Observing IPv6 Scanning at Public Cloud Platforms

Leeuwen, Maarten van

Observing IPv6 Scanning at Public Cloud Platforms

by

M.J.H. van Leeuwen

Supervisor:	Prof. Dr. G. (Georgios) Smaragdakis
2nd Supervisor:	Dr. A.I. (Alex) Stefanov
Project Duration:	September 2023 - January 2024
Faculty TU Delft:	Elektrotechniek, Wiskunde en Informatica, Delft
Faculty Leiden University:	Governance and Global Affairs, Leiden
Student Number:	<redacted> (LU)

Style: TU Delft Report Style, template by Daan Zwaneveld

Preface

Before you lies the master thesis "Observing IPv6 scanning at public cloud platforms". With this thesis, I will conclude my executive master Cybersecurity at Leiden University. I have had the opportunity to do research at the Delft University of Technology. This thesis combines multiple topics I have been interested in for the last 10 years, IPv6, Cybersecurity, and Threat Intelligence.

I want to thank my supervisor, Prof. Dr. G. Smaragdakis, who gave me the chance to do this study at TU Delft. Also, thanks to the people in my reviewpool for the time they spent to improve this thesis.

M.J.H. van Leeuwen
Delft, March 2024

Summary

This thesis is written as part of a master's degree in cybersecurity at the Faculty of Governance and Global Affairs (FGGA) at Leiden University in cooperation with the Cybersecurity Department of the Faculty of Electrical Engineering, Mathematics and Computer Science of Delft University of Technology. It describes the results of a study into the extent and effectiveness of IPv6 scanning on the internet.

IPv6 scanning is becoming more active as the adoption of IPv6 on the Internet grows. The discovery of active IPv6 addresses is a complicating factor in effectively scanning the Internet for active services. Actors therefore use different methods of finding active systems. In this thesis, a comprehensive view is given of the methodology used by these scanners and to what extent IPv6 scanning is happening. Other studies have researched this problem using their own infrastructure. This study focuses on IPv6 scanning behavior within public cloud infrastructure.

A custom-designed and developed data collection platform is used comprised of 39 collection probes, that collect data of scanning activity on the internet. Data is sent to a data processing environment that enriches the data to gather information about the actors that are scanning the probes. The probes are spread across two different cloud platforms that both have a unique method of assigning an IPv6 subnet to a probe.

The results show that the method of assigning addresses that cloud providers use can make a significant difference in the discovery of new addresses by scanners. At one cloud provider addresses are discovered within an average of 2 hours, compared to the other providers of which some probes were never discovered. While previous studies have reported similar behavior that matches what is observed in the second provider, the ease of discovery that is shown at the first cloud provider has never been observed in previous research.

Three experiments ran during the study in which probes, that were still undiscovered, simulated the behavior of regular systems on the internet. The results show that some scanners run legitimate services on the Internet which are used by various kinds of devices. Scanners can use this to collect active IPv6 addresses and use this information to scan these networks. This behavior is observed during this study and has identified three scanners that use this method of discovering active IPv6 addresses.

Additional results of this thesis show that IPv6 scanning is becoming more active and that it is mostly performed by commercial parties. Both the results of the extent and effectiveness matched or even exceeded the results of previous research. Showing an increase in scanning activity by 55% compared to a study in 2022.

Future research is proposed based on the results of this study. These include: (1) creating a methodology for discovering legitimate services that are used to collect active IPv6 addresses; and (2) a comparison of the effectiveness of scanners in different environments, like home networks, business networks, and cloud infrastructure.

Contents

List of Figures

List of Tables

1

Introduction

This thesis is written as part of a master's degree in cybersecurity at the Faculty of Governance and Global Affairs (FGGA) at Leiden University in cooperation with the Cybersecurity Department of the Faculty of Electrical Engineering, Mathematics and Computer Science of Delft University of Technology. It describes the results of a study into the effectiveness of the discovery of new public IPv6 addresses and networks by external parties within public cloud environments. The research is designed together with the TU Delft and is part of a longer research project into potential malicious Internet activity.

IP version 6 (IPv6) is the successor to IP version 4 (IPv4). It is used to identify computers on the internet, and allow communication between internet-connected networks. This is because the Internet is made up of different networks which are connected to each other. These networks are run by businesses, individuals, universities, governments, and more. Through these networks, data and information is exchanged. Computers within these networks communicate using IP addresses. IPv4 is the primary protocol used for addressing the internet. It has been used since the beginning of the internet.

Although everyone uses version 4 there is a practical disadvantage to this version: the number of addresses is limited. In 1995, it was thought that IPv4 addresses would eventually run out since IPv4 has 4,294,967,296 addresses available. As a solution, IPv6 was developed. IPv6 has so many addresses, 3.4×10^{38}, that there are practically an infinite number of addresses available. IPv6 also provides more privacy features compared to IPv4, since because of the size the addresses can be randomized and changed frequently.

1.1. Motivation

Discovery of active public IPv4 addresses by scanners on the Internet has always been 'business-as-usual'. A full scan of all available IPv4 addresses can now be done within hours or even minutes [1]. With IPv6 this isn't possible at this time, due to the large public IPv6 address space available [2].

A lot of information is available on the scanning of IPv4 addresses. Companies have set up services that detect this scanning and report this to organizations that want to protect against this scanning. Examples of these are Greynoise [3] and Spamhause [4]. Other companies are running these scans, this is done to, for example, provide information to organizations on their attack surface. Shodan [5] and Censys [6] are two of the most common organizations that do this.

Information on the scanning of IPv6 is lesser known. There has been a limited number of studies done on the topic, compared to the scanning of IPv4 addresses. Commercially, two providers of scanning data have relatively recently started scanning on IPv6 (Shodan and Censys). Greynoise does not support IPv6 and recognizes that IPv6 knowledge of users needs to improve [53]. Spamhause does have a strategy for blocking IPv6 spam traffic but recognizes that there is more experience to be gained to better understand how this can be done [7].

So, in summary, both scanners and security companies have either only recently started using IPv6 or indicated themselves that this is still too much in their infancy to start doing this, and there has been

limited research on it. This combination ensures that new research can help raise the level of knowledge about IPv6 security. Additionally, an argument is often made that because it is harder for scanners to discover active IPv6 addresses on the internet, it is more secure. The underlying definition for this is 'security-through-obscurity'. Meaning, making a system more secure by obscuring the attack surface of such a system. For IPv4 this doesn't work anymore, but does this still work for IPv6?

This combination of factors, the limited knowledge of IPv6 scanning (both for the attackers and defenders), and the argument of security-through-obscurity created the motivation for this research. To allow for practical information on these issues, it will help to examine the scanning in the wild. Thus, this raised the question: how long can a fresh IPv6 address, which hasn't been active or used before, stay 'hidden' on the Internet to be discovered by scanners? In other words, to what extent and how effective are public IPv6 scans.

1.2. Terminology

Anyone can link a computer to a network that's linked to the entire internet. Once connected, this computer can join other networks or offer services, like hosting a website. The prerequisite for this is that a computer has an IPv4 and/or IPv6 address so that it can be accessed by other computers from other networks. Because the Internet is fundamentally an open network, all computers can access each other, desired and undesired. Thus, there are parties who, for various reasons, try to access all networks to see if services are offered there. By doing this on a large scale, it is possible to "scan" the Internet to identify all available services on the internet. For IPv4, this is relatively easy because the number of addresses is limited. For example, with modern technology, it is possible to scan all IPv4 addresses in less than an hour. Due to this, IPv4 scanning on the Internet is widespread. A recent study showed that 2.17 million internet-wide scans are performed every month [8]. For IPv6, it is not possible to scan all IPv6 addresses due to the large number of addresses available. As a workaround, there are several options for "scanners" to try to "find" IPv6 addresses that are in use. The capabilities used affect the effectiveness of these scanners.

In this thesis, many abbreviations and terminology are used. A table is created with a description for each of these definitions. This table can be found in the appendix of this thesis. This is Table A.1

Section 2.1 further describes IPv6 and its use on the internet. It provides a more technical overview of the information provided in this section.

1.3. Research Problem

For Internet scanners, like Shodan and Censys, it is more difficult to discover new public IPv6 addresses compared to IPv4. It isn't feasible to scan for all public IPv6 addresses since the range for this is too big. IPv6 scanners use different methods for the discovery of public addresses. Scans aren't only done by Shodan or Censys. It is done by many different parties that have different motivations. It can be malicious, for commercial purposes, or research. While studies show the volume of the scanning that is done, and show how public IPv6 addresses potentially can be discovered, the extent and effectiveness of scanning has only been studied to a limited extent.

Also, no research has yet been done on the addressing strategies of public cloud providers. These service providers offer online services that allow companies or individuals to bring services online. Different ways are used to assign IPv6 networks to users. The approach of assigning addresses can potentially impact the exposure of IPv6 networks. This may create privacy or security risks. By making it easier for systems to be found, or by making it easier to profile users.

1.3.1. Main question

The main question for this thesis is: *What is the extent and effectiveness of current public IPv6 scanners in public cloud environments?*

The effectiveness of a scanner is primarily measured by the timeframe in which a new IPv6 address is found. The extent will indicate the different techniques used by scanners to discover IPv6 addresses and how widespread these scans are. Combined with potential patterns seen in the scanning activity, a comprehensive view is created of the scanning activity of IPv6 on the internet.

1.3.2. Sub-questions

The main question is divided into sub-questions that break down the main research question.

1. In what timeframe does a scanner find a fresh public IPv6 address?
2. Which factors influence the timeframe in which a fresh IPv6 address is found?
3. Which scan patterns can be recognized in IPv6 networks running at cloud providers?
4. What protocols reveal the identity of an IPv6 address the fastest?
5. How to mitigate easy discovery?

1.3.3. Expected additional outcome

Apart from the research question, other results can be expected. These results primarily have to do with the setup of a data collection and processing environment for similar studies. Therefore, this thesis takes an in-depth look at the design, implementation, and experiences of setting up such a research environment.

The following questions can be expected to be answered throughout the thesis:

- What requirements can be identified when designing a research environment for Internet activity?
- How to design a data collection probe?
- What data needs to be stored and for what purpose?
- How to ensure the integrity of the data and verify the accuracy of the data?
- What process can be used for the analysis of network data?

1.4. Research contribution

This research provides greater insight into the techniques scanners use to discover new IPv6 addresses and identifies the factors that influence how quickly new networks are found. Previous research has done similar studies using private infrastructure. In this study insight is given into the scanning that is observed at cloud platforms and how addressing strategies from these platforms impact how fast services are found. Additionally, it updates the information on how much scanning in IPv6 takes place on the Internet of which the latest study of this was done in 2022.

With this insight, defenders are better able to protect against reconnaissance activities. It can help protect users' privacy by providing insight into how an address' exposure evolves. Also, more and more services are adopting IPv6, and it is coming into critical infrastructure as well. For this group, it is important to know how high the threat is from Internet scanners.

Because this research also focuses on the addressing strategies of cloud providers, it can also help develop best practices for providers for assigning an address or network to services. Especially as more businesses and individuals begin to use cloud services, the role of these providers will grow. With good, industry-wide, best practices, users of cloud services can be better educated on the use of IPv6 which can benefit security.

Lastly, for this study, a data collection and data processing environment is set up. This thesis describes how this system is set up and what limitations have to be taken into account. The environment runs on public cloud platforms that can be considered an untrusted network. The collection process has to be built so that data integrity is guaranteed. The results of this can be used to support future research into the topic.

1.5. Overview

This thesis consists of three parts: related work, approach, and results. It concludes with a discussion of the results and the conclusion.

Chapter 2 describes the theoretical framework in which this thesis is written. It also provides a more technical description of how IPv6 works, its adoption rate, address assignment techniques, and an explanation of how current Internet scanners work.

Chapter 3 describes how the research was set up and how the research environment is designed and built. It finishes with the limitations experienced during the study which can help improve further research into this topic. This chapter answers the questions that are given in subsection 1.3.3.

Chapter 4 will provide the results of the technical part of this study. It contains results on the observed scanning activity, the profile of the actors that were seen scanning the collection infrastructure, and a link between the discovery of new addresses by scanners and the behavior of the probes. It starts with an overview of the applied settings of the probes.

Chapter 5, the discussion, describes the results based on the research questions and compares them to previous work that is done in the field. It discusses the limitations of the research, presents the implications of the results in a broader technical and organizational perspective, and based on this defines recommendations. It ends with a reflection on the thesis.

Finally, in the conclusion, chapter 6, the answer to the main question is given and suggestions are made for future research.

2

Related work

This chapter presents a baseline level of knowledge that is to be used to interpret the research in this thesis and summarizes the latest and 'state-of-the-art' research into the topic op IPv6, IPv6 security and IPv6 scanning. It is done by first describing IPv6 and some relevant key principles related to this (2.1). Rotation algorithms are used for privacy purposes and thus help keep an address stay hidden on the internet, due to this it can be seen as a protection mechanism against the effectiveness of IPv6 scanners. Section 2.2 describes these algorithms. This is followed by describing previous research into the topic of IPv6 scanning. Previous research can be categorized into two topics: (1): research on discovery techniques (2.3); and (2): research on scanner activity (2.4).

2.1. IPv6

IPv6 is the successor to IPv4. It has been a standard since 1995 [9]. The IPv6 standard has been created because it was expected that the public IPv4 address space would 'run out' in the long term. In 2011 this became true when the Internet Corporation for Assigned Names and Numbers (ICANN), worldwide responsible for assigning IP addresses, reported that all free public IPv4 addresses were handed out [10]. Between 2011 and 2019 the Regional Internet Registries (RIRs), coordination of assigning IP addresses in specific regions (delegated by ICANN), each reported the same [11][12][13][14][15]. With all address space handed out, new and existing organizations will have to resort to either acquiring IPv4 address space from other sources (like from IPv4 address brokers), or by using methods to more efficiently use IPv4 addresses (like using Network Address Translation).

The issue of depleting the entire address space in IPv4 is not a concern with IPv6. This is because there is significantly more IPv6 address space available. The address space is many magnitudes bigger compared to IPv4. While IPv4 has 4,294,967,296 addresses available, IPv6 has 3.4×10^{38} addresses available [9]. There hasn't been a realistic situation thought of that indicates that the IPv6 space is not sufficient for the foreseeable future of the internet.

2.1.1. IPv6 Address Assignment

An IPv4 address consists of 32 bits [16] and an IPv6 address consists of 128 bits [9]. For example, an IPv4 address looks like: 123.123.123.123, and an IPv6 address looks like: 2001:0db8:1234:abcd:5678:ef01:234a:5678. IPv6 consists, like IPv4, of network bits, which identifies the network of which a computer is a part of, and host bits which uniquely identifies the computer on the network. While with IPv4 the network bits generally are 24 bits of the 32 bits of the address, in IPv6 this is usually split in half, so 64 of the 128 bits are used for the network bits and the remaining 64 bits are used as host bits, also called an Interface ID (IID).

RFC7421 describes certain privacy and security issues when using host bits that are smaller than 64 bits. Especially when moving to a bit size of less than 80, any security advantages that the IPv6 standard would create are undone [17].

The usual solution for IPv4 address assignment is a DHCP server which has a pool of addresses that

can be assigned to clients on a network. With the IPv6 standard multiple other standards were created to facilitate address assignment. This consists of retrieving the network bits and retrieving the host bits. The most used solution is Stateless Address Auto-configuration (SLAAC), which provides the network address, subnet, and gateway. It additionally points to a DHCPv6 server to get more information, like the DNS server [18].

DHCPv6 can also be used for retrieving the host bits [19]. Other standards exist for the same functionality. For example, using a modified version of EUI-64 a client generates its own host address based upon its MAC address[20]. Other solutions also exist and are implemented by different operating systems. They often use an address generation algorithm to generate the host bits, like RFC4941 [21] or RFC7217 [18], which are implemented in macOS, iPadOS [22], and NetworkManager which is used within Linux operating systems [23].

2.1.2. Adoption

Reporting on the adoption of IPv6 can be done in two categories. One being the actual adoption rate, and secondly the reasons why organizations adopt IPv6. This subsection describes the research that is done relating to these two categories.

Adoption Rate

Despite being a standard for more than 25 years, IPv6 is still not completely adopted across the world. Various studies have been done in recent years that have measured the adoption of IPv6 across the internet. Adoption rates differ a lot depending on the metric that is used. Metrics that are seen being used in studies are: (1) connection statistics of large websites, (2) traffic analysis in an Internet exchange, (3) analysis on BGP routing information, and (4) various statistics using the Domain Name System (DNS).

