Department of Mathematics and Information Technology: Cloud Platform comparison for malware development

Kamil Janowski

Master’s Thesis in Computer Science

Autumn Term 2018

Department of Mathematics and Information Technology

University of Jyväskylä

ABSTRACT

Janowski, Kamil. 2018. Master’s thesis in computer science. University of Jyväskylä. Department of Mathematics and Information Technology (MIT).

The cloud platforms such as AWS, Google Cloud or Azure are designed to cover most popular cases in terms of web development. They provide services that make it easy to create a new user based on his email address, provide tools for inter-service communication, tools to manage the access rights of different users. Malware development however is more of a corner case, where the client application running on the victim’s machine does not have an email address or a google account to authenticate itself and it does not run directly in the cloud, what can make it more difficult to manage the appropriate access rights. Also, the potential attacker may not want to write his own self-contained service, since, especially when managing a large number of clients, it might be much cheaper to run the backend serverlessly.

//TODO: describe research methods (exploratory?)

The following paper explores possible malware backend architectures for different cloud platforms, aiming to optimise the performance, minimize the development time while keeping the code easy to maintain and to minimize the execution cost.

//TODO: write the final conclusion

Keywords: malware, development, cloud, CnC, backend, serverless

CONTENTS

[1 INTRODUCTION 5](#_Toc521489869)

[1.1 Research Problem 6](#_Toc521489870)

[1.2 Research Objective 6](#_Toc521489871)

[1.3 Research Question 7](#_Toc521489872)

[1.4 Key Definition 7](#_Toc521489873)

[1.4.1 Hacker 7](#_Toc521489874)

[1.4.2 Botnet 7](#_Toc521489875)

[1.4.3 Bot 7](#_Toc521489876)

[1.4.4 Serverless computing 7](#_Toc521489877)

[1.4.5 Cloud Computing 8](#_Toc521489878)

[1.4.6 Malware 8](#_Toc521489879)

[1.4.7 CnC server 8](#_Toc521489880)

[1.4.8 DDoS attack 8](#_Toc521489881)

[1.5 Structure of the thesis 8](#_Toc521489882)

[2 Theoretical Background 9](#_Toc521489883)

[2.1 Common botnet architectures 9](#_Toc521489884)

[2.1.1 Centralised architecture 9](#_Toc521489885)

[2.1.2 Peer to Peer (P2P) Architecture 10](#_Toc521489886)

[2.1.3 Hybrid architecture 10](#_Toc521489887)

[2.2 Common botnet use-cases 11](#_Toc521489888)

[2.3 Command delivery methods 11](#_Toc521489889)

[2.3.1 HTTP notifications 12](#_Toc521489890)

[2.3.2 WebSocket notifications 12](#_Toc521489891)

[2.3.3 IRC notifications 12](#_Toc521489892)

[2.3.4 MQTT notifications 13](#_Toc521489893)

[3 Methodology 13](#_Toc521489894)

[3.1 Purpose of the study 13](#_Toc521489895)

[3.2 Research approach 14](#_Toc521489896)

[3.3 Research method 14](#_Toc521489897)

[3.4 Data collection 15](#_Toc521489898)

[4 Findings – case study on 3 platforms 16](#_Toc521489899)

[4.1 Standalone CnC server 16](#_Toc521489900)

[4.1.1 Design 16](#_Toc521489901)

[5 Conclusion 17](#_Toc521489902)

[6 Discussion 17](#_Toc521489903)

# INTRODUCTION

The popularity of computing clouds have increased drastically during the recent years. It is perfectly understandable, taken into account that renting the infrastructure from a cloud provider tends to be significantly cheaper than maintaining it inside the company. Things like the rental of the server room, electricity consumed by the servers, cooling of the server room and salaries of people responsible for the maintenance of the servers generate unnecessary overhead in terms of costs of maintenance, which can be drastically reduced when switching to the cloud, while in the same time providing higher availability and better monitoring of the hosted services. Furthermore the cloud providers constantly introduce new solutions allowing to reduce the maintenance costs even further. As we can read in “Serverless Computing: Economic and Architectural Impact” by Gajko Adzic and Robert Chatley (2017, p. 884):

Amazon Web Services unveiled their ‘Lambda’ platform in late 2014. Since then, each of the major cloud computing infrastructure providers has released services supporting a similar style of deployment and operation, where rather than deploying and running monolithic services, or dedicated virtual machines, users are able to deploy individual functions, and pay only for the time that their code is actually executing. These technologies are gathered together under the marketing term ‘serverless’ and the providers suggest that they have the potential to significantly change how client/server applications are designed, developed and operated.

