53 (2009), pp. 07–013.

[3] A. Li, X. Yang, S. Kandula, and M. Zhang. "CloudCmp: comparing public cloud providers". In: *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 2010, pp. 1–14.

[4] S. Ltd. *Performance Monitoring API*. https://probeapi.speedchecker.com/. Accessed: 2021-03-09.

[5] ThousandEyes. *ISP Tiers*. https://www.thousandeyes.com/learning/techtorials/isp-tiers. Accessed: 2021-02-03.

[6] ThousandEyes. *Cloud Performance Benchmark 2019–2020 Edition*. Tech. rep. ThousandEyes, 2019.

[7] G. Research. *Gartner Magic Quadrant for Cloud Infrastructure and Platform Services*. Tech. rep. Gartner, 2020.

[8] 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*. 2020, pp. 79–88.

[9] M. Eder. "Analyzing Cloud Reachability on Global Scale". Bachelor's Thesis. Technical University of Munich, 2020.

[10] R. NCC. *RIPE Atlas*. https://atlas.ripe.net/. Accessed: 2021-03-08.

[11] 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: Feb. 2021.

[12] Broadband Internet Technical Advisory Group Report (BITAG). *Interconnection and Traffic Exchange on the Internet(2014)*. Tech. rep.

[13] 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*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 230–246.

[14] A. Ahmed, Z. Shafiq, H. Bedi, and A. Khakpour. "Peering vs. transit: Performance comparison of peering and transit interconnections". In: *2017 IEEE 25th International Conference on Network Protocols (ICNP)*. 2017, pp. 1–10.

[15] T. Böttger, G. Antichi, E. Leao Fernandes, R. di Lallo, M. Bruyere, S. Uhlig, and I. Castro. "Shaping the Internet: 10 Years of IXP Growth". In: (Oct. 2018).

[16] R. Motamedi, B. Yeganeh, B. Chandrasekaran, R. Rejaie, B. M. Maggs, and W. Willinger. "On Mapping the Interconnections in Today's Internet". In: *IEEE/ACM Transactions on Networking* 27.5 (2019), pp. 2056–2070.

[17] B. T. Sloss. *Expanding our global infrastructure with new regions and subsea cables*. `https://blog.google/products/google-cloud/expanding-our-global-infrastructure-new-regions-and-subsea-cables/`. Accessed: 2021-03-24.

[18] Y.-C. Chiu, B. Schlinker, A. Radhakrishnan, E. Katz-Bassett, and R. Govindan. "Are We One Hop Away from a Better Internet?" In: Oct. 2015, pp. 523–529.

[19] J. Koetsier. *Report: Apple Is One Of Amazon's Biggest Customers, Spending Over $350 Million Per Year*. `https://www.forbes.com/sites/johnkoetsier/2019/04/22/report-apple-is-one-of-amazons-biggest-customers-spending-over-350m-per-year/?sh=21166e0f11c4`. Accessed: 2021-03-24.

[20] AWS. *AWS Direct Connect*. `https://aws.amazon.com/directconnect/`. Accessed: 2021-03-24.

[21] Google. *Cloud Interconnect overview*. `https://cloud.google.com/network-connectivity/docs/interconnect/concepts/overview`. Accessed: 2021-03-24.

[22] Microsoft. *Azure ExpressRoute*. `https://azure.microsoft.com/en-us/services/expressroute/`. Accessed: 2021-03-24.

[23] B. Yeganeh, R. Durairajan, R. Rejaie, and W. Willinger. "How Cloud Traffic Goes Hiding: A Study of Amazon's Peering Fabric". In: *Proceedings of the Internet Measurement Conference*. IMC '19. Amsterdam, Netherlands, 2019, pp. 202–216.

[24] V. Giotsas, S. Zhou, M. Luckie, and k. claffy k. "Inferring Multilateral Peering". In: *ACM SIGCOMM Conference on emerging Networking EXperiments and Technologies (CoNEXT)*. Dec. 2013, pp. 247–258.

