



**University of
Zurich^{UZH}**

*Burkhard Stiller, Muriel Franco, Christian Killer, Sina Rafati,
Bruno Bastos Rodrigues, Eder John Scheid, Eryk Schiller (Eds.)*

Internet Economics XIII

TECHNICAL REPORT – No. IFI-2020.01

January 2020

University of Zurich
Department of Informatics (IFI)
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland

Introduction

The Department of Informatics (IFI) of the University of Zürich, Switzerland works on research and teaching in the area of communication systems. One of the driving topics in applying communications technology is addressing investigations of their use and application under economic constraints and technical optimization measures. Therefore, during the autumn term HS 2019 a new instance of the Internet Economics seminar has been prepared and students as well as supervisors worked on this topic.

Even today, Internet Economics are run rarely as a teaching unit. This observation seems to be a little in contrast to the fact that research on Internet Economics has been established as an important area in the center of technology and economics on networked environments. After some careful investigations it can be found that during the last ten years, the underlying communication technology applied for the Internet and the way electronic business transactions are performed on top of the network have changed. Although, a variety of support functionality has been developed for the Internet case, the core functionality of delivering data, bits, and bytes remained unchanged. Nevertheless, changes and updates occur with respect to the use, the application area, and the technology itself. Therefore, another review of a selected number of topics has been undertaken.

Content

This new edition of the seminar entitled “Internet Economics XIII” discusses a number of selected topics in the area of Internet Economics. The first talk “Marketplaces for Networking: a Study of X-as-a-Service Solutions” identifies potentials for service provision in cloud and possible marketplaces for networking. Talk two “An Economic Analysis of the Migration of Geographical Information Systems (GIS) to the Cloud” introduces the role of clouds on immigrating geo-data. Talk three “Economic Assessment of Distributed Denial-of-Service (DDoS) Attacks” introduces economic aspects of distributed attacks such as DDoS. Talk four on “An Overview of Cyber Insurance Models” presents the state of the art policies offered by insurance companies that are supporting Cyber cases such as data loss or attacks. Talk five entitled “An Economic Analysis of Cloud Storage Providers and Private Cloud/NAS Systems” highlights the main aspects data storage. Talk six entitled “Data Collection and Wireless Communication in IoT Using Economic Analysis and Pricing Models” overviews some of best solutions developed in data collections provided for the fifth generation of wireless communications. Talk seven entitled “Survey and Analysis of Existing Cloud SLA Compensation Processes and Values” presents an introduction to the topic of in the last years Cloud SLA policies and analyses their economic aspects. Talk eight entitled “Commercializing Blockchain: Transformation and Emergence of Web 3.0 Business Models” discusses the potentials of Web 3.0 and the

role of blockchain based applications in providing new applications and their economic impact in Web 3.0. Talk nine entitled “Economics of Fifth Generation Cellular Networks” exploits the role of cellular networks in 5G and the international economic competition over reaching out and establishing this technology. Talk ten entitled “The Market Behind Edge Computing: From Content Caching to Autonomous Driving” introduces several use cases based on edge computing with overall economic analysis of them. Talk eleven on “Botnet Economics and Business Models” discusses the role of botnets and their impact on Cyber security from an economic perspective. Finally talk twelve on “The Renaissance of DAOs: Recurring Trends Toward Decentralized Autonomous Organizations” concludes the IntEco seminar in HS 2019 presenting the decentralized applications mainly provided by blockchains which enabled smart contracts and their economic potential.

Seminar Operation

Based on well-developed experiences of former seminars, held in different academic environments, all interested students worked on an initially offered set of papers and book chapters. Those relate to the topic titles as presented in the Table of Content below. They prepared a written essay as a clearly focused presentation, an evaluation, and a summary of those topics. Each of these essays is included in this technical report as a separate section and allows for an overview on important areas of concern, sometimes business models in operation, and problems encountered.

In addition, every group of students prepared a slide presentation of approximately 45 minutes to present his findings and summaries to the audience of students attending the seminar and other interested students, research assistants, and professors. Following a general question and answer phase, a student-lead discussion debated open issues and critical statements with the audience.

Local IFI support for preparing talks, reports, and their preparation by students had been granted by Muriel Franco, Sina Rafati, Bruno Bastos Rodrigues, Christian Killer, Eder John Scheid, Eryk Schiller, and Burkhard Stiller (Eds.). In particular, many thanks are addressed to Sina Rafati for his strong commitment on getting this technical report ready and quickly published. A larger number of pre-presentation discussions have provided valuable insights in the emerging and moving field of Internet Economics, both for all students and supervisors. Many thanks to all people contributing to the success of this event, which has happened in a lively group of highly motivated and technically qualified students and people.

Zürich, January 2020

Contents

1	Marketplaces for Networking: a Study of X-as-a-Service Solutions	7
	<i>Lawand Muhamad, Nicolas Kohler, Tim Schluchter</i>	
2	An Economic Analysis of the Migration of Geographical Information Systems (GIS) to the Cloud.	29
	<i>Stéphanie Wismer, Jan Weber, Silvan Caduff, Sophie Sturzenegger</i>	
3	Economic Assessment of Distributed Denial-of-Service (DDoS) Attacks	63
	<i>Adrian Iten, Artemis Kardara, Vasiliki Arpatzoglou, Timo Schenk, Noah Berni</i>	
4	An Overview of Cyber Insurance Models	99
	<i>Christian Birchler, Michael Nadig, Sandro Padovan and Louis Preisig</i>	
5	An Economic Analysis of Cloud Storage Providers and Private Cloud/-NAS Systems	123
	<i>Clara-Maria Barth</i>	
6	Data Collection and Wireless Communication in Internet of Things (IoT) Using Economic Analysis and Pricing Models	149
	<i>Matej Jakovljevic, Jeremy Kubrak, Dylan Puser, Marc Zwimpfer</i>	
7	Survey and Analysis of Existing Cloud SLA Compensation Processes and Values	183
	<i>Fan Feng, Ruike Wang, Yue Ding and Yuang Cheng</i>	
8	Commercializing Blockchain: Transformation and Emergence of Web 3.0 Business Models	213
	<i>Clive C. Javara, Naël M. H. Prélaz, Syed S. Ahmed, Alphonse Mariyagnanaseelan</i>	
9	Economics of Fifth Generation Cellular Networks	33
	<i>Rabiya Abdullah, Annesha Bhoumik, Dominik Jurilj, Manpreet Singh Sohal</i>	
10	The Economics of Multi-Access Edge Computing	67
	<i>Haishan Fei, Andreas Knecht, Dmytro Polyanskyy</i>	
11	Botnet Economics and Business Models	89
	<i>Famos Tobias, Mannhart Thomas, Tham David, Waltert Gian</i>	
12	The Renaissance of DAOs: Recurring Trends Toward Decentralized Autonomous Organizations	123
	<i>Francesca Monzeglio, Raphael Beckmann, Benjamin Jeffrey and Roberto Baumann</i>	

Chapter 1

Marketplaces for Networking: a Study of X-as-a-Service Solutions

Lawand Muhamad, Nicolas Kohler, Tim Schluchter

Network Function Virtualization (NFV) describes the virtualization of network infrastructures. Nowadays, these functions are made available through marketplaces where developers can publish and customers can acquire network functions without any additional hardware installations. Marketplaces play a crucial role for developers to be able to publish their functions in a store where customers can browse through a catalog of different functions. As there are more and more solutions emerging around this topic, this paper aims to give an overview of different examples of marketplaces representing different approaches to achieve Network-Functions-as-a-Service (NFaaS). As these functions can be acquired as-a-Service, they eliminate the need to have additional hardware installations and maintenance on the customer's side. We see that these solutions greatly differ in their business-models, their quality assurance, their target market as well as in their security aspects. We find that while generally, NFV comes with a lot of benefits such as cost savings and dynamic up- and down-scaling capabilities, they raise questions to new challenges such as dependency on and trust in providers as well as security and control of operations in general.