A study from Google in 2010 showed an adoption rate of 0.35% [24]. Another study in 2017 also used the data from Google and showed an adoption rate of 16%. It also looked at the traffic volume at the Amsterdam Internet Exchange (AMS-IX), which, at the time, was the largest Internet exchange in the world. Of all the traffic going through the Internet exchange, 1.5% was IPv6 traffic [25]. This was compared to another study in 2012 which, from the same vantage point, observed 0.6% of all traffic through the AMS-IX being IPv6.

A report from RIPE in 2022 shows a significant increase in IPv6 adoption over 10 years. This report used BGP, website data, and traffic analysis of the AMS-IX as a metric [26]. The BGP table saw a growth of 1600% in size, compared to 150% of IPv4. Client adoption, which is measured using website data, in this case using Google and Akamai, showed that at the end of 2022 an adoption rate of 40%, 30% respectively. Lastly, the data from the AMS-IX showed that at the end of 2022 the adoption rate is 4.1%. This shows an increase compared to the study in 2017. None of the studies provide an explanation why the adoption rates differ between the different metrics.

Organizational Factors Surrounding IPv6 Adoption

Research is done into the reasons why organizations want to adopt IPv6, or why they don't. A study in 2018 identified three key factors in why organizations implement IPv6. This has to do with three key factors that are relevant for adopting IPv6: (1) new features, (2) relative advantage, and (3) complexity [27]. Other research shows a more in-depth analysis of these factors.

The IPv6 standard does provide some additional features. These are mostly related to the increased size of address space. It is built to be more efficient in larger networks, and due to its larger address space provides users with more options to protect their privacy on their devices. It also has more capabilities for securing network traffic between clients or networks. [28]. Verizon Wireless and T-Mobile have implemented IPv6 on their networks but use translation mechanisms to allow IPv6 clients to connect to IPv4 networks [29]. It is unknown why these organizations have chosen to migrate to IPv6.

Multiple studies have analyzed if there are other benefits when migrating to IPv6. An advantage that may be seen is a decrease in latency and therefore an increase in speed with IPv6 compared to IPv4. Online it is speculated that this may be the case due to the lack of Network Address Translation (NAT) [30] [31]. Research has shown that this is not the case. In the study from Google the latency is seen to be comparable between IPv6 and IPv4 [24]. Another study of 2014 showed similar results [32]. Also,

it found that the use of transition technologies, that allow IPv6 and IPv4 to work together, increase the latency significantly. These technologies can be seen in use by Verizon and T-Mobile, which was reported on in the study by Nikkhah et al. in 2016 [29].

These factors, combined with the experiences from organizations (like GreyNoise) that recognize a gap in knowledge about IPv6, show that a complete adoption to IPv6 on the Internet may not be as straightforward. However, the benefits are there, and everyone who communicates on the Internet will need to move towards a fully adopted IPv6 standard. If it only were to continue the growth of the internet.

2.1.3. Internet Scanning

Scanning the Internet is common practice. It is done very actively by both attackers during the first phase of an attack [33], by organizations that map active services on the Internet in order to use this information commercially [5] [6], or for research purposes [8]. In IPv4 this is very effective since scanning the entire IPv4 space usually takes up to an hour [1]. This is due to the limited size of the IPv4 address space. For IPv6 this isn't feasible due to the large size of available address space. [34]

Some Internet scanners, like Shodan and Censys, try to discover public IPv6 addresses. A quick search on Shodan shows 180,746,295 results (3rd of August 2023, query: *has_ipv6:true*) [35]. Censys, which started scanning later, shows 175,195,573 results (21st of November 2023, query: *labels=ipv6*) [36]. This shows that a significant amount of addresses are scanned and reported on, even compared to IPv4, where Censys has 239,672,028 addresses in its database (query *not labels=ipv6*). While these are a lot of IPv6 addresses, this will never be a complete list, because IPv6 addresses change easily and that the discovery process is not flawless. Section 2.3 further elaborates on the different techniques used by scanners to discover fresh IPv6 addresses.

2.2. Rotation Algorithms

Address generation standards like EUI-64 create privacy problems since it is possible to track devices across different networks [37]. This is done because the host bits are static per device. Many operating systems therefore use alternative address generation algorithms and additionally work with primary and temporary addressing. This means that a client on the network may have multiple IPv6 addresses which are used for different purposes. The host bits, or IID, are often changed in order to reduce the risk of being profiled across networks and on the Internet [38]. The network bits of household networks may also be periodically rotated by Internet Service Providers (ISPs) to further reduce the risk of being tracked on the Internet [39].

Different algorithms are created in order to enhance the privacy of users in networks using IPv6. These algorithms are called rotation algorithms. These algorithms work both for the host bits and for the network bits. An algorithm for this is defined in RFC3972, which describes a method of integrating cryptography into the host bits of an IP address [40]. Other possible solutions to rotate the addresses and to protect the privacy of users are defined in RFC7721 [41] and RFC8135 [42]. Additionally, standard RFC8981, and in an earlier version RFC4941 [21], describes an extension to SLAAC which causes hosts to generate temporary addresses and change them over time. This standard describes this as a solution for, among other things, prevention of scanners finding the active host [38].

IPv6 scanners will therefore be less effective because addresses aren't as static and may only be active for 24 hours. This may be a solution to reduce the effectiveness of IPv6 scanners. While this may be true, many servers are also accessible via a DNS name that is translated to an IPv6 address [43]. When rotating an address for a server it also has to be changed at the DNS record. It therefore is debatable if this is a prevention method for reducing the effectiveness of IPv6 scanners for servers on the internet.

2.3. New IPv6 Address Discovery Techniques

For Internet scanners, like Shodan and Censys, it is more difficult to discover new public IPv6 addresses compared to IPv4. It isn't feasible to scan for all public IPv6 addresses since the range for this is too big. Public IPv6 space is 2000::/3 meaning that there are $4.25 \text{x} 10^{37}$ possibilities [44]. IPv6 scanners therefore use different methods for the discovery of public addresses. This section has the different discovery techniques categorized into: (1) Information leakage via protocols; (2) Target generation algorithms; (3)

Well known host addresses. Lastly an effective solution may be a combination and therefore is also given.

2.3.1. Information leakage via protocols

Many protocols on the Internet leak information about systems connected to the internet. Other protocols can be leveraged in order to collect information about systems. Basically three types of information leakage can be recognized: (1) the scanner runs a service on the Internet to which clients connect to and therefore expose their address; (2) scanners actively look within public datasets which contain public services, like DNS; (3) clients and scanners both connect to legitimate services in which the addresses are exposed. This subsection presents examples for all three of these methods.

This is also possible for the discovery of new IPv6 addresses. For example, Shodan has used NTP servers to collect IPv6 addresses from devices that are connected to those servers. This is not a flawless technique because a client needs to connect to a NTP server for it to be discovered. If a client does not use the NTP server, a client, or the entire network, won't be discovered by the IPv6 scanner operating the NTP server. But if a client does connect to it, it provides information on an active system, and thus an active network. This showed to be very effective in finding Internet of Things (IoT) devices, since these devices often use public NTP servers and have static IPv6 addresses [45].

Another protocol or system that can be leveraged is DNS. DNS maps a name to an IP address. For example, www.google.com maps to the IPv6 address 2a00:1450:400e:803::2004 (21 November 2023). When collecting DNS names and periodically resolving these names to IP addresses, fresh IPs can be found. This technique is effective in finding servers that are online and providing an Internet service [43].

Lastly the peer-to-peer file sharing protocol BitTorrent can be used to collect IPv6 addresses. The BitTorrent protocol requires clients to connect to the BitTorrent network. The IP addresses connected to this network are public and can be collected. This method is very effective in finding personal devices of users, since these systems are mostly used to download or upload files [46]. While this may seem effective, modern devices often run rotation algorithms for the host address for privacy reasons. This however does not protect against detecting an active network.

2.3.2. Target Generation Algorithms

While it is effective to discover new active addresses using information leakage, it does not allow for proactive discovery of new addresses. It is also limited in effectiveness due to it being reliant on the behavior of the victim itself. Thus, more active scanning approaches have also been presented. These are called Target Generation Algorithms (TGAs). Using known active addresses it creates a list of potential active hosts based upon the already known addresses. Examples of these algorithms are Entropy/IP [47], 6Gen [48], 6Gan [49], or 6Tree [50].

Using hitlists generated from, for example information leakage, it is possible to guess new addresses. This may be effective when the algorithm is able to figure out what pattern is used in address assignment or what seed is used in a rotation algorithm. A recent study evaluated 10 different Target Generation Algorithms (TGAs) and found that they are between 1 and 34% effective. It also found that the algorithms have a tendency to favor ISP networks for target generation, which in turn is not very effective on a long term basis since these addresses rotate frequently [34].

2.3.3. Well Known Host Addresses

Lastly, another solution looks into the BGP routes that are on the Internet and uses well known host addresses to make a guess for the full IPv6 address [51]. This reduces the address space to scan drastically since the maximum subnet size that is allowed in public routing is /48, giving 2^{16} possible network addresses. Looking at a report of APNIC it shows that in January 2023 47 percent of all prefixes in the public BGP routing table are /48 [52]. This creates some opportunities for Internet scanners to discover public IPv6 addresses.

In practice this works as follows: the IP range of 2001:db8:ad76::/48 is advertised. This aggregates 65,536 (2^{16}) networks, since the first 64 bits are generally used as network bits, 48 bits are known in the route and the remaining 16 are variable. A potential hitlist can than be generated for any of those

networks. Using existing IPv6 hitlists a set of popular host bits, or IIDs, can be identified. Usually these are lower-byte IIDs. With this set of popular host bits a full address can be created from the 65k network addresses that were initially generated from the /48 BGP route. An example of a full address is then: 2001:db8:ad76:4513::1.

2.3.4. Combination

Often a combination is used in order to discover new or fresh IPv6 addresses. Input can be created for the TGAs by using known active IPv6 addresses from information leakage. Additionally, potential active hosts can be identified by using information from the Internet routing table. This creates a comprehensive list that can be used to discover new IPv6 addresses [43].

2.4. Scanning Activity

Some research on IPv6 scanning activity has been done before. While research dates back to 2006, where no IPv6 scanning activity was detected, and in following studies marginal scanning was seen. The first large scale IPv6 scanning was seen in a study in 2018. In this study scanning activity was measured using DNS Backscatter [2].

More recently two studies reported on the extent and effectiveness of IPv6 scanning. In 2022 Richter et al. reported on large scale IPv6 scanning activity. The research showed that IPv6 scanning is very active. Showing 5,199 scans over a 15-month period originating from 1,326 sources, when aggregated to a /64 prefix. While this is very limited compared to an IPv4 scanning activity, where in 2019 about 2.17 Million scans were detected within a single month [8], it shows that IPv6 scanning is happening and increasing. Besides the scanning activity the paper also reported on how large-scale IPv6 scanning could be detected. While in IPv4 it is usually aggregated by a single IP address, in IPv6 this is not effective. This is because the researchers found that some scanners use multiple scan sources instead of a single IPv6 address. This is relevant for this thesis since it is important to factor in the aggregation level of source IPs to identify a single scan source [53].

The second study in August 2023 by Tanveer et al. ran multiple experiments to measure the effectiveness of IPv6 scanners. It looked at 6 different methods where the IPv6 address of a device was exposed due to information leakage in different protocols. Each experiment took 2 weeks to complete and a baseline as a reference was set over the course of three months. The results show that a significant increase in scanning activity can be seen after networks addresses were exposed. The biggest effects were seen when probes directly connected with scanners using DNS requests or webcrawling. Indirect contact like running an NTP server within the NTP pool or running a TOR node do also show to result in increased scanning activity. This paper closely resembles the original idea for the research in this thesis. The results of this study will therefore be compared against the results of this thesis [54].

3

Methodology

Multiple components had to be set up for this research. Basically consisting of data collection and data processing systems. This chapter describes the approach for the research, consisting of the system and data requirements, planning, configuration and key components. This is elaborated in order to help future research in setting up similar systems.

While other research has focused on the effectiveness of IPv6 scanners in fully controlled environments, where the addresses were fresh and the environment new. In this research the focus will be on public cloud environments, where address assignment is not controlled by the user and therefore additional factors effect how quickly new systems are found.

The approach in a high-level overview was as follows:

- Create the design goals of the platform.
- Design and develop a data collection probe and data processing infrastructure.
- Deploy data collection probes in multiple public cloud environments.
- Collect data from the probes and store on a central location for data analysis.

This chapter first describes the initial system requirements that were used in the designing process. It follows with a section on the research planning. In the section *Probe Design* a detailed technical description is given on how the probe was set up. Following this, a broader technical description is given on the data collection system and the data processing environment. Lastly, limitations were found during the implementation and production of the systems, the final section describes these.

3.1. System and Data Requirements

Requirements were defined based upon previous research, the research proposal, and conversations with the instructor of this thesis. Together a set of 22 requirements were set. These are categorized into: (1) External hosting; (2) Development; (3) Data collection and retention; and (4) Behavior analytics. This section describes these and the argumentation behind them.

3.1.1. External Hosting

The core idea of this research meant that the system had to be able to collect scanning data in untrusted and previously unused/unfound networks. There were known limitations from the start because the data collection part is done in public cloud environments in which the network assignment process is not controlled. Two requirements were then defined that relate to these limitations. Firstly, any probe that is deployed must be checked if it was assigned a network that is suitable for this research. It was expected that this process would require a trial-and-error approach, since the network that is assigned to a probe is not known in advance. Therefore, the deployment of the data collection probe must be automated in order to find suitable networks (req. 1). Secondly, a connection to the data processing

system needs to be secure and individually managed per probe (req. 2), since the probe itself is deployed in an untrusted environment.

The research requires multiple cloud platforms in order to compare different addressing strategies. The selection process of which cloud platform to include in the research is therefore based on how addresses are assigned to the probes that are deployed. It is required that multiple addressing strategies are included in the research and therefore the selection process is to be based on that (req. 3). Additionally, the platform needs to support automation features in order to comply with requirement 1. It also needs to support dual-stack networking with a public IPv4 address in order to communicate to the Internet without exposing the IPv6 address (req. 4). The IPv6 address must not be derived from the IPv4 address, which may expose the IPv6 address anyway (req. 5).

3.1.2. Development

A development environment is required since the data collection environment, which consists of multiple probes, is developed specifically for this research. The integrity and reliability of the data that is collected is the primary reason why a robust development process is needed (req. 6). This is done with a phased approach using DTAP [1] phases. Where the development environment is used to test new software and experiment with it, the testing system is used to test the deployment of the software and the manageability of the features. Lastly the acceptance environment runs a copy of what is run on production and is therefore used to verify the functionality of new features in a production setting. An acceptance system runs on every public cloud platform that is used.

Part of the research is to test if the behavior of probes affect scanning practices. Therefore, requirements are set for the probe.

- The response to TCP scanning on inactive sockets needs to be controllable (req. 7)
- The probe is silent by default and does not connect to any server using IPv6 (req. 8)

3.1.3. Data Collection and Retention

The data collected needs to be stored indefinitely by making it transferrable to some type of long term storage (req. 9). This to support potential reviews on this study or to allow for future research using this data. Additionally, data needs to be collected in two forms, in order to be able to verify the reliability of the data collection (req. 10). These requirements are done to ultimately ensure the integrity of the data and to make sure the data that is analyzed is verified for accuracy.

Requirements are set on the format of the data. All data is extracted from network traffic. The network data needs to be logged by a network analyzer (req. 11) and separately needs to be captured and stored on a central location (req. 12). This ensures that logging data can be verified by capture data. While logging data is easier to process for further analysis. To interpret the capture data a database is kept containing all relevant metadata of the probes that are run in testing, acceptance and production (req. 13).

Multiple types of data are collected in different formats during the research. These formats need to be normalized in order to support automated analysis (req. 14). Some processing and initial analyses has to be done when data is collected in order to make data analysis easier and to retain accuracy of the data. The following analysis needs to be done during data collection:

- Offline IP address profiling to autonomous system number (ASN) (req. 15)
- Interpretation on the state and direction of traffic flows (req. 16)
- Any connection is precisely timed with an accuracy of at least a second in order to correlate data flows between different log formats (req. 17)

3.1.4. Behavior Analytics

Any connection that is set up by the probe is controlled (req. 18). This is in order to prevent any potential contamination and unwanted exposure of the IPv6 address or network. Before any behavior analysis can be performed it is important to verify if the address(es) are not known by other scanners. A

[1] Development, Test, Acceptance, and Production

Date	Reviewers	Document status
11-dec	Georgios	1st draft: Introduction, Related work, Approach
22-dec	Georgios	2nd draft: Thesis complete
30-dec	Georgios, Alex	Final draft
6-jan	Georgios, Alex, LU	Thesis deadline

Table 3.1: Deadlines for the thesis

Month	Activities
August	Preparatory work Development work
September	Theoretical study Selecting cloud providers Development work
Oktober	Probes active Data analysis
November	Probes active Data analysis Thesis writing
December	Thesis writing

Table 3.2: Research planning during the thesis period

verification process must be created that can give a relative degree of certainty on the previous exposure level of an address or network (req. 19). Any research that is performed can only start after this process has finished, and thus that it is determined that a network has not been exposed before.