[25] M. Luckie, B. Huffaker, k. claffy k., A. Dhamdhere, and V. Giotsas. "AS Relationships, Customer Cones, and Validation". In: *ACM Internet Measurement Conference (IMC)*. Oct. 2013, pp. 243–256.

[26] CAIDA. *AS Relationships*. `https://www.caida.org/data/as-relationships/`. Accessed: 2021-03-23.

[27] R. Oliveira, D. Pei, W. Willinger, B. Zhang, and L. Zhang. "The (In)Completeness of the Observed Internet AS-level Structure". In: *IEEE/ACM Transactions on Networking* 18.1 (2010), pp. 109–122.

[28] S. Sundaresan, N. Feamster, R. Teixeira, and N. Magharei. "Community Contribution Award – Measuring and Mitigating Web Performance Bottlenecks in Broadband Access Networks". In: *Proceedings of the 2013 Conference on Internet Measurement Conference*. IMC '13. 2013, pp. 213–226.

[29] V. Bajpai, S. J. Eravuchira, and J. Schönwälder. "Dissecting Last-Mile Latency Characteristics". In: *SIGCOMM Comput. Commun. Rev.* 47.5 (Oct. 2017), pp. 25–34.

[30] SamKnows. *SamKnows*. `https://www.samknows.com/`. Accessed: 2021-05-03.

[31] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch. "Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture". In: *IEEE Communications Magazine* 55.2 (2017), pp. 70–78.

[32] R. NCC. *RIPE Atlas Probes*. `https://atlas.ripe.net/about/probes/`. Accessed: 2021-04-03.

[33] R. NCC. *About Measurements*. `https://atlas.ripe.net/about/measurements/`. Accessed: 2021-05-07.

[34] M-Lab. *Measurement Lab*. `https://www.measurementlab.net/`. Accessed: 2021-05-07.

[35] C. Dovrolis, K. Gummadi, A. Kuzmanovic, and S. D. Meinrath. "Measurement Lab: Overview and an Invitation to the Research Community". In: *SIGCOMM Comput. Commun. Rev.* 40.3 (June 2010), pp. 53–56.

[36] M-Lab. *M-Lab Data Overview*. `https://www.measurementlab.net/data/`. Accessed: 2021-05-07.

[37] CloudHarmony. *CloudHarmony*. `https://cloudharmony.com/`. Accessed: 2021-03-10.

[38] S. Ltd. *Probe API Documentation*. `https://www.speedcheckercdn.com/probe-api/documentation.html`. Accessed: 2021-03-10.

[39] J. S. Labs. *Country and Continent Codes List*. https://datahub.io/JohnSnowLabs/country-and-continent-codes-list. Accessed: 2021-03-23.

[40] Wikipedia. *Private network*. https://en.wikipedia.org/wiki/Private_network. Accessed: 2021-03-23.

[41] CAIDA. *CAIDA IXP Dataset*. https://www.caida.org/data/ixps/. Accessed: 2021-03-23.

[42] H. Asghari and A. Noroozian. *pyasn*. https://pypi.org/project/pyasn/. Accessed: 2021-03-23.

[43] PeeringDB. *PeeringDB*. https://www.peeringdb.com/. Accessed: 2021-03-23.

[44] geoiplookup.net. *GeoIP Lookup XML API*. http://geoiplookup.net/xml-api/. Accessed: 2021-03-23.

[45] AWS. *AWS IP address ranges*. https://docs.aws.amazon.com/general/latest/gr/aws-ip-ranges.html. Accessed: 2021-04-16.

[46] S.-C. Lin, Y. Zhang, C.-H. Hsu, M. Skach, M. E. Haque, L. Tang, and J. Mars. "The Architectural Implications of Autonomous Driving: Constraints and Acceleration". In: *SIGPLAN Not.* 53.2 (Mar. 2018), pp. 751–766.

[47] R. NCC. *Probe Density Map*. https://atlas.ripe.net/results/maps/density/. Accessed: 2021-04-03.

[48] W. Li, D. Zhang, G. Xie, and J. Yang. "TCP and ICMP in Network Measurement: An Experimental Evaluation". In: *Parallel and Distributed Processing and Applications*. 2005, pp. 870–881.