Contents

1.1	Introduction	9
1.2	Background	9
1.2.1	Cloud Computing	9
1.2.2	Networking	11
1.3	Marketplaces for Networking	13
1.3.1	Cisco Marketplace	14
1.3.2	AWS Marketplace	15
1.3.3	T-NOVA	15
1.3.4	FENDE	18
1.3.5	BUNKER	19
1.4	Comparison of Current Marketplaces	23
1.5	Challenges and Opportunities regarding Marketplaces for Networking	23
1.6	Discussion and Conclusion	24

1.1 Introduction

Marketplaces as a place to buy or sell goods have existed since ancient times. While the greek had to be physically present at the agora to conduct trade, this is no longer necessary in todays online marketplaces [22]. With the advent of online storefronts (e.g. AppStore or GooglePlay), a potential customer can browse applications developed by others at ease, no matter the location [17]. Such online marketplaces come in various forms and are not restricted to pure software. For instance, infrastructure capabilities are increasingly managed through the cloud, thanks to the lower costs associated with this virtualization [14]. However virtualization networking capabilities are only partially part of these Infrastructure-as-a-Service(IaaS) solutions [2]. With the European Telecommunication Standard Institute (ETSI) publishing a framework for Network Function Virtualization in 2012 [26], the idea of providing Network Functions as a Service seems to have gained traction.

This study takes a closer look at marketplaces focused on providing networking services or solutions towards full network softwarization. To ensure a common understanding the concepts of cloud computing are introduced first, followed by elaborating on networking concepts. Given the amount of different marketplaces for networking it is expected that these marketplaces differ regarding their services and solutions. Hence the aim of this study is to compare five hand-picked marketplaces based on their approach. To get insights into what is currently already available and what the future might look like it was decided to choose two established providers and three promising prototypes. Based on this comparison, potential opportunities of such networking marketplaces as well as open challenges will be discussed.

There definitely has been a lot of research in the cloud computing space already. The explanation provided by National Institute of Standards and Technology (NIST) [2] has been identified as a go to definition, hence this paper closely elaborates on it, while providing some additional comments from other studies. To introduce the networking concepts the study relies on various sources to provide an overview of the space without diving too much into the technical details. Although certain papers have tried to compare certain marketplaces for networking and highlight challenges, they usually did not make it the focus of their study. Because of this and the novelty of networking marketplaces there is still a lack of research conducted so far. Hence this study tries to take a first step in filling this gap.

1.2 Background

1.2.1 Cloud Computing

The concept of cloud computing proposed by John McCarthy dates back to 1957 [14]. This concept, which today is understood as the ability to provide services through the internet [14], grows in demand rapidly as more and more applications rely on the cloud. According to IBM three out of four non-cloud apps will move to the cloud within the next few years [15]. Such a shift results in an increase in cloud computing investments by companies. IDC [16] shows that the spending for cloud computing grew approximately 4.5 times the rate of overall IT investments since 2009. Looking forward it is expected that this growth accelerates to six times the rate of IT spending overall from 2015 to 2020 [16].

With the growing demand for cloud computing services and its marketplace solutions it is important to understand its concepts. While there have been many different approaches to define cloud computing we will refer to the commonly cited definition by the US based National Institute of Standards and Technology (NIST) [2].

NIST mentions and defines five characteristics which elaborate on the definition of cloud computing namely on-demand self service, broad network access, resource pooling, rapid elasticity and measured service [2]. Referring to NIST [2] they can be explained as follows: **On-demand self service** indicates the ability of consumers to allocate computing capabilities themselves without relying on the help of a service provider while a **broad network access** enables consumers to access these capabilities with standardized means. Another feature is called **resource pooling**. It refers to the fact that the providers resources are combined together to handle multiple users simultaneously. Depending on consumers demand these resources can then be dynamically assigned or released, which is described as **rapid elasticity**. To enable these characteristics it is important to be able to provide a **measured service**. This allows for the monitoring and controlling of the affected resources [2].

1.2.1.1 Deployment Models

When talking about cloud deployment models one is usually concerned about where the infrastructure on which the resources reside is located. NIST differentiates between four different deployment models, Public Cloud, Private Cloud, Community Cloud and Hybrid Cloud [2].

- **Public cloud:** Using a public cloud indicates that the infrastructure is open to the public while residing on the cloud provider side [2]. This approach ensures on demand scalability and technical support from the provider but it might raise data security and privacy concerns since it is not necessarily clear where the data is stored [23].
- **Private cloud:** In contrast to the public cloud, the infrastructure is exclusively used by one organization [2]. A private cloud usually allows more control over the deployment and it can tackle security issues, since only users within the company can access the cloud [24].
- **Community cloud:** In the case of a community cloud the infrastructure is used by a certain group of consumers that share similar interests. Like the private cloud it may be located on or off premise [2]. While it might be cheaper than using a full private cloud a community cloud implies sharing the bandwidth with the other companies [23].
- **Hybrid cloud:** The use of a hybrid cloud combines features of the distinct cloud models as needed [2].

1.2.1.2 Service Models

NIST defines three basic cloud service models: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), which differ from each other in regards to the degree of management the consumer can conduct [2].

- **Infrastructure as a Service (IaaS):** In the case of IaaS the consumer has full control over the deployed software and the management of operating systems, while not managing the underlying cloud infrastructure itself [2]. The consumer can also potentially control networking components like load balancers or firewalls [2].
- **Platform as a Service (PaaS):** PaaS goes further than IaaS. Not only the underlying cloud infrastructure is taken care of by the provider but also the management

of the operating systems, networking and other computing capabilities [2]. The customer can deploy self made or acquired applications, to do so, tools supported by the provider can be accessed [2].

- **Software as a Service (SaaS):** Here the customer can run the providers software on the cloud infrastructure and access these application either be through a client or program interface [2]. The consumer has no control over the underlying cloud infrastructure and capabilities like networking or storage management are not accessible [2]. This approach reduces costs by eliminating the complexity of deploying hardware and employing specialized staff [23].

1.2.2 Networking

A network is a group of two or more devices that can communicate [4]. In practice, a network is comprised of a number of different computer systems connected by physical or wireless connections. The scale of a network can range from a single computer sharing out peripherals to massive data centers or to the Internet itself. Regardless of scope, all networks allow computers or individuals to share information and resources [19]. The term of a computer network is most commonly used in organizations, where they are being used so that data can safely be transferred and accessed within the organizations borders. However, many organizations find the cost of a dedicated standalone appliance per-feature prohibitive, slow and very inflexible to install and to maintain in the long term, since it requires propriety hardware that needs to be setup and maintained by highly qualified specialists. Thus, new network architectures and business models must be found in order to solve these challenges. In this section, we are presenting new networking architectures such as NFV and SDN and discussing the benefits they provide.

