A cloud-native framework for globally distributed capture and analysis of Internet Background Radiation

Fabricio Bortoluzzi[1,3], Barry Irwin[1,2], and Carla Merkle Westphall[3]

[1]Department of Applied Computing, Noroff University College, Kristiansand, Norway
Email: {fabricio.bortoluzzi, barry.irwin}@noroff.no
[2]Department of Computer Science, Rhodes University, Grahamstown, South Africa
[3]Postgraduate Programme in Computer Science, Federal University of Santa Catarina, Florianópolis, Brazil
Email: carla.merkle.westphall@ufsc.br

Abstract*—Among the existing methods for analysing internet traffic, one focuses on unsolicited and often harmful packets, referred to as Internet Background Radiation (IBR). IBR data can be captured by using cloud-based computing instances to listen for incoming packets and recording the received headers and contents in files using the PCAP format. No services are run on these hosts so all traffic can be regarded as non-legitimate. This work presents the research plan for capturing a year- long sample of the IBR arriving to up to 1500 geographically distributed sensors across a major service cloud provider. In combination with collection, a data analysis pipeline will be constructed to enable the querying of most relevant quantitative and qualitative aspects of the resulting dataset. The ultimate goal is to answer "how is the IBR characterised when captured within the context of cloud computing?". Results will include packet distribution according to the properties of the network, transport, and application layers, together with an evaluation of traffic linking to botnet activity such as *Mirai* and *Moobot.

***Index Terms*—internet background radiation, network telescope, network packet analysis, cloud computing**

I. Introduction

Internet Background Radiation (IBR) constitutes the unsolicited traffic targeted towards public IP addresses and devices directly connected to the Internet [1], [2]. IBR traffic is generally regarded as non-productive, and may well be malicious, including flooding, scanning and worm propagation activity. A smaller proportion may be related to network and application-layer misconfigurations [3]. There are several approaches that can be used to capture such radiation. One typical scenario is the temporary allocation of unused CIDR blocks which is then forwarded to capture-enabled devices – sensors – configured to listen to all incoming traffic and record the packets in the PCAP file format for later analysis [1], [3].

Two main types of analysis on the background radiation have been discussed along the years. One is of quantitative nature, relating packet accounting and their distribution according to the network, transport, and/or application layer protocols found. Another is qualitative, in which researchers look for signs of malware activity within the capture sample, correlating the exploitation pattern seen against known threats or to acknowledge the existence of new ones. Malware activity examined in prior works include *Blaster*, *Sasser*, *Samr*, *Welchia*, *CodeRed* and *Conficker* [1]–[4]. More recently, patterns matching denial-of-service attacks were also attributed to *Mirai* botnet on similar studies [5].

This paper summarises research towards doctoral degree being undertaken by the primary author at the Postgraduate Programme in Computer Science within the Federal University of Santa Catarina, Florianópolis, Brazil. The research goal is to investigate the problems associated with current approaches for capturing and analysing Internet Background Radiation, followed by the proposition of a novel cloud-native framework. The research is planned to span 24 months from July 2023 to June, 2025.

The remainder of the paper is organised as follows. Section II examines related work. The proposed research method, problem definition, proposed solution, and research aim and goals are presented in Section III. Sections IV and V respectively discusses the intended framework architecture alongside with key aspects of its implementation and the envisaged resulting data. Finally, section VI summarises the expected results and potential outcomes sought by the research.

II. Related work

The characterisation of the Internet Background Radiation was first formalised on seminal paper by [3]. They were able to capture traffic targeting an unallocated IPv4 `/8` block alongside two `/19` subnets for two months. Their report largely established the methods and style used, and has been followed by others. The report contained a quantitative discussion on packet distribution according to the communication layer, followed by exploratory qualitative analysis related to the outbreak of *Sasser* worm. Backscatter activity was further explored in the form of time-series analysis, mainly related to the abuse of TCP flags, TCP scans and SYN-flood denial-of-service attacks.

A similar experiment [1] was built on a dataset spanning 5 years. The monitored IP address space was a complete `/8` and three partially used `/8` blocks. This resulted in the coexistence of productive and non-productive traffic, requiring filtering and classification of background radiation.