After this process the research can start and different behavioral factors can be tested. In order to support this, any IPv6 traffic initiated by the host must be controlled (req. 20). Different experiments are planned for this study. Therefore, active behavior needs to be generated in a controllable manner, thus the destination, protocol and targets have to be defined before (req. 21). This needs to be done for the following activities: (1) NTP requests; (2) Webcrawling; (3) DNS AAAA queries (req. 22). Combined this creates a probe that allows for testing of different behavioral factors that may expose the address or network to scanners.

3.2. Research Planning

This research is done for the purpose of writing a masters' thesis and therefore it is bound to the time limitations that are set forth by Leiden University. In this case the planned start of the thesis is September 2023 and the deadline for submission is the 6th of January 2024. Additionally, it may support the writing of a research paper, which will be created after the thesis has been submitted.

A research planning has been created based on the timelines set forth by Leiden University. Additionally, together with the supervisors multiple deadlines were set for the review of drafts of the thesis. Tables 3.1 and 3.2 show the planning of the thesis.

3.3. Probe Design

Data has to be collected to research the effectiveness and the extent of IPv6 addresses in practice. The approach will be to install measurement software in different datacenters of public cloud providers that receive an IPv6 address and network. This software will then collect network data that is destined for the IP addresses and the network. With this data it becomes possible to know the timeframe in which the IPv6 address was found and by whom. Together with other data, like the ASN, outgoing traffic (such as NTP requests), or different type of responses on TCP requests a more comprehensive view is created on the use and effectiveness of IPv6 scans on the internet.

3.3.1. Software Tooling

The software will consist of a collection of open source tooling that can be run on the chosen cloud platforms. The probes used an up-to-date version of Debian 11[2] server on which the software is configured and installed during deployment. Any outgoing IPv6 traffic is immediately blocked using IPtables at the beginning of the deployment, this is to prevent accidental exposure of the IPv6 address. Additionally, software that is retrieved from an external location is downloaded over IPv4 for the same reason.

[2]Version (using *lsb_releas a:*): *Description: Debian GNULinux 11 (bullseye)*

In summary the following software and tools will be used during the data collection phase:

- **IPTables**[3]: to control incoming and outgoing network traffic, block unwanted network traffic, and log network traffic for analysis purposes. (version: 1.8.7)
- **Custom software**: custom software is created to collect network data, manage the probe, and generate behavior for research purposes.
- **Log collector**: to collect all logging data and sent that to the data processing server.
- **Wireguard** : used to create a secure tunnel to the data collection environment and to allow for remote management.
- **Daemonlogger**: in order to create PCAPs of all traffic to meet requirement 8. (version 1.2.1)

For a more technical description and information on the custom software see Appendix A.2.

3.3.2. Secure Connection

After deployment a secure tunnel is set up to the data processing infrastructure. This also is done over IPv4 to prevent exposure. This tunnel is used for (1) sending data to the processing environment, and (2) managing the probes. Using the database server other metadata can be added to the ingested data to later help the data analysis part. This is further described in section 3.4. Figure 3.1 shows a representation of the network infrastructure that is set up for this research.

3.3.3. Public Cloud Selection

Critical in the selection of the public cloud providers are the addressing strategies. Ideally three addressing strategies would be used. These were:

1. A probe is assigned a dedicated /64 address space.
2. A probe is assigned dedicated address space but significantly smaller than /64.
3. A probe is assigned a single IPv6 address.

A provider is found for strategy 1 and 2. No public cloud provider has been found that has an addressing strategy where the probe is limited to a single IPv6 address. For addressing strategy 1 the provider Vultr [55] is selected, for strategy 2 DigitalOcean [56] is selected. Where Vultr is assigning an entire /64 to a probe and DigitalOcean assigns a /124, which gives 16 unique addresses, compared to the 2^{64}-1 (1 for the gateway) addresses by Vultr.

3.4. Data Collection

Two primary data streams exist when the probes run. One is the network logging data, and the second is the raw network traffic into a PCAP format. Both data streams need to be stored and processed. This section describes this process and explains what additional external data sources have been used.

3.4.1. Deployment of Probes

Probes are deployed in the two public cloud platforms. Every cloud platform has different regions and datacenters. Probes are deployed in every region of every cloud platform. A completely silent host (even on TCP requests) with IPv6 is used until it is verified that the network hasn't been detected. This is done for a week, while in the meantime the network is checked at public resources for exposure. Resources like hitlists, public scanners (like Shodan and Censys), if an address in the network is used in a DNS record, and lastly if it has a PTR record. If all test fail and no scanning activity has been seen on the probe, or in the entire network space, the network is considered not known by any IPv6 scanner. It therefore is then suitable for use within this research.

The deployment of hosts is done by using a standard Debian 11 image from the cloud provider. Initially a single acceptance probe is deployed using the deployment script. After testing if all functionality and settings are applied correctly multiple probes are deployed over all regions of a provider. Registered and verified accounts were required with both cloud providers to perform these actions.

[3]IPTables allows filtering of IP packets in the Linux kernel.

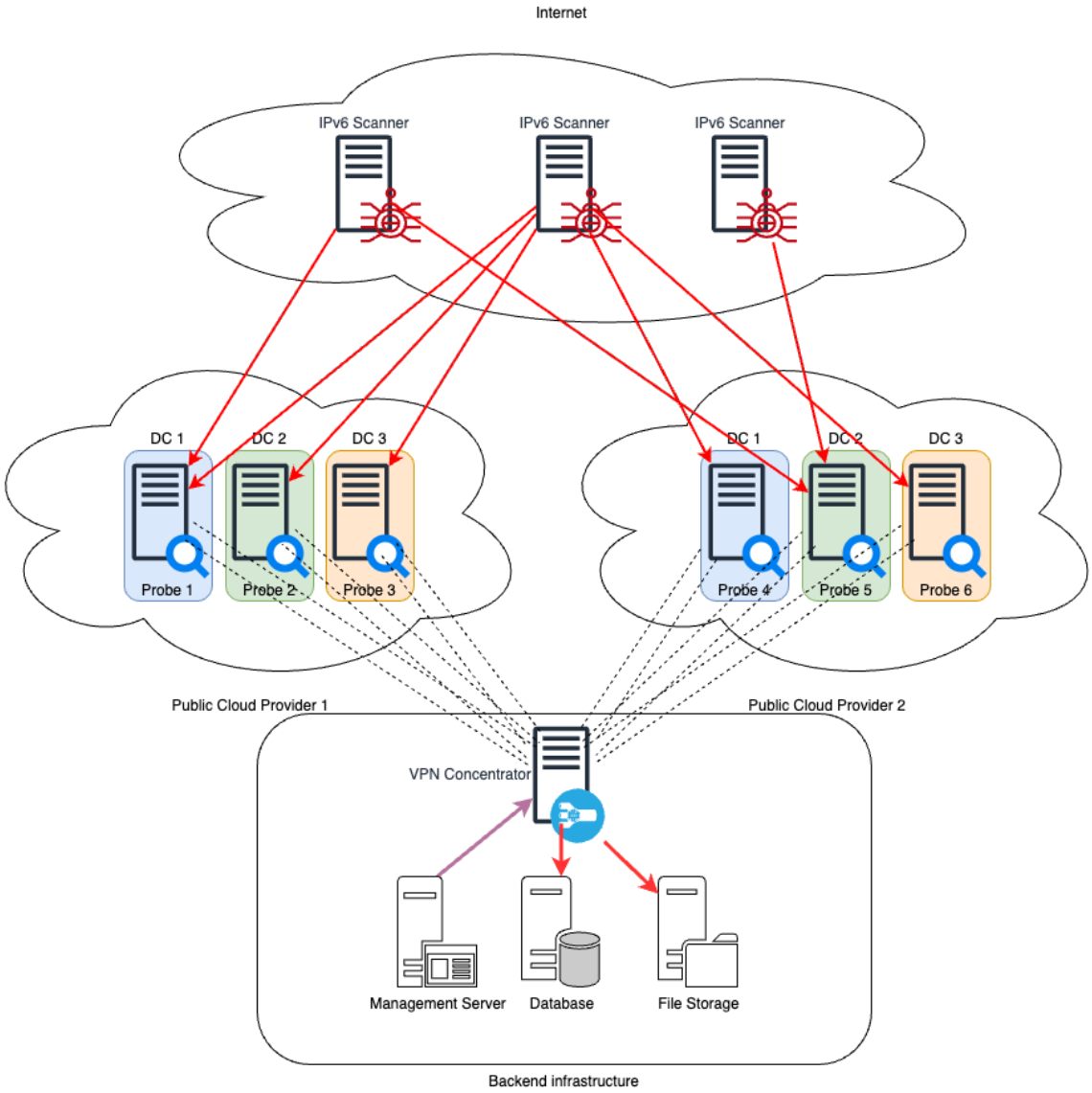

Figure 3.1: Diagram showing the architecture of the probes connecting with a VPN Concentrator through which management, file and database traffic is routed.

After deployment a VPN connection is set up using unique public and private keys to a central VPN concentrator. This system allows the probes to securely connect to management, database and file servers. If any probe is removed or compromised the connection to all internal systems can be removed by revoking its key to access the infrastructure.

3.4.2. Database

All normalized data is stored in an Elasticsearch (version: 8.10.2) [57] database. By using this platform it is possible to store and process large amounts of data with predefined structures. It also has existing data management and processing tools for the log types that were used during this research. This made the initial setup of the platform faster. Additionally, the REST API functionality helped with the integration of the data analysis process.

All logging data is stored in a structured log dataset which is predefined by standard tooling for reading and interpreting IPTables log files. All data is set to be stored indefinitely and is backed up using both system backups and the backup functionality of Elasticsearch.

An agent (Elastic Agent version: 8.10.2 [58]) is installed and configured on every probe. This agent is centrally managed using a Fleet server (Kibana version 8.10.2 [59]), which is run within the management platform of the Elasticsearch platform. Standard policies were pushed to all the probes from the Fleet server [58]. When a new probe is installed it automatically received its policy and started sending logs to the Elasticsearch database.

3.4.3. Collection of Network Logging

Using IPTables the metadata of all in -and outgoing network packets are logged. This is the primary source for data analysis in this research. Some key data needs to be included into the logs. One is the direction of the traffic: if the traffic is originating from the probe or if it is sent to the probe. To achieve this, every log entry contains a prefix based upon the IPTables chain in which a packet is going through. INPUT is for incoming packets, and OUTPUT is for outgoing packets. The FORWARD chain is ignored, since the use of this chain is not relevant for this study.

Secondly, the state of a connection is key. Using the connection tracking system within IPTables a prefix is added to a log for any NEW or ESTABLISHED connections. This helps to identify potential scanner traffic. By filtering on the *incoming* and *new* prefixes it is possible to find any packet that did not originate from the probe and wasn't a response on a connection that was initially set up by the probe itself. This makes the data analysis much easier, since querying the database for certain types of traffic flows becomes very simple.

3.4.4. Packet Captures

All probes collect packet captures of the IPv6 network traffic that is received on the Internet facing interface. This is done by using the software tool Daemonlogger and a custom script that sends capture data to a central collection server for processing. Daemonlogger runs as a service on the probes with a BPF filter of *ip6*. This ensures that solely IPv6 data is collected. All PCAPs are rotated every 15 minutes and are gzipped and periodically sent to the central file server. The file server then combines all PCAPs of a single probe in a PCAP of a single day. In further research it is then possible to use this PCAP data over certain periods of time.

To allow further research to interpret the data, a dataset is extracted from Elasticsearch that provides metadata about every probe. Metadata contains IP addresses, system information and configuration, and simulated behavior during this study. Additionally, if this research needs to be reviewed all raw data is available for verification purposes. Lastly, the PCAP data is used to verify if all network data that is sent to the host is actually logged to Elasticsearch. This way it is possible to check the reliability of log streams.

3.5. Data Processing

Multiple phases of data processing happened during this research. During data ingesting some data processing happens in order to normalize the data stream and to enrich with external data. After ingesting the data is analyzed for two purposes: verification if a host is suitable for research, and

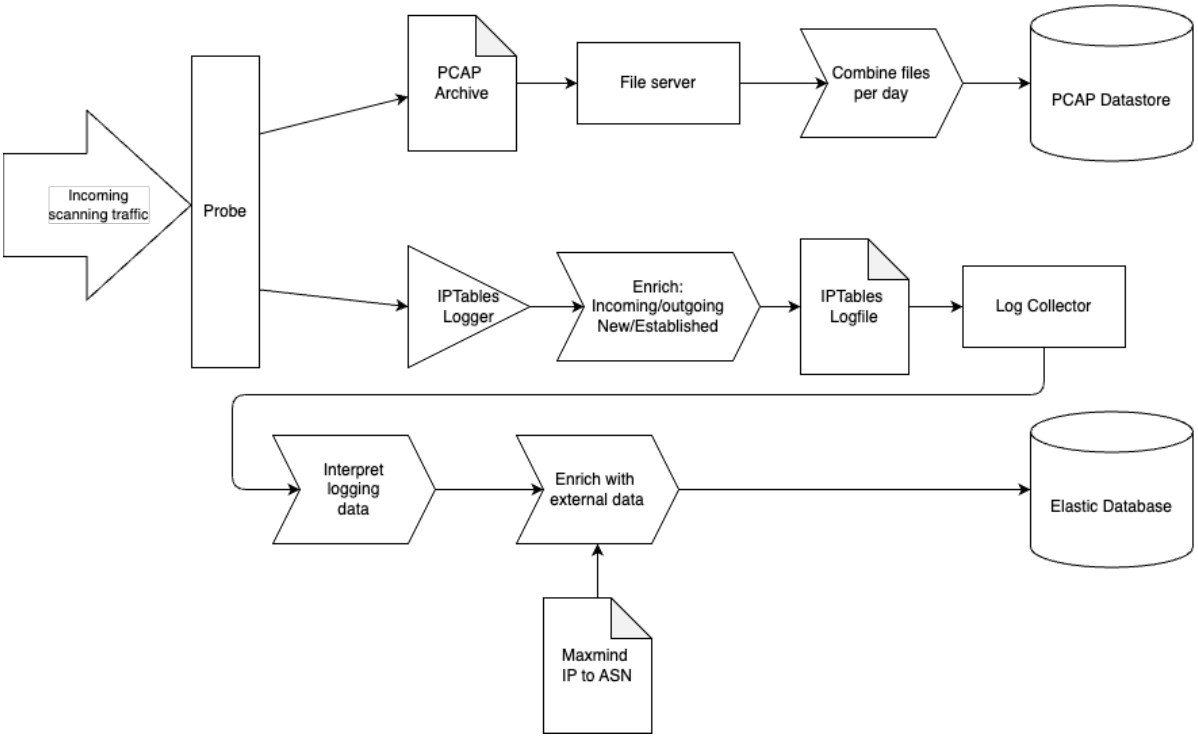

Figure 3.2: Data processing pipeline of the approach.

secondly for analysis of the actual study. Figure 3.2 shows the data processing pipeline until the data is stored. In the subsection 3.5.1 all tooling is described that is used in the processing phase.

During the preparatory phase some factors that may influence the effectivity of IPv6 scanners were predefined. Subsection 3.5.2 describes this. Lastly, subsection 3.5.3 describes the profiling techniques that were used. This is to find the purpose of the scanner.

3.5.1. Tooling

Different tools have been used for data analysis. Primarily Kibana (version: 8.10.2) and Python (version: 3.9.2) code were used. Kibana is used for discovery of anomalous behavior and monitoring of probes. Dashboards have been created for this purpose. Figure 3.3 shows a screenshot of such a dashboard.

Python was used as the main tool for producing the results during the data analysis phase. A GitHub repository [4] is published containing the code that is created specifically for this research.

3.5.2. Predefined Factors

Multiple variables were defined that were used to come to an answer for the sub-questions, which are defined in subsection 1.3.2. This research focuses on the effectiveness and extent of IPv6 scanning in the wild. The effectiveness of a scanner is primarily measured by the timeframe in which a new IPv6 address is found, and what factors may change the timeframe in which an address is discovered. The extent is presented using various different indicators that relate to previous research. The research that this is based on is the paper from Richter et al. [53] and Tanveer et al. [54]. This subsection describes the indicators that are used during the data analysis phase.

Firstly, a comparison should be made between the cloud providers. Relevant factors are therefore (1): is it possible to get a network assigned that isn't found by scanners? (2): in what timeframe do scanners find a network of a cloud provider without any behavior from the probe?