1.2.2.1 Network Functions Virtualization (NFV)

NFV was specified by the ETSI group in 2012 as the concept of deploying independent network components as pure software elements running in standardized virtualized network infrastructure [18]. It is a new way to create, distribute and operate network services and was designed to consolidate and deliver the networking components needed to support an infrastructure, that is totally independent from the underlying hardware, so that network functions can run on off-the-shelf hardware. These functions such as firewalls or intrusion prevention systems become virtual network functions (VNF). The main benefits of moving from the classic network appliance approach to the the network virtualisation approach as seen in Figure 1.1 are:

- It reduces capital expenditure by reducing the need to purchase purpose-built hardware and using pay-as-you-grow models to eliminate wasteful over-provisioning.
- It reduces operational expenditure by reducing cooling, power and space requirements of hardware equipment.
- It simplifies the roll-out and management of network services. It lowers the risks associated with rolling out new services, allowing providers to easily trial and evolve services.
- Delivers agility and flexibility to quickly scale services up or down to address changing demands.

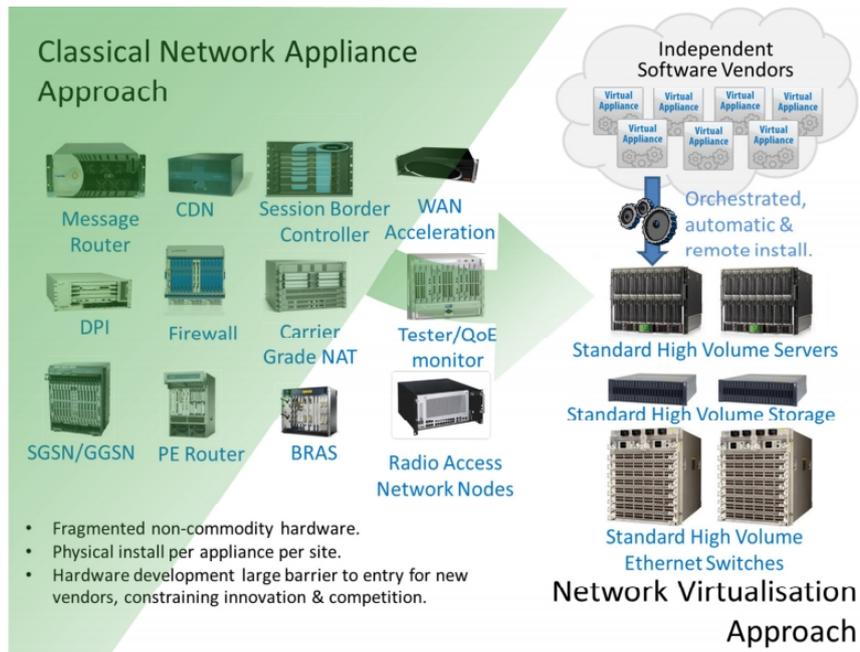


Figure 1.1: Network Function Virtualization [7]

1.2.2.2 VNF

Whereas NFV refers to the process of separating network functions from hardware to create a virtualized network that can run on commodity hardware virtual network functions (VNFs) refer to specific network functions like firewalls, Dynamic Host Configuration Protocols (DHCP) servers or load balancing. Such Individual VNFs can be connected or combined together as building blocks to create a fully virtualized environment as shown in Figure 1.2. VNFs run on virtual machines (VMs) on top of the hardware networking infrastructure. There can be multiple VMs on one hardware box using all of the box’s resources. For the user there shouldn’t be any performance differences whether the function is running on dedicated hardware at the users premise or VMs [28].

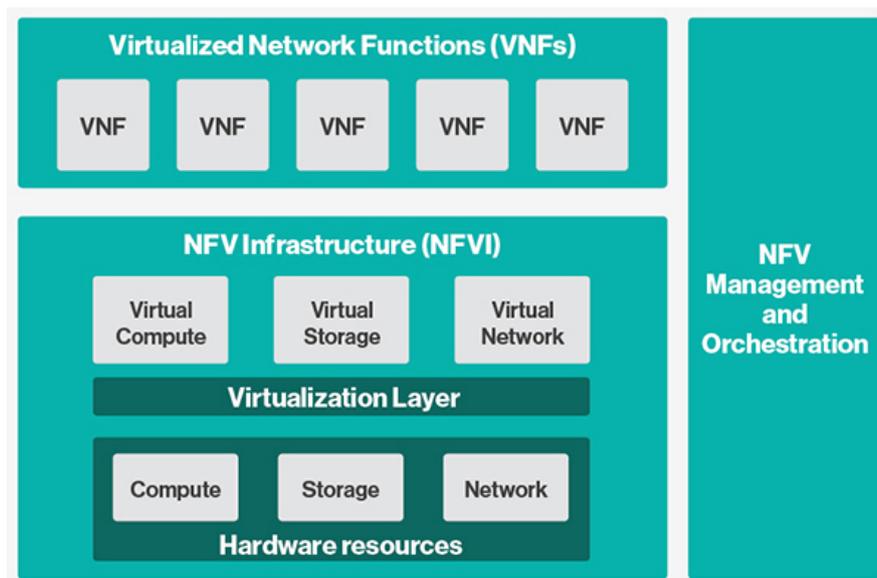


Figure 1.2: VNF building blocks [28]

1.2.2.3 Software Defined Networks

One architecture model that is closely related to NFV are Software-Defined Networking (SDN). SDN is a network architecture, which enables a network to be intelligently, centrally controlled or programmed by using software applications. Operators can manage regardless of the underlying technology, the entire network consistently [9].

The SDN architecture consists of three main layer as shown in Figure 1.3. Firstly, the application layer contains the typical network applications or functions organizations use. This can include intrusion detection systems, load balancing or firewalls. The control layer represents the centralized SDN controller software that acts as the brain of the software-defined network. This controller resides on a server and manages policies and the flow of traffic throughout the network. The control layer interacts with the application layer through an API. An administrator or a network engineer can shape traffic from a centralized control console without having to touch individual switches in the network. The centralized SDN controller directs the switches to deliver network services wherever they are needed, independent of the specific connections between a server and devices. This process is a move away from traditional network architecture, in which individual network devices make traffic decisions based on their configured routing tables. Lastly there exists the third layer, the infrastructure layer, which is made up of the physical switches in the network.

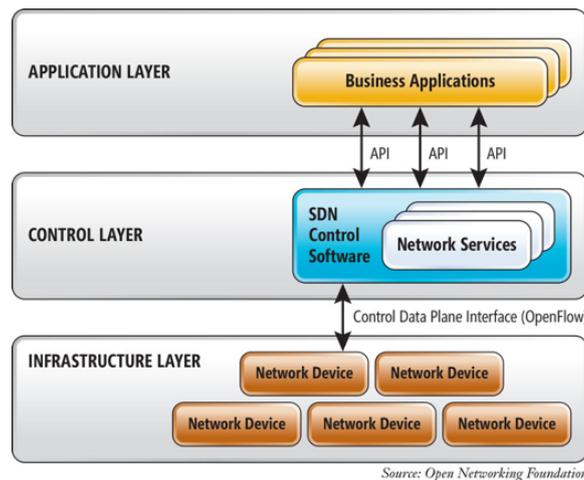


Figure 1.3: SDN Architecture [27]

1.3 Marketplaces for Networking

Online Marketplaces have evolved rapidly in recent years. With examples like the App-Store or GooglePlay Store launching in 2008. Thanks to these stores the smartphone users became able to take full advantage of the computing power the new devices brought with them by easily accessing a growing pool of applications [17].