It is important to note however that those technologies are not only available to big corporations trying to lower their cost of server maintenance, but also to hobby software developers and hackers.

A successful attacker may have thousands of devices under his control. In order to control such a large number of devices remotely a highly scalable Command-and-Control (CnC) server is required. Scaling up the virtual machines (VM) however can be costly, while having only a small number of administrators leads to a situation where most of the resources assigned to those VMs are seriously underutilized. While all the remote malware subscribes to the push notification service, it mostly just waits for a command to be generated by an administrator. Effectively, while our CnC server has to be scalable in order to maintain the connection to numerous clients, it requires fairly low computing power until an administrator decides to generate certain load. This suggests that the serverless approach could be applied in this case, what could potentially not only save the attacker a lot of money, but also make such a large scale attack possible in the first place.

## Research Problem

There are many various cloud providers out there. While they all provide services allowing to easily and quickly build secure web applications, the problem of building a CnC server is more of a corner case, that is not necessarily properly addressed by certain clouds. This might yield it impossible to implement such an application in a serverless manner at all, or require to make some compromises and implement workarounds for services that work in a different manner than desired.

## Research Objective

The main objective of the research is to find a way to use the cloud as a CnC server without implementing any application that requires a constantly running server in a Virtual Machine, as those are the main cost generators of the web applications. For this reason we are going to investigate the serverless solutions provided by various cloud providers as well as other services that come with specific clouds that could potentially allow us to set up the communication between the backend and the client application, enable the file transfer, make it easy to manage the access rights of different clients as well as enable the client management in as a whole. We are also going to take a closer look at how the continuous deployment can be solved in various cloud systems.

Each of the approaches will be backed up by a small Prove of Concept (POC) if possible at all. In order to optimise the development time and ensure multi-platform and multi-cloud support of at least parts of our code, all solutions will be implemented with Node.js.

## Research Question

When focusing on various cloud platforms, such as Amazon Web Services (AWS), Google Cloud Platform (GCP) and Azure the approach to the problem might be completely different and the cost of execution may different significantly as well. The question in this case is, which one of the platforms is the best suited and the cheapest to run our CnC server.

## Key Definition

### Hacker

Hacker is a malicious attacker attempting to access resources of a remote machine. In this thesis the term “hacker” will be used to describe the administrator of the CnC server and in the same time the administrator of the botnet.

### Botnet

A botnet is a network of private computers infected with malicious software and controlled as a group without the owners' knowledge, e.g. to send spam.

### Bot

A bot in this case is a single client application executing (and in some architectures issuing) the commands on the infected device.

### Serverless computing

The “serverless” computing is a marketing term that relates to developing single functions, rather than a large monolithic application and then being charged only for the actual execution time of the function, rather than for the constantly running server that technically is still there, but is hidden from the service user. The concept was originally introduced by Amazon in their AWS cloud in 2014 under the name of Lambda. Since then all major cloud providers introduced various equivalents in their solutions. As many instances of lambda can be triggered in parallel, this solutions is not only cheaper to execute, but also potentially infinitely scalable. This is why it’s commonly used for a wide range of applications, starting with REST API call processing and ending with Big Data event handling.

### Cloud Computing

As Amazon defines it on <https://aws.amazon.com/what-is-cloud-computing> (24-06-2018):

Cloud computing is the on-demand delivery of compute power, database storage, applications, and other IT resources through a cloud services platform via the internet with pay-as-you-go pricing.

### Malware

Malware, or malicious software, is any program or file that is harmful to a computer user. Malware includes computer viruses, worms, Trojan horses and spyware. These malicious programs can perform a variety of functions, including stealing, encrypting or deleting sensitive data, altering or hijacking core computing functions and monitoring users' computer activity without their permission.