Secondly, after determining if it was even possible to get a fresh network assigned, other factors were taken into account. (1) The timeframe in which a probe is scanned after certain behavior has taken

[4] https://github.com/mvlnetdev/thesis2324_code

Figure 3.3: Screenshot of a dashboard that was used to monitor the scanning activity on probes. This shows data of a week from a probe deployed at DigitalOcean.

place, (2) what client behavior of the probe influences the effectiveness of a scanner, (3) if responses on scanning traffic (like TCP RST on a closed socket) change scanning activities by scanners.

Lastly, for any situation it is important to know what type of scanners were seen. (1) This is done by profiling the scanners (see subsection 3.5.3). (2) It also is interesting to see if scanners were seen that scanned other IP addresses in the network than the used IP address that showed behavior. (3) Finally, a breakdown is made into the protocols, ports, source IP addresses, and prefixes used by the scanners.

3.5.3. Profiling

It was possible to profile all the scanners to a certain degree. This is done using different techniques. During data ingesting the IP addresses of scan sources were enriched using MaxMind data (see section 3.5) [60]. With this data the source AS, organization and country could be found. For every IP address that was used during scanning a reverse DNS lookup is done to find a link to a domain which may help identify a service. Further profiling using WHOIS data can help identify the purpose of the scanning activity or describe the type of business that is linked to the scanners' IP address. Not every step was possible for every IP address. Section 4.4 describes the results of the profiling steps.

Additional to the analysis of individual IP addresses, further analysis can be done by aggregating IP addresses into a subnet block. This was also done by Richter et al. [53]. By identifying how many network blocks per ASN were used for scanning purposes, and how many addresses per block helps profile the scanning approach of the scanner.

3.6. Limitations

Some changes to this research had to be made to adapt to the available resources and to work within the time constraints. In the proposal factors were identified that could affect the discovery of new IPv6 addresses by scanners. Due to time constraints this has been limited to two factors: NTP and Web. Also, it would have been ideal to cooperate directly with the public cloud providers to ensure that the IPv6 network that were assigned were previously unused. It was seen as probable that the cloud providers would not participate in this research and therefore the risk was poorly manageable. Due to this, it posed too great a risk to meeting the thesis deadline and therefore was not taken into consideration as a viable option for this study.

During the discovery process in finding unused network blocks it turned out to be fairly difficult to find one. Almost all networks were either scanned just after deployment, or were easily guessed by scanners. This posed a problem that this may not deliver sufficient data to produce relevant results. Additionally, the accounts on one cloud provider, which ran the probes, was suspended due to the activity of repeated and fast deployment and removal of probes. This violated the terms of service of the providers. Therefore, the accounts had to be verified, and deployment processes had to be changed. This delayed the research process and reduced the time to gather data. Therefore, the research period was extended for two additional weeks to ensure that enough data was gathered. In the end this helped but still didn't produce the amount of data that was initially expected for this study.

4

Results

This chapter describes the results of the research that is done for this thesis. This is written in five sections. The first section describes the settings that were configured during this research and for what goal these were set. Secondly statistics are given on the scope of the research platform and what data is observed. In section 4.3 a more in depth view is given on the observed scanning activity. Section 4.4 describes the analysis on what scanners were observed. At last, the section Sensitivity Analysis (4.5) presents possible correlation between probe behavior and scanning activity.

4.1. Setting

The settings on the probes can be categorized into controlling the silent mode, or active mode behavior. With silent mode, the settings define how a probe would respond on scanning activity, and the settings related to the active behavior caused the probe to actively connect to external services on the internet. Two different cloud providers were used, DigitalOcean and Vultr. Both cloud platforms have datacenters in different continents, countries, and regions within countries. Probes were deployed in every datacenter of every cloud platform.

The network and addresses were assigned using the standard process of the cloud platforms. After deployment the address was registered in the data processing environment. While the cloud platforms do allow users to change the address, this wasn't done as an experiment during this study. On one occasion the address was changed, because the scanning activity that was seen on a probe was anomalous compared to the expected behavior. This didn't change the behavior and is used in this study as an observation.

4.1.1. Silent Mode

The probes were initially set to be silent for 7 days. This was done by blocking all outgoing IPv6 traffic using IPTables. In the settings SSH access on IPv6 was disabled. The traffic to port 22 is also blocked as an additional measure to prevent this from interfering in the research. The silent mode continued for at least 7 days until verification of possible previous exposure was finished. This is to ensure that the assigned network and IPv6 address was not exposed by previous users on the cloud platforms.

After exposure, depended on the test case, probes could be changed to respond on communication from scanners. This was done by allowing specific protocols and settings using the filters available in IPTables. Specifically the following two settings exists:

- Allowing ICMPv6 echo replies (type 129) [1]
- Allowing TCP RST ACK as a response on closed sockets [2]

[1] ICMPv6 is used for different functions, type 128 and type 129 are used to test if a device is active. A host sends an echo request (type 128) to another host, which replies with an echo reply (type 129). [61]

[2] TCP is used for establishing a connection between two devices. This protocol uses sockets to identify the application for which the connection is meant. Not all sockets have to be in use. If a client tries to communicate with a server on a socket that is

Reference	Description
A	Assigned public IPv6 scope to Digital Ocean. In this case it is part of a /48 assigned scope. This scope is used for multiple probes. 9 unique /64 prefixes across the 21 probes were assigned to the probes.
B	Always zero.
C	Randomized 24 bits.
D	User customizable 4 bits, but always starts at 1.

Table 4.1: Describing a breakdown of the recognized pattern in a DigitalOcean assigned address. See figure 4.1

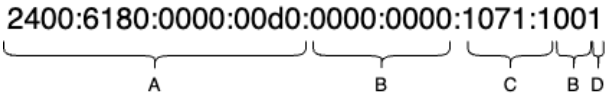

Figure 4.1: Addressing strategy breakdown of a DigitalOcean address that was assigned to a probe.

4.1.2. Active Mode

During the research it showed to be more difficult to control the active components of the probe. All images had automated update installation programs installed (unattended-upgrade) and NTP software running. These components were disabled in order to prevent possible leaks of the network address. The accept environment was used to design a solution for every cloud providers' image that disabled any outgoing traffic that was initiated by the probe itself.

Custom software is written that allowed for controllable active behavior. The software had two functions: (1) an NTP scanner that connected to a specified list of NTP servers, and (2) a web scraper that connected to 300 individual hosts starting from a single website. The NTP scanner connects to static IP addresses, and the webscraper uses DNS to resolve the hostnames of the websites it has to connect to. The default DNS resolvers from the cloud providers are used.

4.2. Baseline Results

Two public cloud platforms are in scope of this research: (1) DigitalOcean, and (2) Vultr. These platforms have a different approach to assigning IPv6 addresses to customer resources. DigitalOcean assigns a /124 network, and Vultr assigns a /64 network. In total 21 probes were run on DigitalOcean and 18 ran on Vultr, all had unique IPv6 networks (either /124 or /64 respectively) assigned to it.

4.2.1. Addressing Strategies

When a probe is deployed on a cloud platform an address is assigned to it by the provider. This process differed between DigitalOcean and Vultr. Vultr generates a random MAC address for the host and uses the EUI-64 standard to generate the IID of the probe, giving 2^{56} possibilities. This follows RFC 7421 [17]. Based on the data that is collected, or accessed, during this research it can be stated that any IPv6 address that was assigned by Vultr was ever assigned to a host before. With this information it can be said that an IPv6 address that is automatically assigned to a probe at Vultr, using this type of address assignment strategy, will always be unique.

With DigitalOcean a different addressing strategy is used. A clear pattern can be recognized within the assigned addresses. Figure 4.1 shows a breakdown of an IPv6 address that was initially assigned to a probe between the 9th of November 2023 and 13th of November 2023. The breakdown contains certain patterns that were recognized across the assigned addresses of all probes. Table 4.1 describes these.

closed the server will respond with a TCP packet that has the flags RST and ACK set. [62]

Selection step	Digitalocean	Vultr
1	21	7
2	0	3
3	0	0
4	0	0
5	0	0
Remaining hosts	0	8

Table 4.2: Results of the probe selection during silent mode, processed from top to bottom. If one step failed, other steps weren't tested.

Cloud provider	Unique AS
Digitalocean	69
Vultr	5

Table 4.3: Unique count of connected autonomous system numbers per provider

4.2.2. Selection of Usable Probes

As discussed in subsection 3.4 any probe that had to be deployed needed to be checked if it could be useful for the research project. This meant checking the network and address against five criteria. These are stated below. After matching one rule, the probe would be automatically discarded and further analysis on other rules would not be done.

1. Is scanning activity seen in the network or address within a week of deployment?
2. Does the network or address exist in the hitlist?
3. Have public scanners (Shodan or Censys) already identified the address?
4. Is the address linked to an existing DNS record?
5. Does a PTR exist for the address?

Usable Probes

Of the 39 probes that were deployed, only 8 probes were marked as usable. DigitalOcean had no usable probes and were all scanned within 4 hours of deployment. Of the 18 probes on Vultr only 8 remained after a week of deployment. A direct correlation was seen when the network was on a hitlist and if scanning was observed. No scanning was ever detected on the address specific to a Vultr probe, but only on other addresses in the assigned network. This scanning activity was detected by monitoring the Neighbor Solicitation messages originating from the router. The full results of this analysis can be seen in table 4.2.

Deployment of Probes

Within a span of 54 days (20[3] of October 2023 until 12[4] of December 2023), 39 probes were deployed. Figure 4.2 shows a graph of the deployment of the probes. Initially the deployment started using DigitalOcean. This was done to start the collection process while, in the meantime, other providers were selected and initial data could be verified. When Vultr was selected and an account was created, the maximum number of possible probes for the account were deployed. This was done because the experiences using DigitalOcean showed that it may be difficult to find 'fresh' IP space. By using a larger amount of probes it was possible to increase the chances of finding unused or undiscovered IP space.

After three days the account at Vultr was suspended and a lengthy process of reinstating the account followed. After this process was finished some probes were destroyed, because the maximum number of probes on the account was reduced by the provider. Meanwhile, all data analysis on the probes of DigitalOcean was finished, and it was concluded that no probe could be used for additional research. Even when changing the default IP addresses to another address within the assigned ranges.

After probes were selected for further research, it turned out that there were insufficient probes that could be used as a reference probe. These probes allowed the data that is gathered from active probes to be compared to probes that weren't active. Therefor, additional probes were deployed in order gather more reference data. This was possible because the maximum number of probes was increased by Vultr after a request from the researchers.

[3] th
[4] th

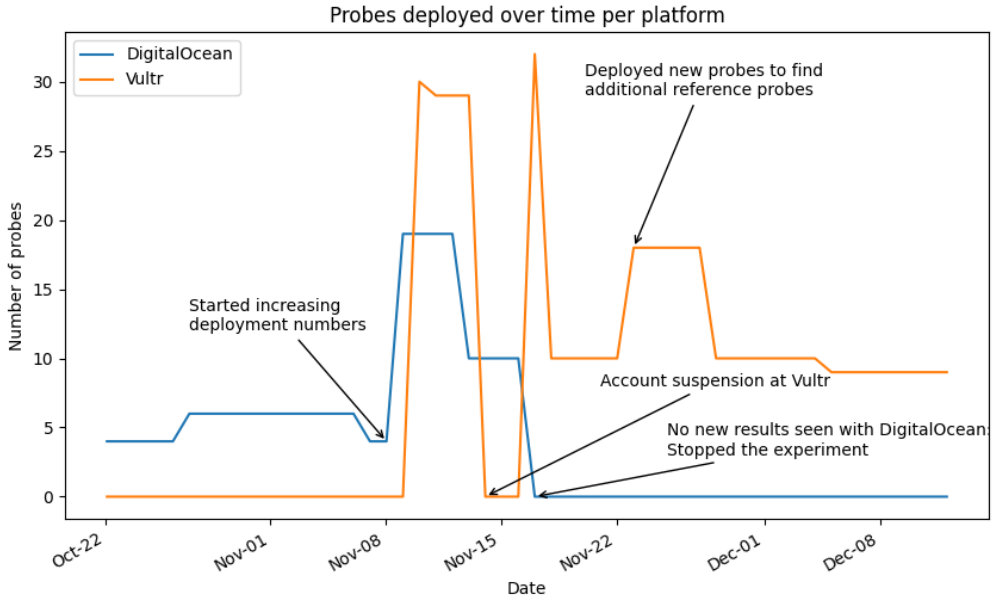

Figure 4.2: Graph showing the deployment of every probe in the data collection environment

4.3. Observed Scanning Activity

In total a number of 39,069 connection attempts to the probes were seen spanning a period of 54 days. This was done by a total of 2424 unique source addresses which originated from 70 unique autonomous system numbers (ASN) located in 45 countries. Table 4.3 shows a breakdown of the unique ASNs per cloud provider. This shows that the DigitalOcean probes were scanned by a significantly higher number of autonomous systems compared to Vultr.

Following sections will further analyze the scanning activity that has been seen on the data collection environment.

4.3.1. Discovery

Discovery of probes by scanners turned out to be different across the two cloud providers. For DigitalOcean all probes were scanned for the first time after deployment in an average of 7648 seconds, which is just over two hours. This behavior was seen at every DigitalOcean probe. An example can be seen in figure 4.3. This shows a timeline starting at the creation of the probe and the following initial connections per ASN that were seen. This potentially shows a relation with certain discovery techniques described in section 2.3

The timeframe of discovery in DigitalOcean is in contrast to the Vultr probes, DigitalOcean probes was discovered quickly, where Vultr probes weren't discovered without any active behavior from the probes itself. It can be said that due to the randomness of the IID the scanners weren't able to find the probes on their own. It is therefore that a timeline is not available for a Vultr probe. Some of the networks that were assigned to the probes were in the hitlists and thus were known publicly. An effect of this could be seen because other addresses of the network were being scanned. The data of this scanning activity could be gathered by monitoring for Neighbor Solicitation messages. 7 networks were scanned within the first seven days of deployment. The timeframe ranged from 139,080 to 293,116 seconds, which is about 1.5 and 3.5 days before a network was discovered. Three other probes were also in the hitlist but weren't scanned within the seven days of testing. These hosts were scanned within 8 to 10 days. Showing that it does not mean that a network is immediately scanned even if a network is in a hitlist.

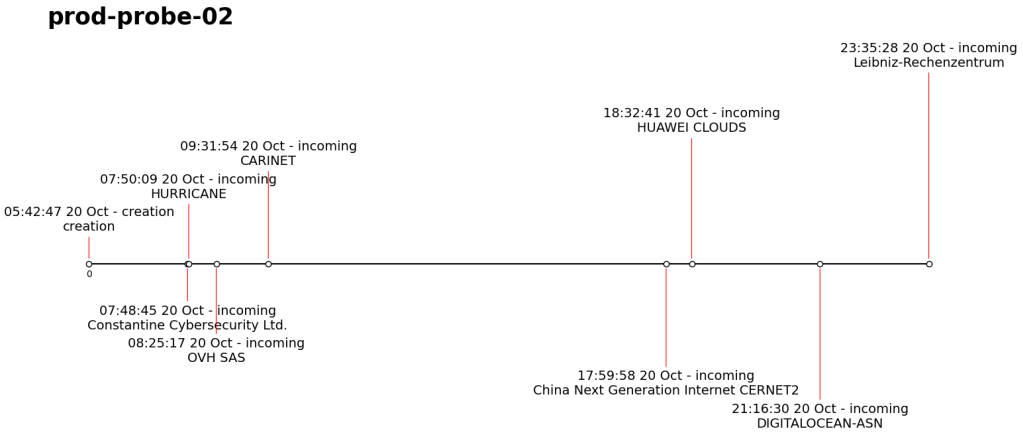

Figure 4.3: Timeline showing ASNs connecting to a probe on DigitalOcean

Protocol	Scan count
TCP	544
UDP	162
ICMPv6	408
GRE	2

Table 4.4: Scan count breakdown per layer 4 protocol

4.3.2. Scanning Analysis

In the research period multiple scans have been detected. This subsection breaks down what type of scans were seen, how many, and what characteristics were seen. The definition of a single scan is as follows: initial packets seen within a day and originating from a single /64 subnet. This definition is based upon the results of previous research where a similar definition is used. Additionally, previous research found that some IPv6 scanners use multiple addresses in a single /64 subnet [53]. Therefor, a scanning source is counted per /64 subnet. A single scan can also span multiple probes and even providers.

The number of scans detected using the definition is 958 in 54 days. This is, compared to the 617 scans in 54 days (normalized from 5199 scans in 455 days) observed by Richter et al., an increase of scanning activity of 55% [53]. The scans originated from 217 sources. The number of source IPs ranged from 1 until 437 for a single scan. 560 scans were done from only a single IP. Graph 4.4 shows the distribution of source IPs per scan, per prefix. Within each scan a breakdown can be made into which protocols were scanned for. Four protocols were seen being in use by the scanners, these were TCP, UDP, ICMPv6, and GRE. Table 4.4 shows the breakdown of every protocol. A combination of protocols were seen to be used with some scans.