Their discussion included a temporal analysis of how malicious activity evolved over the period, followed by a spatial analysis, attempting to answer how traffic could vary based on the address block under observation. Results include the availability of eleven datasets and 10 terabytes of compressed, pcap-formatted data. Qualitative analysis includes a discussion on activity attributed to the *Conficker* malware, VoIP exploitation by means of underlying UDP protocol, and exploitation activity related to a vulnerability associated with domestic DSL routers, aiming at poisoning DNS resolver addresses. [2] focused on the application and implementation of Network Telescopes as a means for monitoring IBR activity at scale. Consideration is given to how sensors should be prepared in terms of storage capacity, logging procedures, sensor security, and data collection procedures. Analysis focused on geopolitical time-series and the creation of a visualisation interface named Inetviz. An observation window of 50 months resulted in capturing 40 million events from a dedicated `/24` block. In addition to characterising the backscatter radiation found, a thorough discussion around the evolution of the *Conficker* worm behaviour was also presented. More recently, [6] explored the use of smaller IP address pools for capturing samples of the Internet Background Radiation. Their results show that smaller and non-contiguous address blocks are able to achieve a high rate of detection and accuracy in comparison to traditionally larger ranges. IBR captures have also been used to determine the occurrence of large scale IoT-based attacks [7], global-scale internet outages [8], and country-level analysis [5].

III. Research Method

The research method applicable to this work fits in what [9] defines as "non-experimental research, in which there is no systematic influence by the researcher". It is further categorised as an objective investigation, where "any competent observer should agree with the findings and the conclusions resulting from the observation".

A. Problem Definition

Globally, the top level IPv4 address pools have been exhausted [10], [11]. Conversely, the number of devices directly connected to the Internet and using IPv4 steadily grows. As a consequence, large unused network blocks are becoming unfeasible for IBR observations. One potential solution seems to be the use of smaller, non-contiguous address sets, following the assumptions and methods defined by [6]. Analysing increasing volumes of collected IBR raises a second problem. Researchers are usually bound to a single computer when conducting their analysis. Even cluster computing is constrained by computing resources made available, and by the fact that many data extraction and processing tools, such as `tcpdump`, cannot easily be parallelised.

B. Proposed Solution

The hypothesis this work aims to investigate speculates the use of the idle capacity on cloud computing platforms to enable the capture, collation, and analysis of the Internet Background Radiation.

Cloud Computing is a commercially available service model in which Cloud Service Providers (CSP) are structured to enable broad, elastic processing power and network connectivity over the Internet [12]. The elastic nature is of special interest as the CSP has to ensure clients will perceive cloud capabilities as "unlimited, and appropriate in any quantity at any time". Therefore, CSPs must be equipped with more computer power than actually allocated to the sum of their clients at any given time. A good example is the availability of *spot*-class computer instances that can be rented under a bidding scheme for a tiny fraction of the standard pricing scheme [13]. The primary goal is to gather enough data to establish an answer to the research question: "How is the IBR characterised when captured within the context of cloud computing?"

Key baseline assumptions requiring investigation are:

- The diversity of regions and availability zones offered by cloud service providers may give unparalleled geographical capillarity to the establishment of cause-and-effect correlations generated by malware activity, denial-of-service attacks and their negative impact affecting the Internet infrastructure;
- It should be very likely that a cloud-originated dataset will contain traces of malware and denial-of-service activity worth analysing; and
- A cloud-native pipeline for network packet capture data analysis could significantly improve, standardise and speed up the processing time that is needed to run inference queries over the dataset.

C. Aim and goals

This research aims at the creation of a cloud-native framework for capturing, collating and analysing the Internet Background Radiation. The aim will be achieved by the accomplishment of the following objectives:

1) Capturing a sample of the IBR within a Cloud Service Provider. The main result will be one dataset representing all captured data. As a byproduct, the resulting architecture will be made publicly available in the form of a repository containing CloudFormation, Terraform and Ansible descriptors of the Infrastructure-as-Code (IaC) artefact for reproducible, independent peer validation;
2) Operationalisation of a dataset querying interface, focusing on the provisioning of two main features: a) the ability to query the dataset using a pattern-matching mechanism, based on an extract-transform-load workflow that includes AWS-based Athena, Glue and Kinesis services [14]; and b) a dataset exploration and observation window to be built upon one of existing cloud- native analytic stacks, based on AWS technologies such as TimeStream, Redshift or Security Lake;
3) Composition of a set of high-level inference questions to be performed upon the dataset allowing for the characterisation of the IBR both under quantitative and qualitative terms.

IV. Framework Architecture

The data collection phase described as the first objective requires the creation of a cloud platform comprised of key elements for capturing, forwarding, collating and storing the captured traffic resulting in the new dataset. The proposed geographical distribution of sensors is depicted on Figure 1.