Not only mobile phone companies made use of such online marketplaces for the distribution of their applications, but also Cloud Computing providers did too. Prominent figures include Amazon Web Service (AWS), Microsoft Azure, and OpenStack. AWS is still considered a marked leader thanks to its first mover advantage for establishing itself in 2006. While Azure and Openstack are big players too, they have not yet been able to catch up to AWS [14].

A more recent kind of marketplace has emerged in tackling services related to networking. Following the NFV framework published by the European Telecommunication Standard

institute (ETSI) in 2012 [26], multiple approaches to push the adoption of NFV have been introduced, such as CISCO, T-Nova, Bunker, and FENDE [1].

First network softwarization could benefit from a online marketplace, by reducing the time to market of new network applications significantly. Furthermore, it may be possible to accelerate the introduction of new network applications by automating tasks related to deployment, management and also. There are two new main X-as-a-Service business models in networking:

- **NFVaaS:** Network Function Virtualization as a Service infrastructure in which you as a Service Provider allow resource slicing to enable multiple virtual providers.
- **VNFaaS:** A service provider operating a VNF instance using its NFV Infrastructure. The VNF is in this case the service providers application. The Enterprise is the consumer of the service. Therefore the enterprise does not have to invest additional capital in advanced network features [25].

As part of this study we picked five marketplaces which we think are worth comparing when talking about networking services. We've picked AWS to cover an established cloud computing provider and Cisco to elaborate on an more NFV specific marketplace. Furthermore we have decided to included 3 NFV marketplace prototypes, FENDE, T-NOVA and Bunker to discuss current research directions and possible marketplace solutions.

1.3.1 Cisco Marketplace

Cisco, most well known for their networking products, introduced their Cisco NFV Infrastructure Solution, which provides the compute, storage, networking infrastructure, and management and assurance capabilities to run network function virtualization. It is a fully integrated solution that is tested and validated by Cisco. It is marketed as a robust solution that delivers high performance, availability, security, and scalability. The Cisco NFVI as shown in Figure 1.4 builds a solid foundation that will improve cost efficiency, enable faster service deployment, allows to integrate hundreds of third party VNFs into their infrastructure [13].

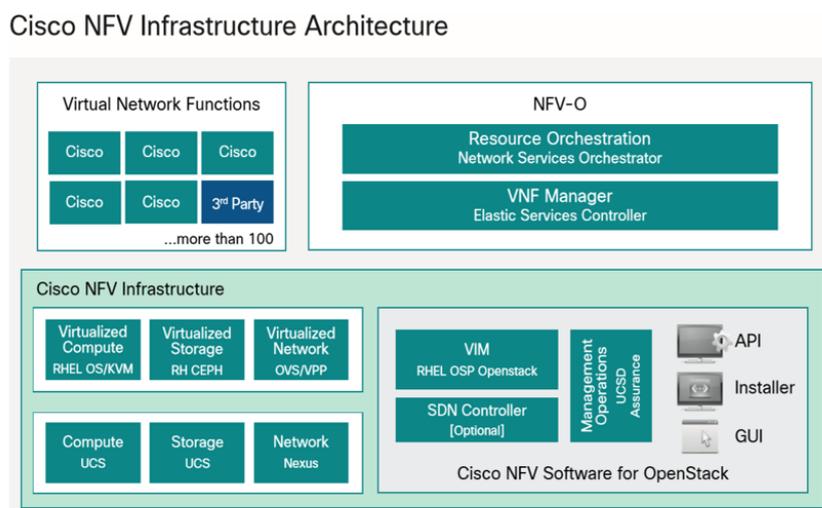


Figure 1.4: Ciscos NFV Infrastructure [13]

Ciscos Marketplace was launched in 2014 as a source for secure, validated enterprise-class apps, products, solutions and services [12]. Generally speaking only certified Cisco Solution Partners are allowed to publish their software on the marketplace. In addition, the software must be compatible to Cisco's Technologies, which excludes many other

network function providers. In contrast to other popular that have emerged in the past few years it does neither resolve conflict and dependency issues nor does it support the whole life-cycle. This means that the users need to deploy, manage and update the solution and services themselves. Furthermore, users can not directly download or buy the solutions on the marketplace itself, but can only select the service or solution that are being shown on the marketplace and are then redirected to the customers web-page on which they can buy the solutions, usually based on a fixed-price model. Thus, given the functionality and range of software products it provides on the marketplace in comparison with the other marketplaces, the additional benefit for users on this marketplace is relatively low.

1.3.2 AWS Marketplace

AWS was launched in 2006 to provide developers with on-demand IT infrastructure, with a focus on enterprises and fewer end users from the outset. The EC2 (Elastic Compute Cloud) are virtual servers running either a Linux distribution or Microsoft Windows Server. They do not have a fixed contract term and are billed by the hour. There are several tariffs to choose from, depending on the amount of memory available and the compute units that correspond to virtual processors. With EC2, the computer capacity can be adapted exactly to the respective requirements and can be reduced or enlarged depending on the requirements of the customer. Amazon EC2 is fully controllable, with full root access for each server instance Amazon[11]. The AWS Marketplace was launched for their EC2 product as a curated digital catalog that makes it easy for customers to find, buy, deploy, and manage third-party software and services that customers need to build solutions and run their businesses [10]. AWS Marketplace contains thousands of software offerings in frequently requested categories including security, networking, storage, machine learning, business intelligence, database, and DevOps, which are being developed and published by either Amazon them-self or by third party developers. In order to be able to publish their software, third party developers need to be either certified or let their product go through an review process. This is to make sure that no malware software is published on the amazon marketplace. One key selling point of the AWS Marketplace is the support of the whole life cycle as well as the dependency and conflict resolve mechanisms. This relieves the customers of many duties as they do not need to waste time and money deploying, managing and updating the third party applications. Amazon choose both fixed-price as well as pay-as-you-go business models for their marketplace. Depending on the solution the provider can select how the service is charged [10]. All in all amazons AWS marketplace benefits all stakeholder in a major way. First amazon itself benefits from the large amount of additional software and features that is being developed for their cloud computing platform, which makes it more attractive. Second the third party developers profit from being able to reach their customers easily and thirdly the customer profits by using a platform that delivers many functionalities, relieves him from many duties and therefore often result in cost savings for the customer.

1.3.3 T-NOVA

T-NOVA is a integrated project of a Network Function as-a-Service infrastructure that was co-funded by the European Commission and the 7th Framework Programme. This project, which is open-source introduced, for example a novel business model and cases for VNF. The aim of T-NOVA is to further promote a NFV through framework that allows for operators to not only deploy VNFs but also offer them to customers. As these services will be provided as-a-service, the customer will not need to worry about installing and maintaining of any kind of hardware anymore to deploy the desired functions. Additionally, on the operator's side, these functions will be of added value as they can sell their

services.