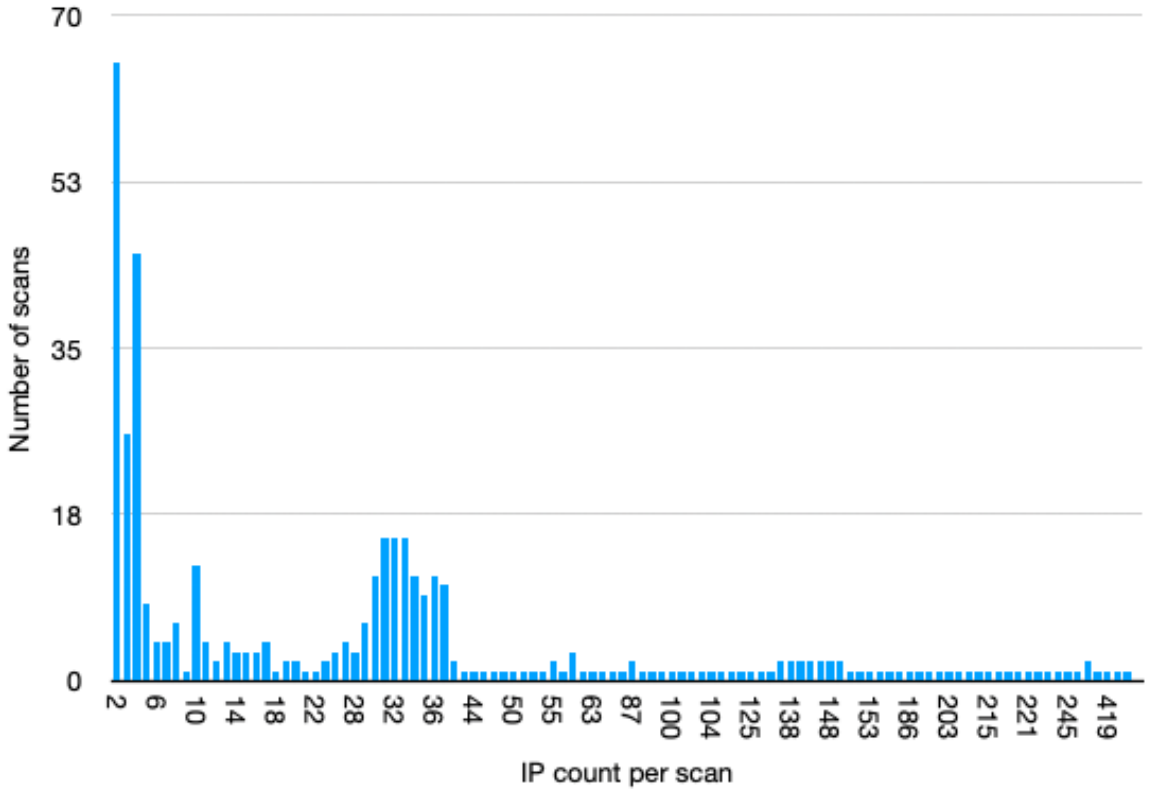

Figure 4.4: Number of scan sources per scan, per /64 prefix. 1 is removed for readability purposes

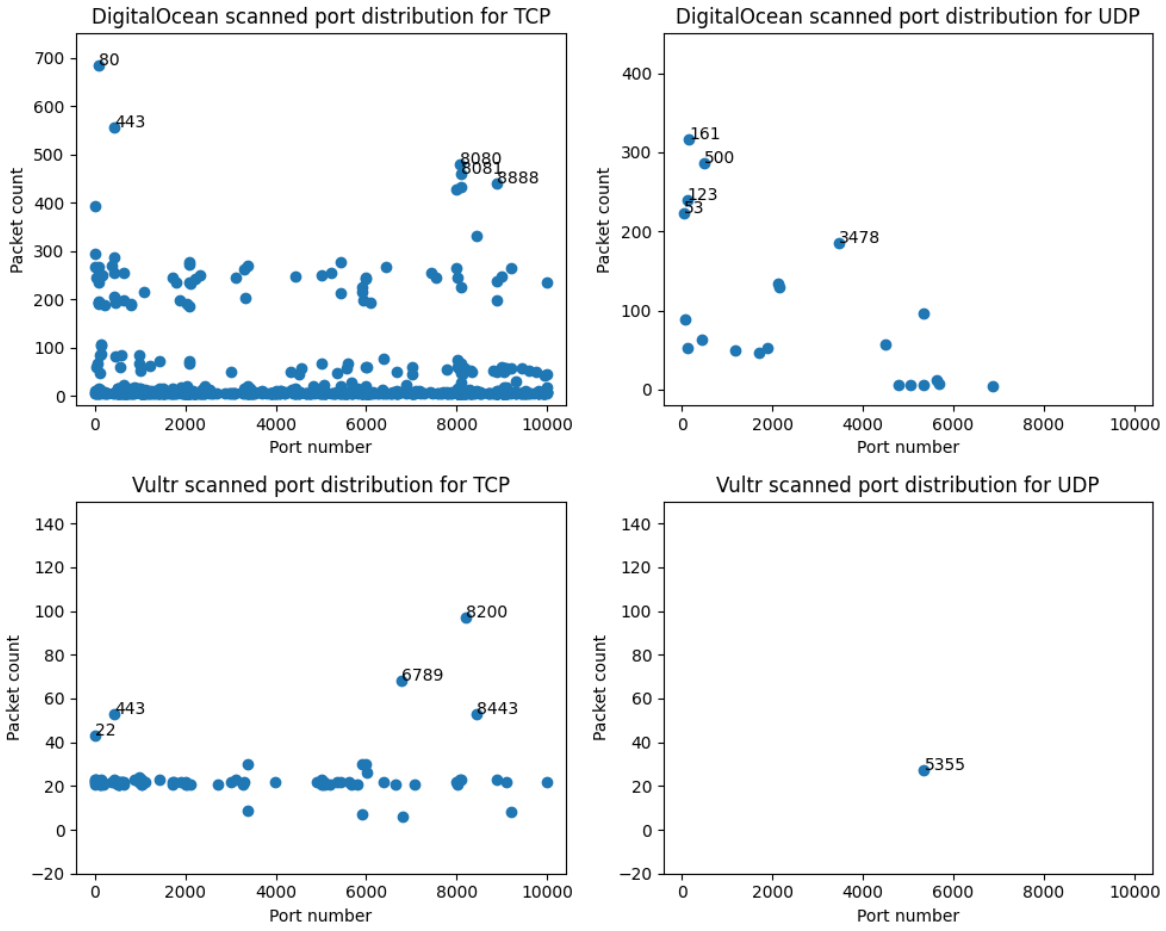

Figure 4.5: Distribution of ports that were scanned per provider. This is showing ports 0-10,000. All ports that were scanned less than 5 times are removed.

4.3.3. Port Distribution

Most scans consisted of TCP and/or UDP scans. With these protocols the connection has to be made to a network port (socket). These range from 0 tot 65535. It is interesting to see the distribution of ports that are scanned per protocol. A breakdown is made for every provider and protocol (TCP and UDP). Figure 4.5 shows the results of this analysis. This figure clearly shows more active scanning on the collection infrastructure on DigitalOcean versus the one on Vultr. It also shows a higher degree of scanning on the lower ports. This can be expected, since most services run on lower ports and scanners look for services on the internet. The port that is scanned most often are ports 80 and 443, which is mostly used for the HTTPS protocol to deliver websites to users. This is also expected since previous research showed that 443 is one of the most commonly scanned ports. Appendix A.4 contains figure A.1, which shows all ports, instead of port 0 through 10,000.

4.4. Scanner Profiles

Profiling the scanners may help understand for what purpose the scanning is performed. This section describes the results of the scanner profiles. Profiling is done per IP address and subsequently aggregate per /64 and ASN. Profiling is done based upon a reverse DNS lookup and an analysis of the organization behind the ASN.

Domain	# IPs	# Subnets	# ASN	# Connections	Providers
gatech.edu	2	2	1	2,276	Vultr; Digitalocean
onyphe.net	8	1	1	1,110	Digitalocean
rwth-aachen.de	2	1	1	361	Digitalocean
cernet2.net	3	1	1	278	Digitalocean
internet-measurement.com	282	16	2	9,866	Digitalocean
tum.de	12	1	1	300	Digitalocean
shadowserver.org	1,332	3	1	14,941	Digitalocean
mpg.de	1	1	1	15	Digitalocean
hinet.net	1	1	1	6	Digitalocean
vps.hosting	5	5	1	18	Digitalocean
time4vps.cloud	1	1	1	4	Digitalocean
appliedprivacy.net	1	1	1	4	Digitalocean
googleusercontent.com	12	12	1	32	Digitalocean
binaryedge.ninja	82	2	1	109	Vultr
nextdns.io	1	1	1	3	Digitalocean
glesys.net	2	2	2	4	Digitalocean
contaboserver.net	1	1	1	3	Digitalocean
stark-industries.solutions	6	6	1	15	Digitalocean

Table 4.5: Aggregation of the domains that were seen to be related to the IP addresses of IPv6 scanners.

4.4.1. Domains

Of the 2424 addresses seen, 1754 resolved to a domain name, which is 72%. Table 4.5 shows a list of all domains that were seen to be linked to the IP addresses of scanners. A couple of interesting statistics can be deducted based upon this data. All but two domains originated from only a single ASN. Only one ASN runs scanners from two domains, this is DigitalOcean. Table 4.6 shows a list of these ASNs with the domains that originated from them.

The top three domains account for 1704 unique IP addresses, meaning 96% of all IP addresses that could be resolved and 70% of all scanner IP addresses seen. These domains are all linked to commercial scanners. These are shadowserver.org, driftnet.io (internet-measurement.com), and binaryedge.io (binaryedge.ninja). Of those three domains the top two, shadowserver.org and driftnet.io account for 65% of all connection attempts (total number of 39,069 attempts).

Interestingly, Shodan has not been seen scanning the collection environment during this research and Censys was seen contacting only a single probe for a single time. It was expected that these IPv6 scanners would be seen extensively connecting to the collection environment, since these are widely used and recognized for their scanning activities. However, this was not the case. A reason for this wasn't found.

4.4.2. Autonomous System Number Analysis

In total, 70 unique ASNs have scanned the data collection infrastructure. Most of the scanning traffic originated from just a couple of ASNs. The first two ASNs account for 65% of all scanning activity. Table 4.7 shows a top 15 of an analysis of all ASNs that were seen. Appendix A.3 contains a full table with all ASNs that were seen during the research period. The third top ASN that scanned the infrastructure is CARINET, which had 5,510 connection attempts. There are no domains linked to the IP addresses that originated from this AS. Lastly, a large source of scanning activity originates from an education network of Georgia Tech University. The source IPs resolved to a DNS name which, when browsing to the host, reported it being part of a research project[5]. This scanner was the only scanner with significant numbers of connection attempts to the Vultr probes.

ASNs have a type which describes the activities for which the autonomous system is used. For this research this data is retrieved from *ipinfo.io* and linked to the AS numbers in the database. A breakdown

[5]The IPs resolved to scanner4.cc.gatech.edu and scanner5.cc.gatech.edu which, when browsing to this host, mentions: *This is a research scanning machine from the Georgia Institute of Technology.*

ASN	ASN Type	AS Organization	Domains
2637	education	GEORGIA-TECH	gatech.edu
16276	hosting	OVH SAS	onyphe.net
47610	education	RWTH Aachen University	rwth-aachen.de
23910	education	China Next Generation Internet CERNET2	cernet2.net
211298	isp	Constantine Cybersecurity Ltd.	internet-measurement.com
12816	hosting	Leibniz-Rechenzentrum	tum.de
6939	hosting	HURRICANE	shadowserver.org
680	education	Verein zur Foerderung eines Deutschen Forschungsnetzes e.V.	mpg.de
14061	hosting	DIGITALOCEAN-ASN	binaryedge.ninja; internet-measurement.com
3462	isp	Data Communication Business Group	hinet.net
3214	hosting	xTom GmbH	vps.hosting
62282	hosting	UAB Rakrejus	time4vps.cloud
208323	business	Foundation for Applied Privacy	appliedprivacy.net
396982	hosting	GOOGLE-CLOUD-PLATFORM	googleusercontent.com
42473	hosting	ANEXIA Internetdienstleistungs GmbH	nextdns.io
42708	hosting	GleSYS AB	glesys.net
51167	hosting	Contabo GmbH	contaboserver.net
44477	hosting	Stark Industries Solutions Ltd	stark-industries.solutions
48618	hosting	Oulun DataCenter Oy	glesys.net

Table 4.6: Domains that were seen per ASN

AS Organization	ASN	Country	# Connections	# IPs	# 64	# 56	DO	VU
HURRICANE	6939	US	15,345	1,481	3	3	10	0
Constantine Cybersecurity Ltd.	211298	GB	10,037	280	14	14	10	0
CARINET	10439	US	5,510	4	1	1	10	0
GEORGIA-TECH	2637	US	2,276	2	2	2	3	3
GOOGLE-CLOUD-PLATFORM	396982	US	1,626	42	15	15	4	0
OVH SAS	16276	FR	1,111	9	2	2	10	0
Leibniz-Rechenzentrum	12816	DE	611	22	1	1	10	0
AS-CHOOPA	20473	US	438	9	9	9	6	2
RWTH Aachen University	47610	DE	361	2	1	1	6	0
Akamai Connected Cloud	63949	US	337	263	7	7	7	3
China Next Generation Internet CERNET2	23910	CN	279	4	2	2	10	0
DIGITALOCEAN-ASN	14061	US	201	122	9	8	6	4
HUAWEI CLOUDS	136907	SG	153	8	1	1	10	0
RICAWEBSERVICES	26832	CA	98	1	1	1	10	0
LLC Baxet	51659	RU	80	1	1	1	6	0

Table 4.7: Top 15 list of all organizations seen that scanned the probe infrastructure during the research period

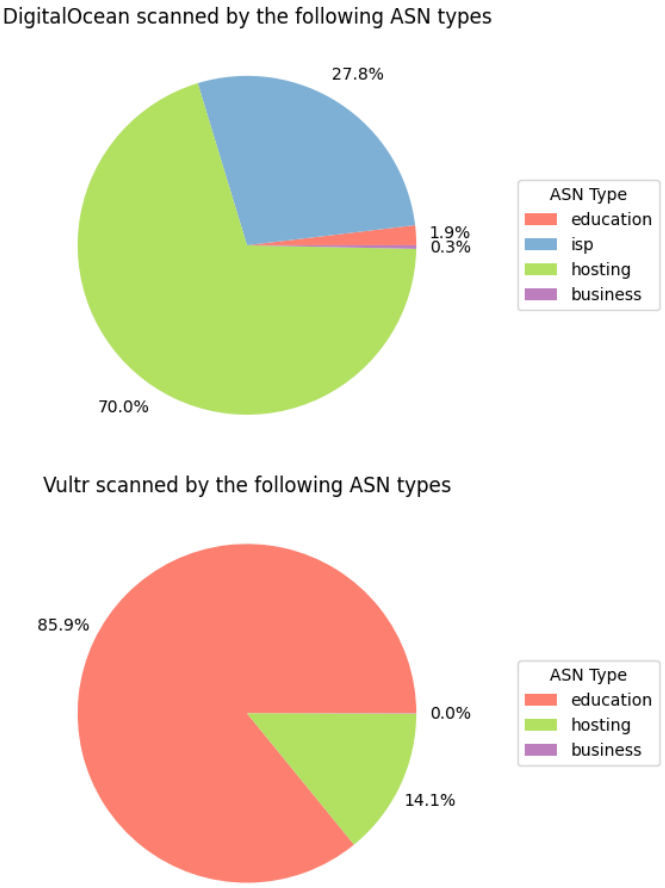

Figure 4.6: Breakdown of AS types that scanned the data collection environment. Broken down into the two providers.

can be given into what type of organizations tried connecting to the probe infrastructure based upon this data. Figure 4.6 shows this breakdown per provider. A big difference can be seen between the two providers. While the scan sources on the DigitalOcean platform mostly originated from ASNs with a hosting type, Vultr was primarily scanned by ASNs that are linked to education. Of which Georgia Tech is the only source, which scanned probes running in active mode. These probes were scanned only in limited numbers by other scanners, see section 4.5.1 for more information on this. Additionally, this large number of connection attempts seen from Georgia Tech weren't observed on DigitalOcean probes.