### CnC server

In “Survey on botnet: its architecture, detection, prevention and migration” by Ihsan Ullah et al. (2013) the CnC servers are defined as centralised servers allowing the malicious attacker to remotely control a number of clients applications that connect to it.

### DDoS attack

DDoS stands for Distributed Denial of Service. It’s one of the common use-cases of a botnet, where a number of bots are instructed to simultaneously send requests to a specific server, yielding the server inaccessible for other users.

## Structure of the thesis

//TODO: Do it last. Explain what the reader can expect to read in each chapter.

# Theoretical Background

## Common botnet architectures

As we can read from “Survey on botnet: its architecture, detection, prevention and migration” by Ihsan Ullah et al. (2013, p. 661-662), as well as “Botnet Communication Patterns” by Gernot Vormayr et al. (2017, p. 2772) there’s a number of different architectures that can be developed depending on the attacker’s needs.

### Centralised architecture

The architecture assumes that there’s one CnC server that all the clients can connect to. It tends to use either Internet Relay Chat (IRC) or HTTP as the communication protocol. This solution tends to be the most commonly seen due to the ease of implementation as well as high efficiency. The main drawback of the approach is that it is fairly easy to detect. Each of the clients of the botnet needs to have a hard-coded address of the server that it is going to communicate with. Effectively simply editing the byte code of the application (or decompiling it, if possible) allow you to quickly read the address of the CnC server and then block all the traffic to it. The address can also be seen through network sniffing. This problem however can be mitigated through the use of Domain Generator Algorithms (DGA).

DGAs generate different domain names based on a changing input. For instance, a different domain could be used based on the current time. This then requires all clients to have synchronized time down to one hour. While relying on the system time might not necessarily be a good idea, as the system time largely depends on the user-specified settings, it can be easily achieved by polling popular websites that contain such information.

### Peer to Peer (P2P) Architecture

The approach allows to hide most of the network traffic by introducing the supervisor-bot, who becomes responsible for delivering the command to other clients, who later on can forward the command even further. While the source of the command becomes fairly difficult to detect in this case, the actual delivery as well as the delivery of the result takes significantly more time than in the centralised architecture. This makes such botnet difficult for the attacker to manage. Also, it is important to note that the architecture is prone to the Sybil attack, where the attacker subverts the reputation system of a P2P network by creating a large number of pseudonymous identities, using them to gain a disproportionately large influence.

In a fully meshed botnet every client is linked to every client. This way it is possible to reduce the latency as well as ensure that the removal of any number of bots does not disrupt the communication. This solution however is not scalable due to the number of required connections in larger botnet. Additionally, the larger number of connections increases the visibility of the botnet. Also adding or removing a single client generates a significant network traffic as all other clients have to register the information about the new bot.

The topology unfortunately is difficult to implement due to the challenges of finding the initial peers and reliably distributing commands to every bot.

The list of peers can be hard-coded directly in the executable or provided by a cache server. The first solution however can work only in a very targeted attack and should the botnet be detected, the list can be easily extracted from the code.

In the second case, the server is visible to the public internet and that brings back all the issues related to the centralised architecture.

Finally, depending on the NAT configuration, not every computer has direct access to the internet making it difficult to access from external network.

### Hybrid architecture

Hybrid architecture combines both centralised architecture and the P2P one. Instead of bots connecting directly to the CnC server, an additional proxy layer consisting of bots connected in a P2P topology is added. Determining whether a certain bot should behave only as proxy or P2P accessed worker can be done based on the connectivity properties (such as when some of the infected devices don’t have the direct access to the CnC server). In order to lower the probability of detection of the CnC server, additional layers of P2P connection can be added, although that comes with the cost of increased latency.

## Common botnet use-cases

According to the definition of botnet provided by Norton on https://us.norton.com/internetsecurity-malware-what-is-a-botnet.html (07-07-2018), a botnet can be used for purposes like:

* Executing a DDoS attacks
* Emailing spam to millions of internet users
* Generating fake Internet traffic on a third-party website for financial gain.
* Replacing banner ads in your web browser specifically targeted at you.
* Pop-ups ads designed to get you to pay for the removal of the botnet through a phony anti-spyware package.