The core team of the T-NOVA project has been made up of people from the academic field as well as SMEs. This also explains why it is aimed at SMEs as well as academic institutions. To facilitate the involvement of a lot of different actors to contribute their NFVs, T-NOVA introduces a *NFV Marketplace*. The aim of this marketplace is to have several developers publish their network services and functions. Customers on the other end can then look for and select the desired services. T-NOVA's innovative design doesn't end here, as they also introduced a way for customers to bid for the NFV's associated Service Level Agreement (SLA) [29].

T-NOVA realizes this NFVaaS concept by introducing an integrated management platform. The idea behind the platform is to have a central instance that is responsible for the configuration, automated provision, optimization, and monitoring of the deployed VNFs. This means that T-NOVA aims to allow for elastic provision and allocation of IT resources wherever needed most in the hosting network of the deployed functions [6].

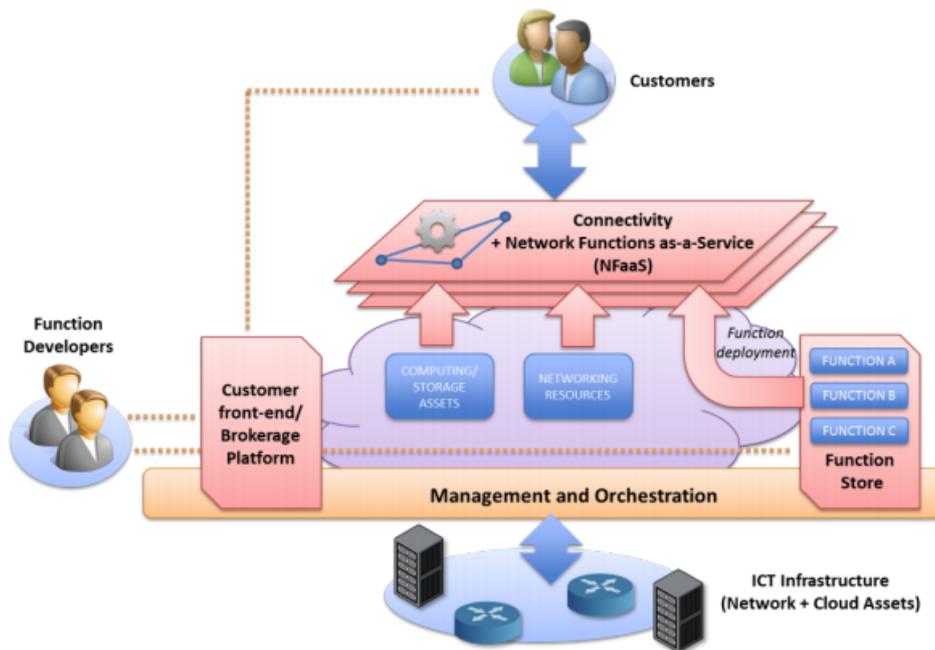


Figure 1.5: High-level architecture of the T-NOVA platform [5]

As mentioned above, T-NOVA introduces a *NFV Marketplace*, depicted in Figure 1.5 as "Function Store". This store allows for developers, including third party developers to develop and publish their services through the marketplace [5]. The idea behind allowing for independent entities to enter the market is the hope of a rapid growth of services available in the store, which should in turn boost the introduction of VNFs into the market [31]. The customers that want to acquire services can then browse through the store which and purchase services in a customer specific front-end platform [6].

The process of managing and orchestrating resources seen in Figure 1.5 is done by the previously mentioned *Orchestrator platform*. A more in-depth visualization can be seen in Figure 1.6. The main functions of the *Orchestrator platform* can be summarized as the automated deployment and configuration of Network Functions as well as the management and optimization of resources needed for them. As seen in Figure 1.6, the platform's tasks consist of a resource repository, a resource mapping, a connectivity-, cloud and NF management, the monitoring and optimization as well as high availability. The resource repository shows a network topology with all the available IT resources for the VNFs. Resource mapping means deciding on as well as optimizing resources for NF. The connectivity management depends on the given SLA and instantiates virtual machines (VMs)

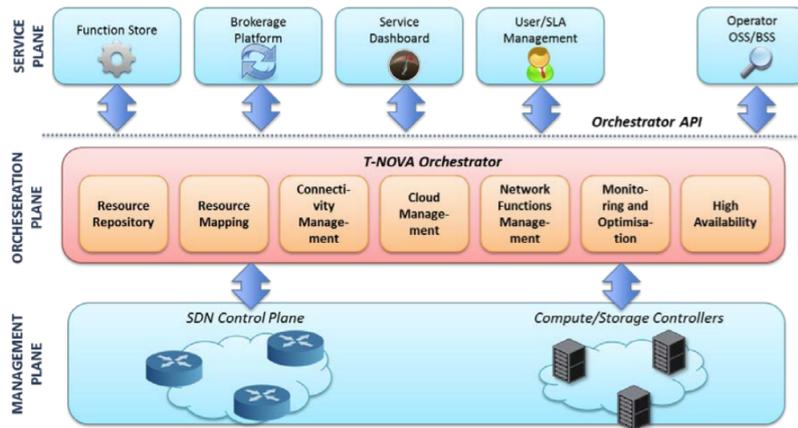


Figure 1.6: Visualization of the T-NOVA *Orchestrator platform* [5]

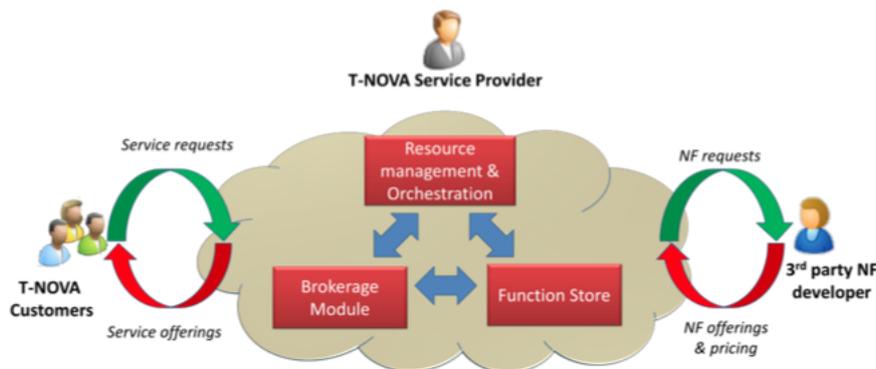


Figure 1.7: Visualization of marketplace [30]

and ensures quality of service (QoS). Cloud management means the communication with the given cloud controller API, where computing is allocated and resources for NFs are stored. All the deployed NFs are parameterized at the NF management. Monitoring and Optimization means monitoring the network and storage assets as well as optimizing them. Finally, high availability means the detection and prediction of operational anomalies. The whole management of all these resources within T-NOVA will be based on open APIs, this means that the *Orchestrator platform* is done through an open API which has to be defined[5].

Furthermore, in order to attract third parties to participate in the NFV market, T-NOVA's NFV Marketplace allows for a variety of billing models. Apart from the more traditional payment methods like *pay-as-you-go* and *fixed-price*, T-NOVA store introduces a key requirement through a more novel approach that incorporates auction-based model where customers can bid to acquire a certain service [5]. Through the brokerage module, depicted in Figure 1.7, customers can acquire VNFs. It also enables the trading of VNFs and facilitates the auctioning between the Function Providers [31]. The way in which a customer can acquire a service is by placing a request through the brokerage module where they declare what VNFs exactly they want to acquire. Upon that request, they will receive offers with the corresponding SLAs [30].