4.5. Sensitivity Analysis

One of the goals of this research is to find what behavior of a system may result into it being discovered faster. During this research 8 probes were found to be suitable for this part of the research (section 4.2.2 describes how these 8 were selected). These probes would either be used as a 'active' probe, where it would mimic behavior from a normal server or client, or continue as a 'silent' probe, which would continue to collect data as a reference to compare to the results of the active probes. The active/silent split was 50/50, meaning 4 probes were used as an active probe and an equal amount was used to collect reference data.

The number of probes was less than initially expected. With only 4 probes being available for testing, the number of tests that were possible reduced significantly. It was therefore decided that three tests would be conducted. Two of these experiments would each run on a single probe and the last experiment would run on two probes, but spread out across different datacenters. The following tests were defined:

1. Webcrawler which would run a single time, starting at the hostname *https://www.cnn.com* and

continue connecting to different hosts up until 300 unique hostnames.

2. NTP crawler which would run once and connect to 260 servers that were linked to various subdomains of pool.ntp.org.
3. NTP crawler which would run hourly and connect to 266 servers that were linked to various subdomains of pool.ntp.org.

The passive probes did show to have some outgoing activity to servers of the cloud provider and to telemetry servers of the operating system. This behavior was also seen on the active probes. Based on the data seen at the reference probes all of this traffic was removed during analysis. Differences were seen in the reference data across the different datacenters that the probes were deployed to. In future research this factor needs to be taken into account for accuracy purposes and reference probes should therefore be placed into the same datacenter as the active probes. This however did not affect the outcome or reliability of the results shown in this research.

4.5.1. Analysis on Active Probes

The results of the tests can be seen in figure 4.7. The data shows almost immediate scanning traffic after the probes started becoming active. This only applied to active probe 2[6], 3 and 4. Active probe 1 did not show any scanning traffic after the activity had occurred and therefore stayed 'hidden'. The reference probes continued to show no incoming traffic throughout the testing phase and showed that without activity the probes continued to stay hidden.

On active probe 2, 3, and 4 scanning patterns were observed almost immediately (within half an hour) after connecting to the NTP servers. About a day later two other actors were shown to scan the active probes. Figure 4.8 shows a timeline of one of the active probes. This exact behavior was observed on all other active probes. Appendix A.4.2 contains timelines of the other 2 active probes on which scanning activity was observed after becoming an active probe.

Scanning was observed from three parties, Georgia Tech, DigitalOcean (actor: binaryedge.io), and Akamai (of which the actor could not be ascertained). The three actors showed different behavior when scanning continued and stopped. Georgia Tech stopped scanning after the active probe stopped connecting to NTP servers. This is in contrast to the other two actors, which continued scanning after active 2 stopped. Active probe 3 and 4 observed continued scanning throughout the experiment. Showing that some parties may employ a timer for populating their own hitlist.

[6]Active probe 2 was deployed before probe 3 and 4 and could therefore be used earlier for testing.

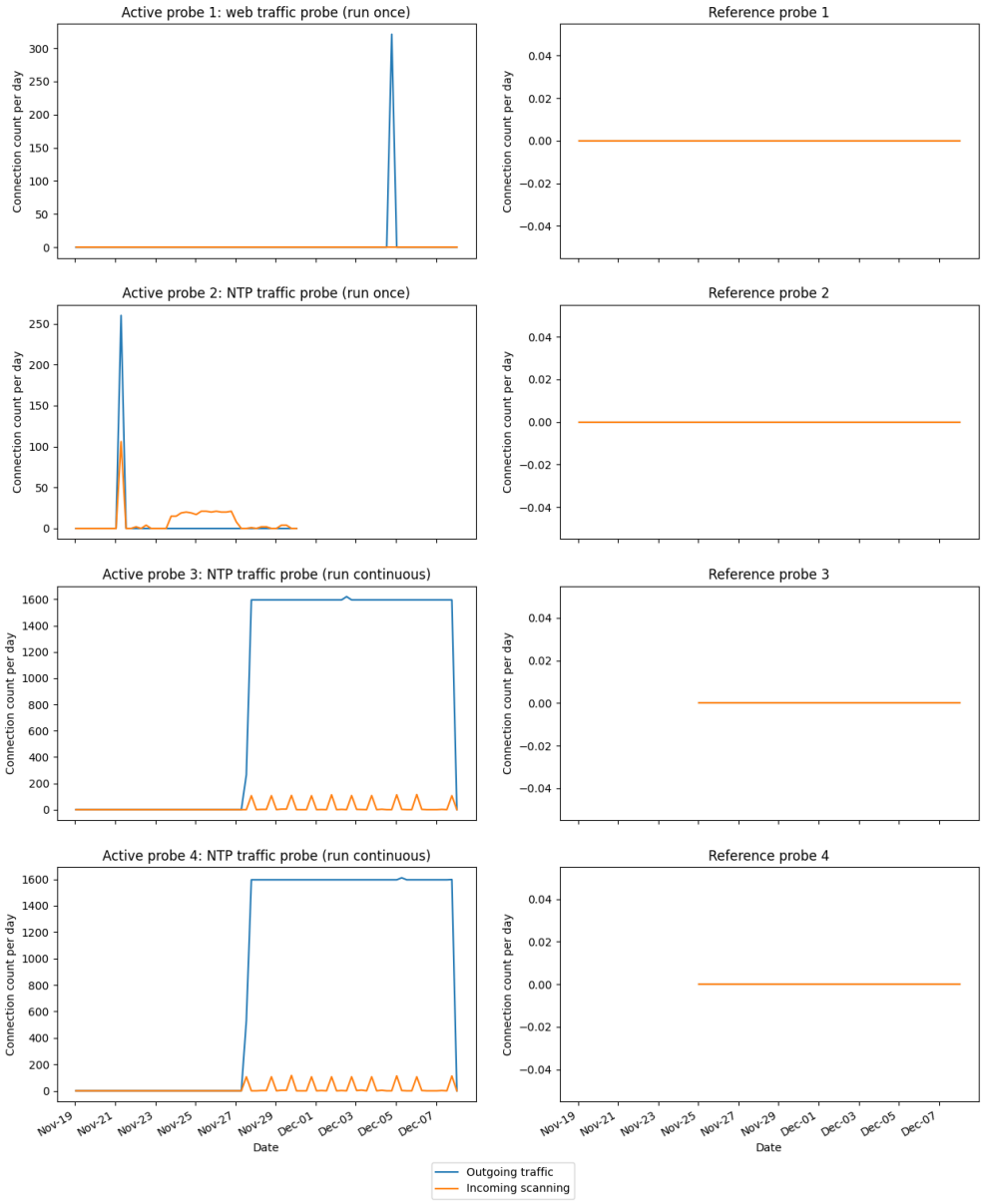

Figure 4.7: Comparison between active and passive probes, showing outgoing behavior and incoming scanning traffic. Data points are buckets of 6 hours. Reference probes 3 and 4 were later deployed after determining that there were too little reference probes.

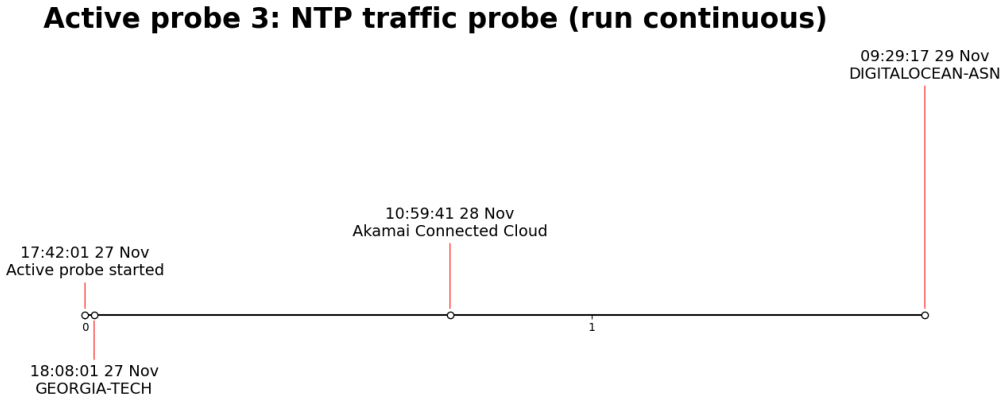

Figure 4.8: Timeline showing active probe 3. The timeline starts when the first outgoing connection was initiated.

5 Discussion

This research looked at IPv6 scanning on the internet. It tried to identify if, and for how long, it is possible to stay 'hidden' on the Internet before Internet scanners find a new, unused (fresh), or undiscovered address. It further tested what factors may leak the IP address to scanners and therefore may expose the attack surface to a broader group of actors. Finally, it looked at the addressing strategies used by cloud providers in assigning addresses to deployed services by customers.

This chapter will interpret the results and compares it to previous work, it also describes the limitations of the results and the dataset. It follows with a section on the implications of the results on the security of devices that are connected to the internet. Based upon this, section 5.3 gives recommendations for administrators of Internet connected devices and for cloud providers that assign addresses to customers' services. It finalizes with a reflection on the writing of this thesis.

5.1. Interpretation of results

In this study looked at both the effectiveness and extent of IPv6 scanning at cloud platforms. The research started with a theoretical study and followed with research to observe the scanning in the wild. This section will discuss the results of this research and compares this with the results of the theoretical study, which is described in section 2, Related Work.

5.1.1. Effectiveness

The results of previous work showed that the effectiveness of IPv6 scanners is dependent on the behavior of the targets itself. When hosts start connecting with external services the IP address may leak and expose the device and network to potential scanners. These results were affirmed in this research. These results were based upon addresses that were hidden before and where address assignment was in control by the administrator of the devices.

IT infrastructure is increasingly deployed on cloud platforms. These cloud providers are responsible for assigning addresses to the services that customers deploy on their platforms. Cloud providers use different addressing strategies for this. This research showed that the addressing strategy used by the cloud provider heavily influences the timeframe in which an IPv6 address is found on the internet. A comparison was made by deploying probes on two cloud platforms that had different addressing strategies. One platform had an approach that was relatively easy to predict, and the other used random addressing based upon the EUI-64 standard and giving the customer a /64 network block, so addresses would be highly customizable.

The results show that probes that were deployed using a predictable approach were scanned within an average of two hours. This is in contrast to the randomized approach where probes would not be found by scanners. This matches the privacy and security issues raised by RFC 7421 regarding smaller subnet sizes. The paper from Tanveer et al. showed that addresses were discovered when using external services. This research used two methods that were discussed in the paper. By connecting to public NTP servers associated with pool.ntp.org, and by connecting to various websites. When connecting

to NTP servers the IP address was discovered by multiple scanners, affirming the results of Tanveer et al.. The probe that connected to various websites wasn't found by other scanners. While this is in contrast with the results from that paper, this may be caused by the limited number of webtraffic that was generated. The probe connected to 300 websites compared to the 2.6K websites that were accessed in the paper.

5.1.2. Extent

In total, 958 scans were observed during the research period consisting of 39,069 connection attempts. These originated from 2424 unique source addresses and 70 unique autonomous systems. These showing an increase of 55% in scanning activity compared to the results from Richter et al. in 2022. Scans were done using TCP, UDP, ICMPv6 and GRE protocols. Where TCP and ICMPv6 were the most active protocols. The scans mostly focused on the lower number ports, like 80 and 443, also corresponding with the activity seen in previous studies.

Based upon the results of previous research, a scan was defined as a /64 prefix that made a connection attempt to one or more addresses of the probe infrastructure. Many scanners use a single, or a couple, addresses during a scan. Some scanners may use many hundreds of unique addresses. A scan should therefore never be classified based on a single address, but based on the /64 prefix or higher.

By using reverse DNS lookups it was possible to attribute 72% of the scanning activities to an actor by using reverse DNS lookups. This showed that most of the traffic originated from commercial scanners or from educational institutions. When analyzing the individual organizations it was observed that 65% of all scanning traffic originated from two commercial scanners and 70% of all source IPs originated from the top 3 scanners.

5.1.3. Unexpected results

With the results of this research much of the results of previous work was affirmed. Some results showed new insights into the scanning in the wild. A key finding, which was unexpected based on previous work, was that the address assignment strategy heavily impacts how effective an IPv6 scanner can be. Newly assigned address were scanned very quickly when using predictable addressing. This compared to previous studies and the results from another cloud platform which used a different addressing strategy. Additionally, the extent of scanning grew significantly. This study observed an increase by 55% compared to previous research in 2022.

5.2. Implications

The results presented in this study show that it is possible to stay 'hidden' for longer periods of time. Even when deploying services on public cloud platforms. This requires a couple of key conditions, one being that the addressing strategy of the cloud provider needs to be based on a random approach, and secondly the address needs to be silent and limit the interaction with external services. Especially the last condition can be very difficult, since to use the Internet it is imperative that connections with external services will be made. Previous research and this study showed that there are numerous ways that expose the IPv6 address to potential scanners.

IPv6 scanning is becoming more active. Especially by commercial scanners who are getting more effective in finding IPv6 addresses on the internet. In IPv4 this scanning is common practice and is on a much larger scale compared to IPv6. While the IPv6 scanning is by far not reaching the levels of IPv4, the increase of scanning traffic can be observed. It is therefore also questionable if staying 'hidden' is a viable security measure. Based on this, and previous research, it can be stated that it is not. So while it may provide a slight benefit for security on the short-term, it should never be used as an argument for mid-to-long-term security because the attack surface will eventually be discovered anyway. This is the case for devices on the Internet providing a service, which usually requires a static IP address, so clients can connect to it.

Rotation algorithms do actually provide additional privacy measures for devices and may improve security since devices have to be re-discovered after a certain amount of time. These are less useful for services on the internet, since this causes addresses to change periodically which may impact the availability of the services. However, originally it was said that these may be useful for clients or

Internet of things (IoT) devices. Based on the results of this study it showed that scanners that have direct access to external services that capture IPv6 addresses are effective in quickly re-discovering new addresses. This does have an impact on the exposure of devices and therefore the attack surface is known to a broader range of threat actors. While this may seem serious, it again relies on the principle of security-through-obscurity which in the end doesn't protect the client, but slows down an attacker.

5.3. Recommendations

The implications show that there is little change to the way IPv6 security is linked to the increased difficulty of the discovery of new addresses by IPv6 scanners. While the effectiveness of the IPv6 scanners are lower compared to IPv4, addresses will always be found due to the behavior of devices that are assigned an IPv6 address. Therefore, the recommendations for running Internet connected services and clients/IoT devices do not change: Administrators should expect these devices to be found anyway and protect these devices accordingly. For devices like clients and IoT devices it is still beneficial to rotate the addresses frequently just for the added level of privacy.

A new recommendation can be defined for cloud providers. This study showed a significant difference between the ease of discovery between the two addressing strategies. One provider used a strategy that used predictable addresses and therefore the scanners were very effective in finding new addresses. The other provider used a random approach to assign a network and address to a service and follows the recommendations set out in RFC 7421. This rendered IPv6 scanners ineffective.

The second provider also assigned a /64 network to a customer service, which allows the customer to change the address frequently without being predictable for scanners. The first provider assigned a /124 address which was not random and could be easily predicted. With a /124 network the customer can change the address to 16 different addresses. This makes it easy for scanners to enumerate all possible addresses of a network and doesn't change the effectiveness op IPv6 scanners. Such a limited number of IPv6 addresses per service can be expected when IPv6 usage grows. Cloud providers may opt for smaller network blocks per service in order to use the space more efficiently. However, the network bits should then be randomized as much as possible, to reduce the effectiveness of scanners and always be bigger than a /80, in order to follow the recommendations set forth in RFC 7421.

5.3.1. Limitations

While the results of this research match previous research, the number of data points are significantly less compared to other research in this field. This was due to a number of factors, one being the limited timeframe in which the research had to be conducted, and secondly the additional complicating factor of using external infrastructure. For example, the account which was used to perform the research on one of the cloud platforms was suspended when automatically deploying probes. It was seen as malicious behavior. This slowed down the research, since a limit was set on the number of probes that could be run at the same time.

Another limitation was the communications between the cloud providers and the researchers. It may be possible that IP addresses are leaked to external parties by the cloud providers. No evidence was found during this research that indicates this happens, but this cannot be ruled out. In future research it would be beneficial to contact the cloud providers and request this type of information. A collaboration between the researchers and the cloud platforms would even be more valuable and provide more accurate results.

6

Conclusion

The extent and effectiveness of IPv6 scanners have been studied in this thesis. This is done by collecting data using monitoring probes at different cloud providers. Previous study analyzed the effectiveness of scanners using own infrastructure, which created an environment that ensured that all variables were controlled. The goal of this research was to monitor the timeframe in which a new IPv6 address was discovered, how often, and what type, of IPv6 scanning is performed. Additionally, three experiments have been conducted where the probes set up connections to external services to expose the addresses to potential scanners. At last, some actors that are running the scans have been identified by enriching the data that is collected.