## Command delivery methods

There’s a number of different ways that a command can be delivered by CnC server to a bot. As already mentioned in the introduction, HTTP and IRC protocols are the most commonly used for this purpose, however those are not our only options. As mentioned by Inmaculada Ayala et al. in “An empirical study of power consumption of Web-based communications in mobile phones” (2017) WebSockets are also a common option for the message delivery both in case of mobile applications as well as websites (and effectively botnet client). Also what is available in most clouds are IoT services that can enable the communication with a remote client over the MQTT protocol. Let’s take a closer look at each one of these approaches now

### HTTP notifications

As mentioned by Inmaculada Ayala et al. the command delivery over the HTTP protocol can be handled in two different ways: polling and long polling.

Inmaculada Ayala et al. defines the polling approach in the following way:

The polling mechanism is the simplest way to receive asynchronous data. The client polls the server periodically (polling interval) for new content by sending HTTP requests, allowing the server to respond with an HTTP response if new data is available. Each request attempts to pull any available data. If no data is available, the server returns an empty response and the client waits for some time (polling interval) before sending another (poll) HTTP request.

Whereas the long polling is defined as follows:

In order to alleviate client continuous polling, there exist different web models in which a longheld HTTP request allows a web server to push data to a browser only when new data is available. One of the most common server push mechanisms is HTTP “Long Polling”, in which the server “holds open” (not immediately reply to) each HTTP request, responding only when there is new data to deliver. Then, there is always a pending request to which the server can reply for the purpose of sending data as it is available, thereby minimizing the latency in message delivery, and the use of processing/network resources.

### WebSocket notifications

Inmaculada Ayala et al. describes also the WebSocket-based approach to the problem. With WebSocket protocol it is possible for the client to create a full-duplex persistent TCP connection to the server.

Based on this connection, the Web server is able to actively send data to the client whenever it is available. Prior to data/message exchange, the WebSocket protocol requires an initial handshake and the message exchange. The initial handshake uses the HTTPUpgrade-request, which allows to switch from the HTTP to the WebSocket protocol. The message exchange is executed in form of frames, which contain either text or binary data.

### IRC notifications

IRC protocol is a simple plain text protocol operating over a persistent TCP connection. Effectively, similarly to the WebSocket approach, the message is delivered to the client as soon as it is available on the server.

### MQTT notifications

As Konglong Tang et al. Define the MQTT protocol in ” Design and Implementation of Push Notification System Based on the MQTT Protocol” (2013), it’s a protocol originally designed and developed by IBM, that allows the delivery of push messages. MQTT can work in one of three modes of message delivery:

* At most once – the actual delivery depends only on the TCP connection and as a result some messages can be lost on the way
* At least once – the server ensures that the message is delivered, but duplicates can happen
* Only once – the server ensures that the message is delivered exactly one time

It is a particularly interesting protocol in our case, as the MQTT push notification service is provided by every major cloud through IoT services.

# Methodology

This part describes the methodology used for the study. It will describe the full approach of conducting the study, the data collection methods, the research methods and the purpose of the study.

## Purpose of the study

The purpose is exploratory. The exploratory research, as the name already implies, aims to explore the research questions rather than provide the ultimate solution to the problem. This is important in this case, as there is are hundreds of ways to implement a malware. It simply wouldn’t be feasible to go through them all to find the one best solution.

The paper will compare various cloud platforms, services they provide and the cost of their usage to find various possible architectures for our Command and Control server and effectively the malware communicating with it. In the end we will also compare the cost of maintenance of different architectural approaches. After all the very reason why designing a Cloud Platform-specific CnC server makes sense is because it can drastically lower the execution costs.

It is also worth mentioning here that malware development is not among any of the target applications of those cloud platforms. While they provide a number of very convenient features useful for building robust web applications, management of IoT devices and AI data processing, services that might turn out to be essential to achieve our goals might simply not be in place. Should that happen, the only way to execute our solution is to spawn a virtual machine inside that cloud, running a standalone CnC server, what defeats the purpose of using that specific cloud.

## Research approach

In the paper we will use the deductive approach. When utilizing the deductive research approach we want to start with a hypothesis and the through data collection we want to build a proven theory. In this case our hypothesis is that it is possible to build a CnC solution using only the serverless technologies provided by various cloud platforms and therefore minimize the cost of execution of the CnC application while keeping it scalable, what is necessary to manage a large number of clients.