There is currently no information about market adoption of T-NOVA available since the project has ended and from there on never left the prototype phase. However, the concept of T-NOVA still provided a solid basis for future research in that area. It introduced new concepts of billing methods and alongside that, provided a solid basis for future research especially through its innovative *Orchestrator platform* and the configurability and adaptability of its NFs. Another topic of discussion concerns the definition of the

SLA. Who is to be made responsible for the SLA? Should it be the supplier of the hardware that is running the software made by someone else, or should it be the supplier of the software that has no say about on what hardware their software runs on[5]?

1.3.4 FENDE

FENDE not only provides NFV marketplace features but also promotes itself as a whole ecosystem for NFVs [1]. While FENDE was deployed in three network domains in Brazil and is currently running as a prototype, development and research is still ongoing [1]. Four main requirements important to a NFV marketplace have been identified during the development of FENDE [1]. The researchers highlight that offering, execution, accounting and management are key. They note that these requirements are currently only partially covered by the competition therefore FENDE wants to capitalize on them and tries to integrate them all into one marketplace. Hence FENDE tries to combine marketplace, management and infrastructure capabilities [1].

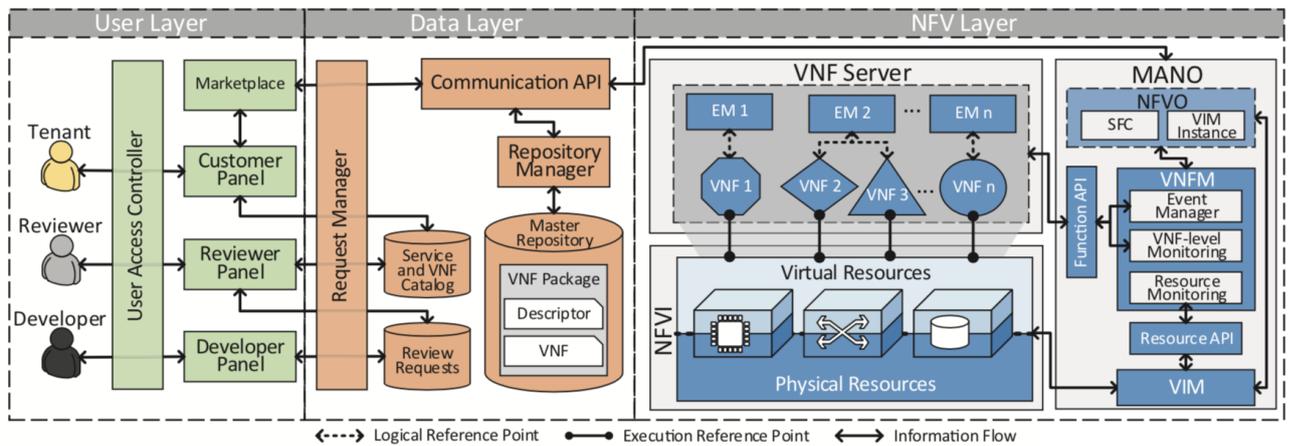


Figure 1.8: The Architecture of FENDE [1]

The researchers behind FENDE developed an architecture, as presented in Fig. 1.8, which closely follows the NFV framework introduced by ETSI [26] and is split into three different layers (User Layer, Data Layer, and the NFV Layer) [1]. This architecture will be discussed in the next few paragraphs. To explain these three layers and their components this study will refer to the paper proposing FENDE [1].

User Layer: As the researcher stated in their paper this layer distinguishes between three different users (the developers, reviewers and customers), which all communicate through a specific panel with the system. They explain that the developers can register their VNF which will then be reviewed by the reviewer, before eventually becoming accessible to the customers (tenants) on the marketplace. Next the customers can start instances, perform life cycle management and create SFCs. While looking at the user layer one can quickly see that FENDE was designed with the idea in mind that 3rd party developers can easily deploy their VNFs onto the platform, similar to developers putting their apps on the AppStore [1].

Data Layer: According to the researcher introducing FENDE the data layer is responsible for managing data related to the users of the platform as well as the data related to the SFCs and VNFs. To achieve this they include three different databases: the service catalog, the master repository and the review requests. They explain that working on the master repository is the repository manager which oversees the VNFs descriptors that are accessible in the catalog. According to them any information from or to the repository manager is forwarded through the communication API. The paper continues to explain that the communication API enables and facilitates the interaction between

the NFV layer and the users of FENDE. Looking at the development side, the review request manager manages VNF registration inquiries and keeps the review request and service catalog databases up to date [1].

NFV Layer: The NFV layer consists of three sub layers as can be seen in Fig. 1.8. The paper introducing FENDE explains that within the NFV Management and Orchestration (MANO) sub layer the Virtual Network Manager (VNFM) manages the life cycle operations and the creation of SFCs. The paper proceeds on illustrating that as part of the VNFM the event manager receives the requests from the user layer, the resource monitoring oversees the physical resources of a created VNF and the VNF-Level monitoring gathers data regarding the function usage of a VNF. The researcher state that in order to guide the VNFM in its life cycle management, the NFV Orchestrator orchestrates the process of service allocation and composition. Noteworthy is that, according to the researcher, FENDE is not restricted to only one VIM (to manage its available resources), because thanks to its communication API FENDE facilitates different VIMs, enabling the composition of different infrastructures like the ones based on CloudStack and OpenStack. Briefly touching upon the other two remaining sub layers they outline that the VNF server is mainly responsible for the local maintenance of VNFs operations while the virtual and physical resources, that can be used for deployment, are referred to by the NFV Infrastructure [1].

In it's current prototype state it's already visible that 3rd party developers shall play a huge role in the ecosystem of FENDE. While currently reviewing is still done manually, potential autonomic revision mechanics have been discussed [1]. No matter the review process it has to be ensured that no malicious code gets onto the ecosystem. The customers ability to conduct life cycle management and SFC through a user friendly interface all by themselves [1] shows the commitment of FENDE to make NFV more accessible. Another note worthy feature is that the VIM can use the communication API to abstract technology specific commands, by doing so FENDE supports the composition of different infrastructures [1]. The paper introducing FENDE is not specific about what kind of business models they eventually want to apply. However they elaborate on methods like pay-as-you-go, fixed price, auction based and custom built as potential solutions [1].

1.3.5 BUNKER

BUNKER is a prototype that is currently being developed by the Communication Systems Group CSG at the Department of Informatics at the University of Zurich. BUNKER stands for "Blockchain-based trUsted VNF pacKagE Repository." Security is a major factor to consider when it comes to designing and employing NFV. As an end-user of these services, security is, apart from functionality, one of the main aspects when it comes to deciding on which provider of services to go for. As of now, in order to ensure that packages are neither malicious, nor have they been tampered with, only a few solutions have been introduced to tackle these problems. Currently, Trusted Platform Modules (TPM) as well as remote attestation services are responsible for verifying the quality and security of these packages through centralized databases. This means that ultimately, the user of these products has to believe in the integrity of that provider [32].