The data itself was collected and stored using two separate methods to ensure the integrity of the collected data and to allow for verification during the data analysis process. All data was stored in portable file formats to allow the data to be shared. This creates the opportunity to review the results and to use it for further research. Data analysis was done based upon previous research and by finding anomalous behavior in the data. All experiments were compared against reference systems. Finally, all results were compared between the two cloud providers and checked against the results of previous studies.

6.1. Research question

The primary research question of this thesis is: *What is the extent and effectiveness of current public IPv6 scanners in public cloud environments?*. This question was divided into sub-questions:

1. In what timeframe does a scanner find a fresh public IPv6 address?
2. Which factors influence the timeframe in which a fresh IPv6 address is found?
3. Which scan patterns can be recognized on IPv6?
4. What protocols reveal the identity of an IPv6 address the fastest?
5. How to mitigate against easy discovery?

6.2. Summary

This study showed that IPv6 scanning is becoming more active. The discovery of addresses varies significantly between the two cloud providers. At one cloud provider addresses are discovered within an average of 2 hours, compared to the other providers of which some probes were never discovered. Provided that these probes were silent and did not connect to external services. This is mostly due to the technique used by the provider when assigning an address. One provider uses a technique which is predicable for scanners and limits the number of guesses a scanner has to do. The other provider uses a technique which creates a more random address which therefore makes it significantly more complicated for scanners to guess the addresses.

Three experiments were performed that exposed the addresses of the probes to different external services. This was done using probes that were assigned a network that wasn't discovered by scanners before. The results show that by exposing the address to other services, the probes were found by IPv6 scanners. The scanners were most effective when the probes connected to NTP servers. These NTP servers were potentially run by the scanners itself to collect IPv6 addresses that connect to them. At least three scanners were shown to use this technique to collect active IPv6 addresses and networks.

All probes on one provider were assigned predictable addresses and were scanned very actively. This created the opportunity to, more easily than expected, analyze the extent of IPv6 scanning. In total the research period lasted for 54 days in which 958 scans were detected which originating from 217 sources, which in total used 2424 different IP addresses. Half of the scans that were observed used a single IP address for scanning, other scans used multiple (between 2 and 437) addresses for a single scan. Scanners scanned using the transport layer protocols TCP, UDP, ICMPv6 and GRE. For TCP and UDP the lower bound ports were scanned the most. 72% of the scanning activities could be attributed to an actor by using reverse DNS lookups. This showed that most of the scanning is done by commercial scanners and educational institutions. This corresponds with the data collected from the ASNs that were seen and the type of business that linked to that ASN.

In this thesis the extent and the effectiveness of scanners was studied. Where previous research used self-managed networks and addresses, this research focused on IPv6 scanning at cloud platform. This showed that the technique of assigning addresses, or addressing strategy, of cloud providers can make a significant difference for the discovery of new addresses by scanners. Additional results of this thesis show that IPv6 scanning is becoming more active and that it is mostly performed by commercial parties. Both the results of the extent and effectiveness matched or even exceeded the results of previous research, even when this is done in a different environment.

6.3. Future work

Seeing that the IPv6 scanning is becoming more active on the Internet and that addresses are getting discovered based upon various factors, further research may help better understand the topic of IPv6 discovery and scanning. A further expansion on the scanning practices in cloud, business, critical infrastructure and home networks will help determine the different factors that attribute to the discovery of new addresses. Great progress into the quality of the research can be achieved by cooperating with (internet) service providers and by collecting network data at various strategic capture points on the internet. Additionally, the type of scanning, IPv6 specific attacks, and attribution can give more insight into this topic.

Furthermore, research into the topic of IPv6 security in general helps to protect organizations that implement IPv6. It may also help create a safer IPv6 internet. An example of this is research into methods of detecting address harvesting systems. For example, legitimate solutions, like the NTP pool, are abused for the collection of new and active IPv6 addresses. It would help to create a method that detects these collection systems and removes these from the NTP pool. This will also help improve the privacy of users that run IPv6.

Lastly, there are big differences seen in the adoption of IPv6 between countries. Research is needed to understand why this is the case and how organizations can be motivated into implementing IPv6. This helps further advance the Internet as a whole.

References

[1] Zakir Durumeric, Eric Wustrow, and J. Alex Halderman. "ZMap: Fast Internet-wide Scanning and Its Security Applications". In: *22nd USENIX Security Symposium (USENIX Security 13)*. Washington, D.C.: USENIX Association, Aug. 2013, pp. 605–620. ISBN: 978-1-931971-03-4. URL: %5Curl%7Bhttps://www.usenix.org/conference/usenixsecurity13/technical-sessions/paper/durumeric%7D.

[2] Kensuke Fukuda and John Heidemann. "Who Knocks at the IPv6 Door?: Detecting IPv6 Scanning". In: *Proceedings of the Internet Measurement Conference 2018*. Boston MA USA: ACM, Oct. 2018, pp. 231–237. ISBN: 978-1-4503-5619-0. DOI: 10.1145/3278532.3278553. (Visited on 07/31/2023).

[3] *GreyNoise Is the Source for Understanding Internet Noise*. https://www.greynoise.io/. (Visited on 01/09/2024).

[4] *The Spamhaus Project*. https://www.spamhaus.org/. (Visited on 01/09/2024).

[5] *Shodan*. https://www.shodan.io. (Visited on 12/31/2023).

[6] *Exposure Management and Threat Hunting Solutions*. https://censys.com/. (Visited on 12/31/2023).

[7] *Spamhaus IPv6 Blocklists Strategy Statement*. https://www.spamhaus.org/organization/statement/-012/spamhaus-ipv6-blocklists-strategy-statement. (Visited on 01/02/2024).

[8] Philipp Richter and Arthur Berger. "Scanning the Scanners: Sensing the Internet from a Massively Distributed Network Telescope". In: *Proceedings of the Internet Measurement Conference*. Amsterdam Netherlands: ACM, Oct. 2019, pp. 144–157. ISBN: 978-1-4503-6948-0. DOI: 10.1145/3355369.3355595. (Visited on 11/22/2023).

[9] Steve E. Deering and Bob Hinden. *Internet Protocol, Version 6 (IPv6) Specification*. Request for Comments RFC 1883. Internet Engineering Task Force, Dec. 1995. DOI: 10.17487/RFC1883. (Visited on 06/22/2023).

[10] ICANN. *Available Pool of Unallocated IPv4 Internet Addresses Now Completely Emptied*. https://itp.cdn.icann.org/-en/files/announcements/release-03feb11-en.pdf. (Visited on 12/31/2023).

[11] *APNIC - APNIC IPv4 Address Pool Reaches Final /8*. https://web.archive.org/-web/20110807162057/http://www.apnic.net/publications/news/2011/final-8. Aug. 2011. (Visited on 12/31/2023).

[12] *LACNIC Enters IPv4 Exhaustion Phase | The Number Resource Organization*. https://www.nro.net/lacnic-enters-ipv4-exhaustion-phase. (Visited on 12/31/2023).

[13] *ARIN IPv4 Free Pool Reaches Zero*. https://www.arin.net/vault/announcements/20150924/. Sept. 2015. (Visited on 12/31/2023).

[14] Publication date: 25 Nov 2019- NEWS, IPV4, Ipv4 Depletion, IPV6, and Press Release. *The RIPE NCC Has Run out of IPv4 Addresses*. https://www.ripe.net/publications/news/the-ripe-ncc-has-run-out-of-ipv4-addresses. (Visited on 12/31/2023).

[15] *IPv4 Exhaustion - AFRINIC - Regional Internet Registry for Africa*. https://web.archive.org/-web/20200915081117/https://afrinic.net/exhaustion. Sept. 2020. (Visited on 12/31/2023).

[16] *Internet Protocol*. Request for Comments RFC 791. Internet Engineering Task Force, Sept. 1981. DOI: 10.17487/RFC0791. (Visited on 12/31/2023).

[17] Brian E. Carpenter, Tim Chown, Fernando Gont, Sheng Jiang, Alexane Petrescu, and Anew Yourtchenko. *Analysis of the 64-Bit Boundary in IPv6 Addressing*. Request for Comments RFC 7421. Internet Engineering Task Force, Jan. 2015. DOI: 10.17487/RFC7421. (Visited on 12/31/2023).

[18] Fernando Gont. *A Method for Generating Semantically Opaque Interface Identifiers with IPv6 Stateless Address Autoconfiguration (SLAAC)*. Request for Comments RFC 7217. Internet Engineering Task Force, Apr. 2014. DOI: 10.17487/RFC7217. (Visited on 11/22/2023).

[19] Tomek Mrugalski, Marcin Siodelski, Bernie Volz, Anew Yourtchenko, Michael Richardson, Sheng Jiang, Ted Lemon, and Timothy Winters. *Dynamic Host Configuration Protocol for IPv6 (DHCPv6)*. Request for Comments RFC 8415. Internet Engineering Task Force, Nov. 2018. DOI: 10.17487/RFC8415. (Visited on 12/31/2023).

[20] Steve E. Deering and Bob Hinden. *IP Version 6 Addressing Architecture*. Request for Comments RFC 4291. Internet Engineering Task Force, Feb. 2006. DOI: 10.17487/RFC4291. (Visited on 12/31/2023).
[21] Thomas Narten, Richard P. Draves, and Suresh Krishnan. *Privacy Extensions for Stateless Address Autoconfiguration in IPv6*. Request for Comments RFC 4941. Internet Engineering Task Force, Sept. 2007. DOI: 10.17487/RFC4941. (Visited on 12/31/2023).
[22] *IPv6 security*. https://support.apple.com/nl-nl/guide/security/seccb625dcd9/web. (Visited on 11/22/2023).
[23] *Ipv6: NetworkManager Reference Manual*. https://developer-old.gnome.org/NetworkManager/stable/settings-ipv6.html. (Visited on 12/31/2023).
[24] Lorenzo Colitti, Steinar H Gunderson, Erik Kline, and Tiziana Refice. "Evaluating IPv6 Adoption in the Internet". In: *International Conference on Passive and Active Network Measurement*. Springer, 2010, pp. 141–150.
[25] John Pickard, John Southworth, and Dale Drummond. "The IPv6 Internet: An Assessment of Adoption and Quality of Services". In: *Journal of International Technology and Information Management* 26.2 (2017), pp. 48–64.
[26] *IPv6 10 Years Out: An Analysis in Users, Tables, and Traffic*. https://labs.ripe.net/author/wilhelm/ipv6-10-years-out-an-analysis-in-users-tables-and-traffic/. June 2022. (Visited on 06/22/2023).
[27] Xuequn Wang and Sebastian Zander. "Extending the Model of Internet Standards Adoption: A Cross-Country Comparison of IPv6 Adoption". In: *Information & Management* 55.4 (June 2018), pp. 450–460. ISSN: 03787206. DOI: 10.1016/j.im.2017.10.005. (Visited on 11/21/2023).
[28] RIPE. *IPv6Security-Slides-2.Pdf*.
[29] Mehdi Nikkhah. "Maintaining the Progress of IPv6 Adoption". In: *Computer Networks* 102 (June 2016), pp. 50–69. ISSN: 13891286. DOI: 10.1016/j.comnet.2016.02.027. (Visited on 11/21/2023).
[30] *IPv4 vs IPv6, What Is IPv4, What Is IPv6, IPv6 to IPv4 Basis - FS.COM | FS Community*. https://community.fs.com:7003/article/ipv4-vs-ipv6-whats-the-difference.html. Sept. 2021. (Visited on 01/05/2024).
[31] *IPv4 vs IPv6 – What's the Difference and Which Is Better?* https://blog.servermania.com/ipv4-vs-ipv6. (Visited on 01/05/2024).
[32] Ali Albkerat and Biju Issac. "Analysis of IPv6 Transition Technologies". In: *arXiv preprint arXiv:1410.2013* (2014). arXiv: 1410.2013.
[33] *Active Scanning, Technique T1595 - Enterprise | MITRE ATT&CK®*. https://attack.mitre.org/techniques/T1595/. (Visited on 12/31/2023).
[34] Lion Steger, Liming Kuang, Johannes Zirngibl, Georg Carle, and Oliver Gasser. *Target Acquired? Evaluating Target Generation Algorithms for IPv6*. July 2023. arXiv: 2307.06872 [cs]. (Visited on 11/21/2023).
[35] *Shodan Search*. https://www.shodan.io/search?query=has_ipv6%3Atrue. (Visited on 08/03/2023).
[36] *Labels=ipv6 - Host Search*. https://search.censys.io/search. (Visited on 11/21/2023).
[37] Said Jawad Saidi, Oliver Gasser, and Georgios Smaragdakis. *One Bad Apple Can Spoil Your IPv6 Privacy*. Mar. 2022. arXiv: 2203.08946 [cs]. (Visited on 07/31/2023).
[38] Fernando Gont, Suresh Krishnan, Thomas Narten, and Richard P. Draves. *Temporary Address Extensions for Stateless Address Autoconfiguration in IPv6*. Request for Comments RFC 8981. Internet Engineering Task Force, Feb. 2021. DOI: 10.17487/RFC8981. (Visited on 12/31/2023).
[39] Erik C. Rye, Robert Beverly, and kc claffy. "Follow the Scent: Defeating IPv6 Prefix Rotation Privacy". In: *Proceedings of the 21st ACM Internet Measurement Conference*. Nov. 2021, pp. 739–752. DOI: 10.1145/3487552.3487829. arXiv: 2102.00542 [cs]. (Visited on 12/31/2023).
[40] Tuomas Aura. *Cryptographically Generated Addresses (CGA)*. Request for Comments RFC 3972. Internet Engineering Task Force, Mar. 2005. DOI: 10.17487/RFC3972. (Visited on 12/31/2023).
[41] Alissa Cooper, Fernando Gont, and Dave Thaler. *Security and Privacy Considerations for IPv6 Address Generation Mechanisms*. Request for Comments RFC 7721. Internet Engineering Task Force, Mar. 2016. DOI: 10.17487/RFC7721. (Visited on 12/31/2023).
[42] Mans Nilsson and Magnus Danielson. *Complex Addressing in IPv6*. Request for Comments RFC 8135. Internet Engineering Task Force, Apr. 2017. DOI: 10.17487/RFC8135. (Visited on 12/31/2023).
[43] Oliver Gasser, Quirin Scheitle, Sebastian Gebhard, and Georg Carle. *Scanning the IPv6 Internet: Towards a Comprehensive Hitlist*. July 2016. arXiv: 1607.05179 [cs]. (Visited on 07/31/2023).

[44] *Internet Protocol Version 6 Address Space*. https://www.iana.org/assignments/ipv6-address-space/ipv6-address-space.xhtml. (Visited on 08/03/2023).

[45] SANS Internet Storm Center. *Targeted IPv6 Scans Using Pool.Ntp.Org* . https://isc.sans.edu/diary.html?storyid=0. (Visited on 07/31/2023).

[46] Fernando Gont and Tim Chown. *Network Reconnaissance in IPv6 Networks*. Request for Comments RFC 7707. Internet Engineering Task Force, Mar. 2016. DOI: 10.17487/RFC7707. (Visited on 07/31/2023).

[47] Pawel Foremski, David Plonka, and Arthur Berger. "Entropy/IP: Uncovering Structure in IPv6 Addresses". In: *Proceedings of the 2016 Internet Measurement Conference*. Nov. 2016, pp. 167–181. DOI: 10.1145/2987443.2987445. arXiv: 1606.04327 [cs, math]. (Visited on 08/03/2023).

[48] Austin Murdock, Frank Li, Paul Bramsen, Zakir Durumeric, and Vern Paxson. "Target Generation for Internet-Wide IPv6 Scanning". In: *Proceedings of the 2017 Internet Measurement Conference*. London United Kingdom: ACM, Nov. 2017, pp. 242–253. ISBN: 978-1-4503-5118-8. DOI: 10.1145/3131365.3131405. (Visited on 07/31/2023).

[49] Tianyu Cui, Gaopeng Gou, Gang Xiong, Chang Liu, Peipei Fu, and Zhen Li. "6GAN: IPv6 Multi-Pattern Target Generation via Generative Adversarial Nets with Reinforcement Learning". In: *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*. May 2021, pp. 1–10. DOI: 10.1109/INFOCOM42981.2021.9488912. arXiv: 2204.09839 [cs]. (Visited on 11/21/2023).

[50] Zhizhu Liu, Yinqiao Xiong, Xin Liu, Wei Xie, and Peidong Zhu. "6Tree: Efficient Dynamic Discovery of Active Addresses in the IPv6 Address Space". In: *Computer Networks* 155 (May 2019), pp. 31–46. ISSN: 13891286. DOI: 10.1016/j.comnet.2019.03.010. (Visited on 07/31/2023).