Fig. 1. Cloud regions where sensors will be launched

Sensors will be equally distributed across 15 (fifteen) Amazon Web Services regions: one in central Canada; four in the United States of America: (Seattle, California, Virginia and Ohio); one in South America (São Paulo); one in Africa (Cape Town); one in the Pacific (Sydney); three in Asia (Tokyo, Mumbai and Seoul); the final four regions will be in Europe (Stockholm, London, Frankfurt and Paris). The Stockholm region will also host the global bucket to which all sensors will forward captured data every hour. Stockholm region was chosen as the target for data storage due to the fact Sweden has strong GDPR (General Data Protection Regulation) enforcement implemented by *Datainspektionen* [15] within the European Union. The more sensors simultaneously capturing IBR, greater the statistical significance. Budget is the constraint that will determine the size of the sensor fleet that can be launched, multiplied by the expected duration of the capture. The budget that was defined for this project takes into consideration the coverage of expenses detailed on Table I.

TABLE I
Summary of cloud service costs

Region	Max Assets	USD Month	USD Year
Cape Town	100x t3.nano spot EC2	279	3358
Mumbai	100x t3.nano spot EC2	236	2840
Seoul	100x t3.nano spot EC2	256	3076
Sydney	100x t3.nano spot EC2	264	3174
Tokyo	100x t3.nano spot EC2	268	3227
Central Canada	100x t3.nano spot EC2	237	2844
Frankfurt	100x t3.nano spot EC2	250	3005
London	100x t3.nano spot EC2	245	2943
Paris	100x t3.nano spot EC2	245	2943
Stockholm	100x t3.nano spot EC2	222	2673
Sao Paulo	100x t3.nano spot EC2	374	4488
US N. Virginia	100x t3.nano spot EC2	213	2567
US Ohio	100x t3.nano spot EC2	213	2567
US California	100x t3.nano spot EC2	255	3069
US Oregon	100x t3.nano spot EC2	213	2567
Stockholm	Main bucket	424	5141
Stockholm	Bucket data transfer	327	3932
Stockholm	Athena interface	800	9600
Stockholm	Redshift instance	222	2664
		USD	**66,684**

A. Project Timeline

We plan to launch a maximum of 1.500 EC2 T3.micro instances with 100 per region. These will be operated over a period of 12 months, starting July 2023. Storage is expected to grow linearly over the period of observation in accordance to the estimation represented by $Storage = T * 2MB * 24h * 30d * 12m$. In this calculation, T is the number of EC2 instances allocated for the project, each with a single Public IP. 2 MB is the average size of incoming traffic seen in trial experimentation, per day, per month, during 12 months, resulting in a potential dataset of 16 terabytes in size. These in combination with the predicted costs related to the data analysis toolset, result in the maximal need for cloud credits accounting USD 66 684.

The design and implementation of the dataset query interface will be guided and constrained by the following principles, technical instrumentation, and resources:

- An Athena-based interface will be developed to allow interactive, straightforward analysis of the dataset.
- Data exploration will be enabled by the implementation of an analytical pipeline that has not yet been designed, but for which the following components are anticipated at this point:
  - Amazon Redshift: A cloud-native data warehouse solution; or
  - Amazon Time Stream: A cloud-based time-series platform; or
  - AWS Security Lake: A native solution for automating and centralising the management of security-related data, enabling the creation of security-oriented data lakes in accordance with the Cybersecurity Schema Framework [16].

In order to achieve the third research objective, it will be necessary to create a set of ready-made analytical queries to address the most essential quantitative and qualitative research questions. These questions are detailed on Section V, encompassing traffic characterisation of network, transport and application layers. The fourth and final objective is to make the collected dataset available to other researchers. Access to the full dataset should be possible under individual agreements. Time allocation for each objective is summarised on Table II.

TABLE II
PROPOSED PROJECT TIMELINE

	Project Timeline. 24 months. Q=Quarter							
Objective / Year	2023		2024				2025	
	3Q	4Q	1Q	2Q	3Q	4Q	1Q	2Q
O1: Platform Design	o							
O1: IBR capture	o	o	o	o				
O2: Athena Interface			o	o	o			
O2: Analytic Interface					o	o	o	
O3: Analytic queries						o	o	
O4: Report results					o	o	o	o
O4: Publish dataset							o	o

V. EXPECTED RESULTS

It is expected this work will enable the collection and examination of the characterisation of the Internet Background Radiation in relation to the following aspects:

- Traffic volume categorised by protocol within the network layer, transport layer and/or the application layer;
- Service port number distribution for TCP and UDP transport-layer protocols;
- Traffic that relates to port scanning activity;
- Traffic that relates to both scanning and denial-of-service attacks, particularly the ones attributable to *Mirai*, or the more recent *Moobot*; and
- Prevalence of DoS/DDoS activity by target region.