This is where BUNKER comes in. BUNKER aims to improve the aspect of security of NFV without relying on a central trusted authority and instead to make use of blockchain technology [8]. Blockchain was first introduced in 2009 with the introduction of Bitcoin [33]. The distributed ledger of the blockchain is comprised of blocks of data, where each block contains the hash of the previous block and is cryptographically fixed, as well as information about transactions. What this means is that the data stored in such a block cannot be changed. As the hash of each block is stored in the previous one, changing just one block would mean all other blocks would become invalid as well [32]. Therefore,

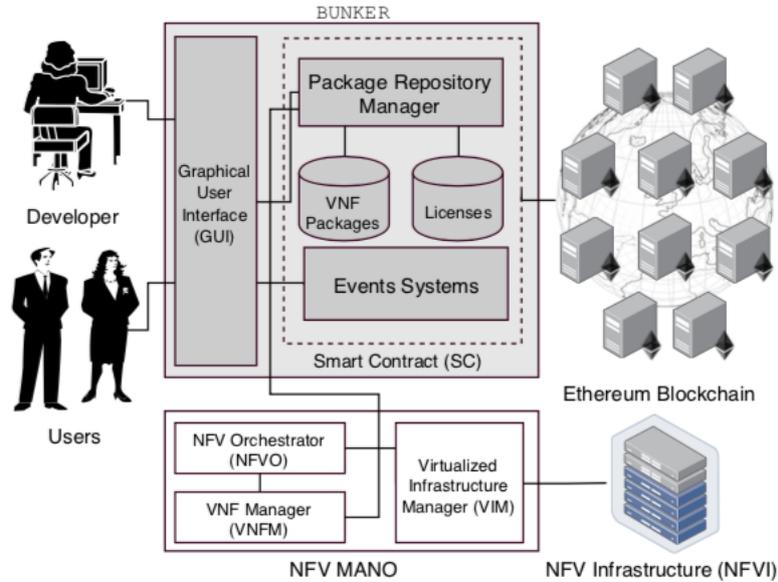


Figure 1.9: Visualization of marketplace [6]

data immutability and data decentralization are some of the most important properties of a blockchain [8]. Another crucial element of blockchains are the smart contracts (SCs). Contrary to regular contracts, a SC does not rely on any third party to authorize them, it is dependent on blockchains to facilitate, execute and enforce the terms of an agreement between untrusted parties” [34]. This means they facilitate trusted exchanges between untrusted parties [8].

BUNKER makes use of the Ethereum blockchain [8]. The Ethereum SC are, in contrast to e.g. Bitcoin SC, Turing-complete, which introduces some security risks but also allows for complex functionality. Nodes in Ethereum store the most recent state of each contract [32]. In the case of VNF, BUNKER represents a concept of an idea to incorporate blockchain-based trusted VNF packages. End-users and developers can interact with each other without any Trusted Third Party (TTP) involved. Fees will be automatically transferred from customer to developer without any TTP. End-users also are no longer forced to trust any central authority on whether their package could be malicious, as BUNKER stores the hash of the acquired package so that users can verify if it has been tampered with [8]. The combination a VNF repository with the blockchain technology makes the idea of having an internal or external trusted authenticator superfluous, as the authentication process can easily be done through a distributed ledger. All of the requests are sent to a blockchain, where they are authenticated and only after a successful authentication, the virtualization begins. The status of the virtualization is again sent back to the blockchain, which is again stored in the blockchain where it can not be tampered with [8].

BUNKER has two main components: the previously mentioned **Smart Contracts** and the **Graphical User Interface** (GUI), which can be seen in Figure 1.9. The *SC* of BUNKER are built on the Ethereum blockchain and have 4 main parts [35]: *Events System*, *License Database*, *Package Repository Manager* and *VNF Package Database*.

The *Events System* is where all the events are emitted and managed. An event is stored in the log of a transaction and other applications can wait for and react to specific events. BUNKER has defined a specific event *license* which takes place whenever a package is bought. *License* contains information about the transaction, the hash address as well as a link to the VNF [8]. The application that waits for this event is in this case the GUI. Other possibilities would be for the NFV MANO to wait for a specific event and then automatically deploy the VNF [32].

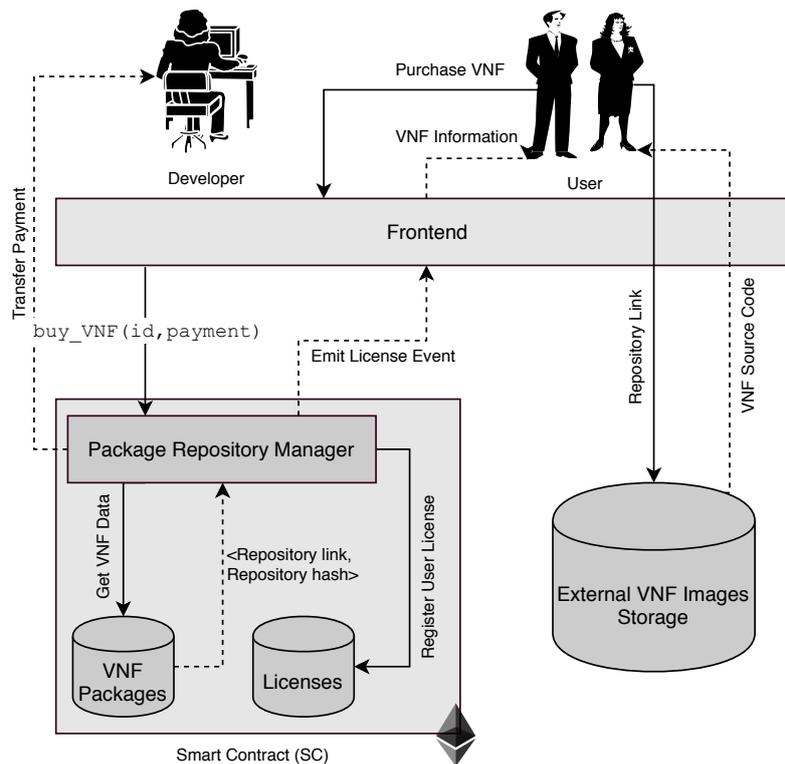


Figure 1.10: Acquiring a package [6]

Package Repository Manager is where all the requests to the repository go through and if necessary, accesses the repository. It is therefore responsible for creating and maintaining VNF package entries. This means that the repository also acts as an authenticator of each user before execution [8].

The *VNF Package Database* keeps track of package information such as the license, ratings, hashes etc. and can only be accessed through the repository manager. Ideally, all of the information would be stored in the SC, but as this gets more and more expensive with the more data is being stored, BUNKER makes use of an external data storage to store the package code itself. Security is still guaranteed as the packages can be verified by comparing the hashes [8].

The *GUI* is where the user interacts with the system and where he/she gets all the relevant information displayed. Users can see the packages available to buy and the packages they have already acquired. Alongside with every package, users can access ratings of other users and prices. Developers on the other end are able to edit the information displayed for their package. BUNKER specifies the following functions: *Registration and Upgrade System*, *Licensing System*, *Verification System* and *Rating System*.

Registration and Update is where developers can submit new VNF packages to the repository. They can also update existing packages properties, such as the information or functions. The *Licensing Systems* is where customers acquire the license of a package. The user places a request for a license of a package through the front end by creating a blockchain transaction, where the different fees are included. The license event is emitted after the SC checks if the funds are sufficient. The license event contains the information needed to retrieve and deploy the VNF.