## Research method

We will use the qualitative case study research method. The qualitative case study method is used to collect the data through in-depth investigation of multiple cases within one context.

The study will focus on three cases of three different clouds:

* Amazon Web Services (AWS)
* Google Cloud Platform (GCP)
* Microsoft Azure

These are currently the most commonly used cloud platforms. Each one of these clouds support serverless computing in one way or another, whether those are lambdas, Google App Engine or other form of serverless logic executor and after all those are the services this study puts a lot of emphasis on. They all also provide various ways of message delivery through various custom push notification services to HTTP and MQTT-based IoT services. Some of them also provide other services that can allow us to make our malware more effective (like for instance the P2P services).

//TODO: find a paper describing cloud market shares to prove that these are actually the 3 biggest clouds

## Data collection

In exploratory case studies data often is collected through questionnaires, interviews and experiments. While the questionnaires and interviews make very little sense in terms of technology-related studies, the experiments do.

In the study we will collect the data through:

* Already existing research papers, official cloud documentations and blogs on related topics. Especially the blogs may prove to be very useful as most framework and technology providers as well as data security companies tend to describe on their blogs various approaches to various problems related to architecture, implementation and security threats.
* Empirical implementation, to validate that the approach is actually feasible. In the software development it is a very common case that a certain technology appears to solve the proposed problem, whereas during the implementation of the solution it turns out that the selected technology imposes certain limitations, yielding it inapplicable for the specific problem. Effectively the only way of ensuring that the solutions we will propose in this study are valid is to implement the proof of concept for each one of them.

Data collection through documentary analysis and empiric implementation.

Describe the study framework. (check the link to that research from Nguyen, 7.2.3 – Study Framework)

# Findings – case study on 3 platforms

## Standalone CnC server

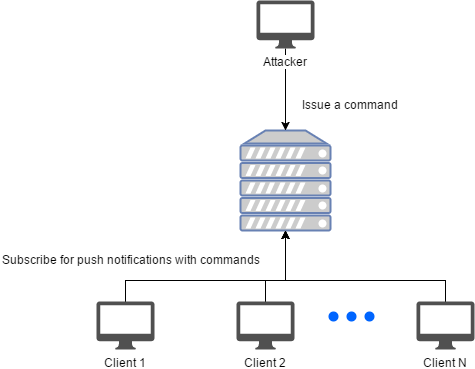
In order to better understand the complexity of CnC applications as well as evaluate the cost of their execution, let’s first analyse the standalone approach where we try to create our own CnC application running on a server. Let us however not focus on any extreme examples just to prove the point on the thesis. Technically we could create a Java application running on a Tomcat server, but according to Oracle documentation (https://docs.oracle.com/cd/E13169\_01/ales/docs22/installadmin/prepare.html, 07.08.2018) we would need 512 MB of memory just to run the server and then there are memory requirements of our application on top of that. For this reason we’re going to build a small application in Node.js instead. One that can integrate the whole server in it, without relying on a third party one.

### Design

While the list of common use cases of botnet is fairly long, most of them can be handled in a similar way:

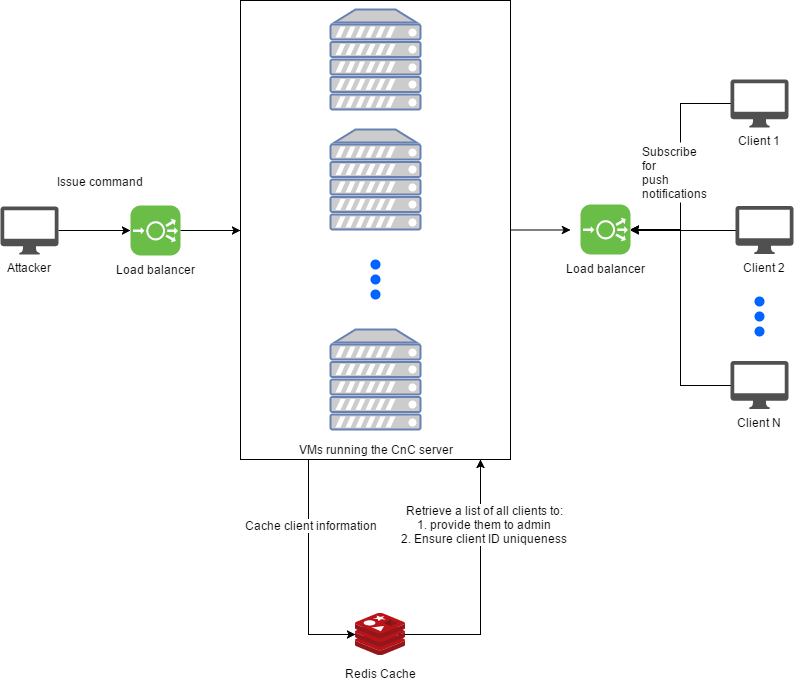
* + - 1. Client subscribes for the push notifications from the server over HTTP or IRC protocol (as already mentioned before)
      2. A request is issued by the administrator to the server
      3. The server dispatches appropriate commands to the client