The resulting data analysis pipeline will also provide information to shed light over key qualitative concerns, namely:

- Probing activity generated by well known open source intelligence platforms, particularly Shodan, Censys and Greynoise;
- Malware dissemination activity detected within the dataset, accompanied by detection trends per region; and
- Correlation between the release of CVE alerts and the increase of malicious activity probing and exploiting vulnerable services.

VI. CONCLUSION

This paper documents the research plan towards building a framework enabling the capture, collation and analysis of the Internet Background Radiation focusing on the geographically distributed context of cloud computing environments.

As discussed on related works, the field is of great interest for researchers willing to better understand the properties of the Internet Background Radiation. At the same time, the scarcity of IPv4 addresses poses as a major challenge to the extraction of future capture samples. This work proposes a viable solution that relies on the idle capacity inherent to the business model of cloud service providers in which a dataset will be captured, collated, and analysed.

The use of cloud-native technologies to interact with the data could also represent a valuable contribution. Peers will be able to inspect and reproduce the queries leading to the IBR characterisation, both in quantitative and qualitative terms.

Furthermore, the upcoming availability of infrastructure-as-code artefacts will allow interested peers to conduct new observations and compare results, contributing to a broader understanding of the Internet Background Radiation.

REFERENCES

[1] E. Wustrow, M. Karir, M. Bailey, F. Jahanian, and G. Huston, "Inter- net background radiation revisited," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, IMC '10, (New York, NY, USA), p. 62–74, Association for Computing Machinery (ACM), 11 2010.

[2] B. V. W. Irwin, *A framework for the application of network telescope sensors in a global IP network*. PhD thesis, Rhodes University, 2011.

[3] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson, "Characteristics of internet background radiation," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, IMC '04, (New York, NY, USA), p. 27–40, Association for Computing Machinery (ACM), 10 2004.

[4] B. Irwin, "A baseline study of potentially malicious activity across five network telescopes," in *5th International Conference on Cyber Conflict (CYCON 2013)*, pp. 1–17, Institute of Electrical and Electronics Engineers (IEEE), 2013.

[5] B. Irwin, "Evaluation of mauritian ipv4 address space within internet background radiation data," in *International Conference on Intelligent and Innovative Computing Applications (ICONIC' 2022)*, 2022.

[6] S. D. Chindipha, B. Irwin, and A. Herbert, "Quantifying the accuracy of small subnet-equivalent sampling of IPv4 internet background radiation datasets," in *Proceedings of the South African Institute of Computer Scientists and Information Technologists 2019*, SAICSIT '19, (New York, NY, USA), p. 1–8, Association for Computing Machinery (ACM), 9 2019.

[7] F. Shaikh, E. Bou-Harb, N. Neshenko, A. P. Wright, and N. Ghani, "Internet of malicious things: Correlating active and passive measurements for inferring and characterizing internet-scale unsolicited IoT devices," *IEEE Communications Magazine*, vol. 56, pp. 170–177, Sep. 2018.

[8] A. Guillot, R. Fontugne, P. Winter, P. Merindol, A. King, A. Dainotti, and C. Pelsser, "Chocolatine: Outage detection for internet background radiation," in *Network Traffic Measurement and Analysis Conference (TMA)*, pp. 1–8, Institute of Electrical and Electronics Engineers (IEEE), June 2019.

[9] R. Wazlawick, *Research Methodology for Computer Science*. LTC, 3 ed., 2020.

[10] P. Richter, M. Allman, R. Bush, and V. Paxson, "A primer on ipv4 scarcity," *SIGCOMM Comput. Commun. Rev.*, vol. 45, p. 21–31, apr 2015.

[11] L. Prehn, F. Lichtblau, and A. Feldmann, "When wells run dry: The 2020 IPv4 address market," in *Proceedings of the 16th International Conference on Emerging Networking Experiments and Technologies*, CoNEXT '20, (New York, NY, USA), p. 46–54, Association for Computing Machinery, 2020.

[12] National Institute of Standards and Technology, "Special publication 800-145: The nist definition of cloud computing," Tech. Rep. Recommendations of the National Institute of Standards and Technology, U.S. Department of Commerce, Washington, D.C., 2011.

[13] Amazon Web Services, "Amazon ec2 spot instances run fault-tolerant workloads for up to 90% off," 2023.

[14] H. Rozestraten, "Create real-time clickstream sessions and run analytics with amazon kinesis data analytics, aws glue, and amazon athena," 2019.

[15] Integritetsskyddsmyndigheten, "GDPR ra¨ttslig grund," 2023.

[16] Amazon Web Services, "Amazon security lake (preview)," 2023.