[51] Guanglei Song, Jiahai Yang, Zhiliang Wang, Lin He, Jinlei Lin, Long Pan, Chenxin Duan, and Xiaowen Quan. "DET: Enabling Efficient Probing of IPv6 Active Addresses". In: *IEEE/ACM Transactions on Networking* 30.4 (Aug. 2022), pp. 1629–1643. ISSN: 1063-6692, 1558-2566. DOI: 10.1109/TNET.2022.3145040. (Visited on 07/31/2023).

[52] Geoff Huston. *BGP in 2022 – the Routing Table*. Jan. 2023. (Visited on 08/02/2023).

[53] Philipp Richter, Oliver Gasser, and Arthur Berger. "Illuminating Large-Scale IPv6 Scanning in the Internet". In: *Proceedings of the 22nd ACM Internet Measurement Conference*. Nice France: ACM, Oct. 2022, pp. 410–418. ISBN: 978-1-4503-9259-4. DOI: 10.1145/3517745.3561452. (Visited on 07/31/2023).

[54] Hammas Bin Tanveer, Rachee Singh, Paul Pearce, and Rishab Nithyanand. "Glowing in the Dark: Uncovering IPv6 Address Discovery and Scanning Strategies in the Wild". In: *32nd USENIX Security Symposium (USENIX Security 23)*. Anaheim, CA: USENIX Association, Aug. 2023, pp. 6221–6237. ISBN: 978-1-939133-37-3. URL: %5Curl%7Bhttps://www.usenix.org/conference/usenixsecurity23/presentation/bin-tanveer%7D.

[55] *SSD VPS Servers, Cloud Servers and Cloud Hosting*. https://www.vultr.com/. (Visited on 01/09/2024).

[56] *DigitalOcean | Cloud Hosting for Builders*. https://www.digitalocean.com. (Visited on 01/09/2024).

[57] *Elasticsearch Guide [8.10] | Elastic*. https://www.elastic.co/guide/en/elasticsearch/reference/8.10/index.html. Learn/Docs/Elasticsearch/Reference/8.10. (Visited on 01/09/2024).

[58] *Fleet and Elastic Agent Overview | Fleet and Elastic Agent Guide [8.10] | Elastic*. https://www.elastic.co/guide/en/fleet/8.10/fleet-overview.html. Learn/Docs/Fleet/Guide/Elastic Agent/8.10. (Visited on 01/09/2024).

[59] *Kibana Guide [8.10] | Elastic*. https://www.elastic.co/guide/en/kibana/8.10/index.html. Learn/Docs/Kibana/Reference/8.10. (Visited on 01/09/2024).

[60] *GeoLite2 Free Geolocation Data*. https://dev.maxmind.com/geoip/geolite2-free-geolocation-data. (Visited on 01/09/2024).

[61] Mukesh Gupta and Alex Conta. *Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification*. Request for Comments RFC 4443. Internet Engineering Task Force, Mar. 2006. DOI: 10.17487/RFC4443. (Visited on 12/19/2023).

[62] Wesley Eddy. *Transmission Control Protocol (TCP)*. Request for Comments RFC 9293. Internet Engineering Task Force, Aug. 2022. DOI: 10.17487/RFC9293. (Visited on 01/09/2024).

A

Appendix

A.1. Definitions

Term	Definition
Scanner	An external party that tries to find active hosts on the Internet
IP Address	a series of numbers separated by dots that identifies a particular computer connected to the Internet
NTP	Network Time Protocol: used by computers to accurately update their clock
DNS	Domain Name System: used to translate names to IP addresses and back
ICANN	Internet Corporation for Assigned Names and Numbers: it is responsible for the worldwide cooodination of IP addresses and Autonomous System Numbers
IANA	Internet Assigned Numbers Authority: The global coordination of the DNS Root, IP addressing, and other Internet protocol resources. Subsidiary of ICANN.
RIR	Regional Internet Registry: delegated numbers authority for a specific region in the world.
ARIN	American Registry for Internet Numbers: The RIR for North America and the Caribbean
RIPE	The RIR for Europe, the Middle East and Central-Asia
APNIC	The RIR for Asia Pacific region.
DHCP	Dynamic Host Control Protocol, used to assign IP addresses and other information to computers in order to operate on a specific network. Defined in RFC 2131
DHCPv6	DHCP for IPv6 addresses. Defined in RFC 8415.
SLAAC	IPv6 Stateless Address Autoconfiguration, used to for address assignments in IPv6 networks. Can work together with IPv6. Initially defined in RFC 4862.
BGP	Border Gateway Protocol, the protocol used to distribute routing information on the internet. This protocol is used to tell other networks on the Internet where IP addresses can be found. Initially defined in RFC 4271 (IPv4) and RFC 7454.
BGP Routing table	The table containing all routing information of the entire internet.
ASN	Autonomous System Number, a number which BGP uses to identify networks that contain a set of IP networks.
Addressing Strategy	Used in this thesis to define what addresses and networks are assigned to customers' servers.

Table A.1: Definition table

A.2. Public Code

All code of that is developed for this paper specifically is made public on the following repository: `https://github.com/mvlnetdev/thesis2324_code`. Documentation is provided in this repository on the different scripts that were used, and it will describe the infrastructure that was set up. In this process additional steps will be taken to clean up the code and make it usable for other projects.

Lastly, all data will be backed up and stored. This data can be requested by sending an e-mail to m.j.h.vanleeuwen@tudelft.nl.

A.3. Data Tables

AS Organization	ASN	Country	# Connections	# IPs	# /64	# /56	DO	VU
HURRICANE	6939	US	15,345	1,481	3	3	10	0
Constantine Cybersecurity Ltd.	211298	GB	10,037	280	14	14	10	0
CARINET	10439	US	5,510	4	1	1	10	0
GEORGIA-TECH	2637	US	2,276	2	2	2	3	3
GOOGLE-CLOUD-PLATFORM	396982	US	1,626	42	15	15	4	0
OVH SAS	16276	FR	1,111	9	2	2	10	0
Leibniz-Rechenzentrum	12816	DE	611	22	1	1	10	0
AS-CHOOPA	20473	US	438	9	9	9	6	2
RWTH Aachen University	47610	DE	361	2	1	1	6	0
Akamai Connected Cloud	63949	US	337	263	7	7	7	3
China Next Generation Internet CERNET2	23910	CN	279	4	2	2	10	0
DIGITALOCEAN-ASN	14061	US	201	122	9	8	6	4
HUAWEI CLOUDS	136907	SG	153	8	1	1	10	0
RICAWEBSERVICES	26832	CA	98	1	1	1	10	0
LLC Baxet	51659	RU	80	1	1	1	6	0
AMAZON-02	16509	GB	49	28	10	10	8	0
Petersburg Internet Network ltd.	44050	RU	49	3	3	3	10	0
CHINANET SiChuan Telecom Internet Data Center	38283	CN	49	1	1	1	2	0
ANEXIA Internetdienstleistungs GmbH	42473	FR	43	15	15	15	1	0
MICROSOFT-CORP-MSN-AS-BLOCK	8075	GB	31	10	10	10	4	0
Stark Industries Solutions Ltd	44477	MD	26	10	10	10	1	0
M247 Europe SRL	9009	AT	24	8	8	8	1	0
Melbikomas UAB	56630	LT	22	8	8	8	1	0
MULTA-ASN1	35916	US	20	1	1	1	4	0
xTom GmbH	3214	DE	18	5	5	5	4	0
China Education and Research Network Center	4538	CN	17	1	1	1	3	0
Verein zur Foerderung eines Deutschen Forschungsnetzes e.V.	680	DE	15	1	1	1	10	0
HOSTHATCH	63473	SE	12	3	3	3	4	0
Akenes SA	61098	CH	11	4	4	4	1	0
ScaleUp Technologies GmbH - Co. KG	29014	DE	10	10	10	10	2	0
FREE RANGE CLOUD	53356	CA	8	3	3	3	3	0

EDGEUNO SAS	7195	BR	6	3	3	3	2	0
Contabo GmbH	51167	DE	6	2	2	2	1	0
KDDI CORPORATION	2516	JP	6	1	1	1	1	0
Data Communication Business Group	3462	TW	6	1	1	1	2	0
Serverius Holding B.V.	50673	NL	6	1	1	1	4	0
Chinanet	4134	CN	4	3	2	2	2	0
Akton d.o.o.	25467	MK	4	2	2	2	1	0
31173 Services AB	39351	NL	4	2	2	2	1	0
COMCAST-7922	7922	US	4	1	1	1	1	0
UAB Rakrejus	62282	LT	4	1	1	1	2	0
Foundation for Applied Privacy	208323	AT	4	1	1	1	1	0
COGENT-174	174	CZ	3	1	1	1	1	0
Nine Internet Solutions AG	29691	CH	3	1	1	1	1	0
GleSYS AB	42708	SE	3	1	1	1	1	0
Nuno Felgueiras	44222	PT	3	1	1	1	1	0
Optimus IT d.o.o.	48894	SI	3	1	1	1	1	0
LLC Baxet	49392	RU	3	1	1	1	3	0
EDIS GmbH	57169	AT	3	1	1	1	1	0
EIS	36868	MU	2	1	1	1	1	0
UK Dedicated Servers Limited	42831	GB	2	1	1	1	1	0
MISAKA	917	US	1	1	1	1	1	0
SYNAPSECOM S.A. Provider of Telecommunications and Internet Services	8280	GR	1	1	1	1	1	0
Melbikomas UAB	8849	AE	1	1	1	1	1	0
NEXRIL	13830	US	1	1	1	1	1	0
AMANAHA-NEW	32489	CA	1	1	1	1	1	0
Trabia SRL	43289	MD	1	1	1	1	1	0
proinity GmbH	44239	CH	1	1	1	1	0	1
Shenzhen Tencent Computer Systems Company Limited	45090	CN	1	1	1	1	1	0
Oulun DataCenter Oy	48618	FI	1	1	1	1	1	0
Globalhost d.o.o.	200698	BA	1	1	1	1	1	0
CENSYS-ARIN-01	398324	US	1	1	1	1	1	0

Table A.2: Full list of all organizations seen that scanned the probe infrastructure during the research period

A.4. Figures

A.4.1. Port Scan Distribution

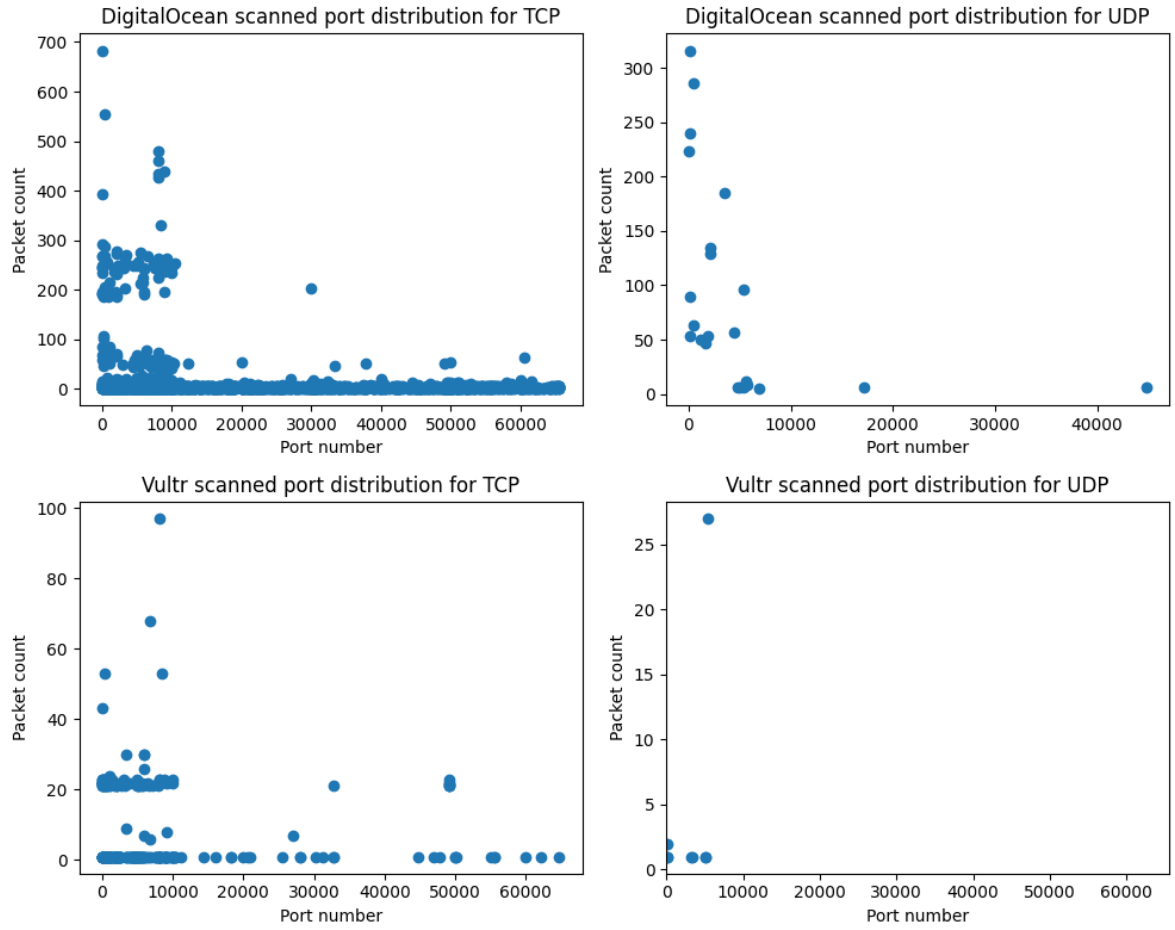

Figure A.1: Distribution of ports that were scanned per provider.

A.4.2. Active Probe Timelines

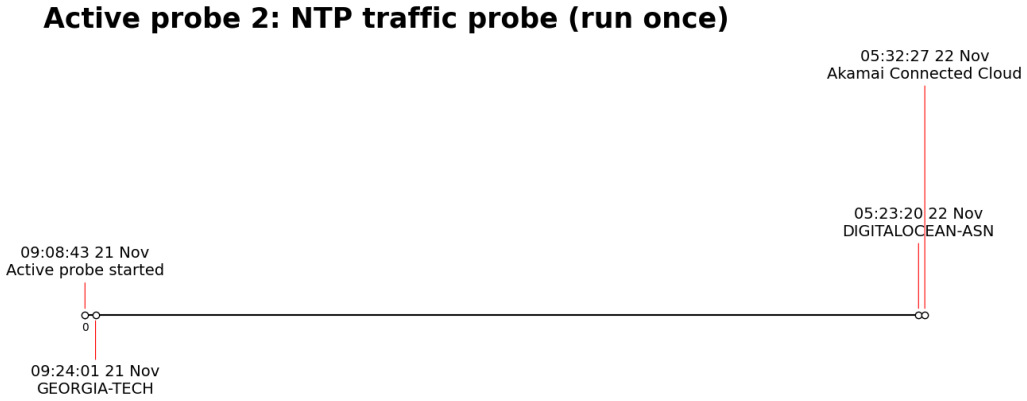

Figure A.2: Timeline showing active probe 2. The timeline starts when the first outgoing connection was initiated.

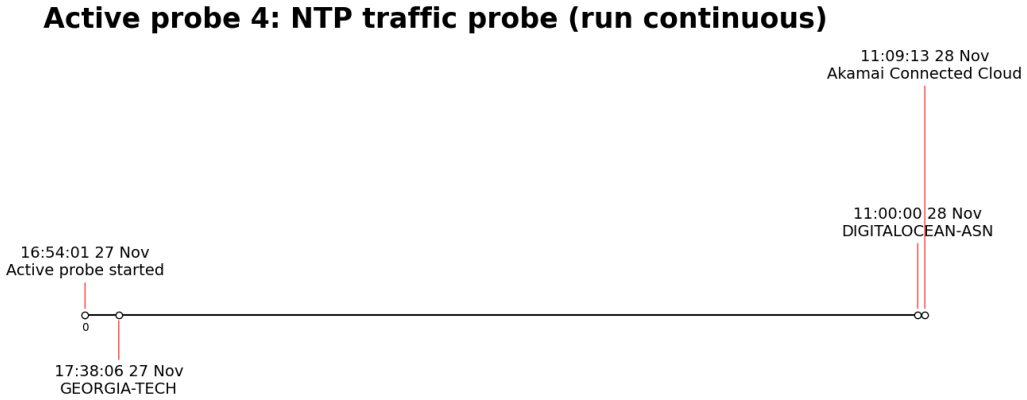

Figure A.3: Timeline showing active probe 4. The timeline starts when the first outgoing connection was initiated.