Effectively the most trivial CnC application could be essentially just one server with all the clients connecting to it and waiting for the attacker to issue a command.



//todo look up how to insert figures properly

This approach however has a drawback – there’s only so many clients that can connect to the server in the same time. They all have to maintain an open TCP connection in order to be able to react to the command as soon as possible and once some data has to be transferred between the client and the server, there’s also a limit imposed by the connection speed of the virtual machine running our server application. We can obviously always configure the virtual machine giving it higher bandwidth, but then we would end up paying for it at all times, even when we don’t really use it. The same goes for all the other resources required to run the application. For this reason it seems more cost effective to have a number of VMs with lower amount of resources, that can be spawned automatically by a load balancer when they’re needed. This however introduces a difficulty. If there are multiple servers hidden behind a load balancer, then they need to be able to exchange the information about the connected clients between each other. Luckily there are multiple caching services out there that can be used for this purpose. One of the most popular ones and provided out of the box by most major cloud providers is Redis. Having that in mind, let’s update the application design.



This way we limited the cost of VMs required to run our CnC application, however in the same time we introduced the necessity of using the load balancer and the Redis cache, which do not come for free either.

In the next section let’s try to evaluate how much resources are needed in both approaches in order to calculate the approximate cost of execution of the server-based CnC application.

//todo: sequence diagram of what happens there

### Resources and cost

In order to evaluate the resources actually needed to execute I wrote a minimalistic proof of concept in Node.js that can work either with or without the Redis support.

//TODO: can I somehow reference the code written? Like… “you can see it in appendinx 1” and then paste the entire code to the appendix? o\_O – pinged Fergal Carolan and O. Khriyenko

In order to evaluate the required resources, I will simulate 10000 client connections and measure the memory consumption and the processor usage of the CnC application process.

Market overview of cloud computing / 3 platforms. All the info about all 3 platforms. (in separate sub-points)

Comparison of the platforms.

# Conclusion

# Discussion

references

Journal articles with doi:

Ihsan Ullah, Naveed Khan, Hatim A. Aboalsamh (2013). Survey on botnet: Its architecture, detection, prevention and mitigation. *10th IEEE International conference on networking, sensing and control (ICNSC),* 660-665.

doi: 10.1109/ICNSC.2013.6548817

Vormayr, G., Zseby, T., & Fabini, J. (2017). Botnet communication patterns. IEEE *Communications surveys and tutorials*, 19(4), 2768-2796.

doi: 10.1109/COMST.2017.2749442

Gajko Adzic, Robert Chatley (2017). Serverless computing: economic and architectural impact. *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering,* 884-889.

doi: 10.1145/3106237.3117767

Ayala, I., Amor, M., Fuentes, L., & Muñoz, D. (2017, November). An empirical study of power consumption of Web-based communications in mobile phones. In Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence & Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2017 IEEE 15th Intl (pp. 861-866). IEEE.

doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.144

Journal articles without doi:

Tang, K., Wang, Y., Liu, H., Sheng, Y., Wang, X., & Wei, Z. (2013, September). Design and implementation of push notification system based on the MQTT protocol. In International Conference on Information Science and Computer Applications (ISCA 2013) (pp. 116-119).

appendices