The *Verification System* ensures the package's quality and verifiability by comparing the hash of the package in the repository with the one that was downloaded. This takes place whenever new packages are acquired or during run time, which ensures proper life cycle operations. Finally, the *Rating System* describes the rating mechanism available on packages so that end-users can review the acquired products. This should stimulate the quality of packages available in the store as opinions of other users can be considered

Table 1.1: Comparison of different networking marketplaces

Provider	Technology	Provided Services	Status	Business Model	Developer Access	3rd Party Developers	Deployment Model
FENDE	VNF	VNF (Publishing & Deployment), Management and Orchestration (life cycle management, SFCs etc.), Infrastructure (composition of different infrastructures)	Prototype	Fixed-price, Pay-as-you-go, auctions, custom-built (mentioned as potential use case)	Review Process	Allowed	NA
T-NOVA	VNF	VNF (Publishing & Deployment), Management and Orchestration (Orchestrator Platform), adaptable to underlying cloud API	Prototype	Fixed-price, Pay-as-you-go, auctions, auctions (mentioned as potential use case)	NA	Allowed	NA
BUNKER	VNF	block-chain based VNF (Publishing & Deployed), SC to manage packages, underlying MANO implementation dependent	Prototype	NA	NA	Implementation dependent	Implementation dependent
AWS	Cloud Computing	Reliable, Scalable, and Economical Cloud Computing Services (Publishing & Deployment), Management and Orchestration through whole life cycle (Orchestrator Platform)	In use	Fixed-price, Pay-as-you-go	Review Process	Allowed	Instantiate VMs and install 3rd party software
Cisco Marketplace	Cloud Computing, NFV	Solutions built on Cisco product and architectures, Cloud and Managed Services, Global Partner Network	In use	Fixed-price	Certification	Allowed	Redirects to downloads in developer's site

when buying a package. An issue all marketplaces have yet to find a solution for is the credibility of these ratings as they could be fake [8].

While BUNKER looks like a promising concept, there are some areas that still need further research. Storing the package's code in an external database somewhat defeats the purpose of having the repository being built on a decentralized system. Furthermore, the limitations of language can be a problem when it come to verifying the quality of the offerings through the ratings, but this is not just a concern of BUNKER but more a concern of marketplaces in general. One big advantage of BUNKER remains that it can be integrated into existing NFV solutions.

1.4 Comparison of Current Marketplaces

As can be seen in Table 1.11, the variety of different solutions is quite large in the way they are carried out. We have seen different technologies being implemented such as blockchain. But the providers also differ in their business models, some allowing dynamic pricing and even enable auction like pricing models. Another important aspect to note is the access criteria for 3rd party developers to publish their services on the marketplaces, as some solutions have defined criteria whereas others do not specify how they would handle it. Overall, it is evident that there is a lot of development happening in the area with a lot of the proposed solutions being in the prototype phase still. Many of the providers try to find and experiment with new approaches to add to their solutions.

1.5 Challenges and Opportunities regarding Marketplaces for Networking

We were able to show that prototypical marketplaces for networking such as FENDE, BUNKER and T-NOVA would offer significant benefits to organizations, that are currently investing large sums of money in their own network infrastructure. Marketplaces for networking would reduce the time to market not only for the network function developers and providers but would also be highly beneficial for startups that do not need to invest high capital expenditure to start growing their business and are also able to extend their existing network quickly by renting or buying new network functions depending on their needs. Moreover organizations are able to save money in the long term, since the network provider is mainly responsible for deploying, managing the network functions and infrastructure.

Despite the advantages marketplaces for networking provide and its appeal as shown in the examples before, there are still many challenges that prevent a fast adoption of networking marketplaces. Four main challenges have been identified , that are currently slowing the process of wide adoption of network marketplaces:

- Network end users must be able to understand the capacities of network devices hosting network applications with the requirements different network applications have.
- Network end-users need to manage and sort out issues related to dependencies and conflicts among different network applications themselves.
- End-users need to manually tune NetApps before running them, due to the heterogeneity of execution environments.
- Another challenge emerges when defining the responsibility of the SLA. As elaborated it has to be decided whether the developer of the network function or the network host will be held accountable
- Usually, a network applications is written having a specific network architecture in mind such as NFV or SDN. Therefor network applications are usually not agnostic with regards to the underlying networking technology.
- Note that, despite the goal to maximize the compatibility with widely adopted cloud and network controllers, the integration of an orchestrator with other control frameworks is possible as long as appropriate plugins are developed. Such an example can be seen in the T-Nova platform

- Given that various providers allow 3rd party developers to provide network function through their marketplace and given the fact that network services must fulfill the highest security standard for organizations, it has to be ensured that no malicious code find its way onto the platforms. Introducing a fast and secure reviewing process might be a challenge.
- Various business models have been discussed. Pay-as-you-go, fixed-price, custom-build and auction based are potentially solutions. Yet the right business models which benefits both parties the most has not been identified. Therefore, future research on these different models is recommended.
- To mention Bunker specifically one prevalent issue is that the goal of pure decentralization is currently not achieved given the fact that storing certain parts of the packages requires the use of a trusted third party database. However this problem is not limited to BUNKER's implementation specifically but stems from the architecture of blockchain and its combined costs for storing information in a SC, which might be subject to change in the future.

1.6 Discussion and Conclusion

We have seen a lot of different approaches concerning the virtualization of network functions. A lot of the proposed solutions looked promising in terms of their technology used as well as in their business models, all with the overall goal to attract more people to the NFV market. Furthermore, new network architectures exist such as SDN, which can make network deployments and operations much faster than today, hence being one of the enablers of fast adoption of network marketplaces.

Looking back at early development of application marketplaces, these solutions have solved a lot of problems and adapted their overall design well to the domain, creating dynamic business models that calculate price depending on usage as well as ensuring quality and integrity of the packages through encryption. This innovative thinking generally indicates that there is a lot of interest and research going on around NFV.

On the other side we discussed open research challenges, which are slowing the adoption of networking marketplaces. We showed issues regarding heterogeneity, auditing, recommendation and security, which require considerable research efforts. The approach followed by BUNKER to have a decentralized repository only works to a certain degree that way as some parts of the packages still rely on being saved on an external database. The overall issue of auditing is, while by some, not specified by all of the proposed solutions. As it is with any online marketplace, the genuity of the reviews remains to be suspect of future research as it is still difficult to avoid fake reviews. We discussed research challenges to the adoption of network marketplaces. In particular, we found that issues regarding heterogeneity, auditing, recommendation, placement, security, and NetApps description still require considerable research efforts. Therefore, the future of networking marketplaces involves more in-depth research in the areas mentioned, even more so with the increase of new technology used. While this development is highly dependent of the innovation and research of new technologies which may or may not be adapted into marketplaces, the relevance of network marketplaces does not seem to decrease any time soon.

After elaborating on the concepts of cloud computing and networking the study compared five different marketplaces for networking. FENDE, BUNKER and T-Nova have been chosen to represent projects in the prototype stage while Cisco and AWS serve as examples for already established marketplaces. While they differ in their approaches they share the same goal of providing networking services through a marketplace. These different approaches have been discussed while highlighting potential challenges and opportunities.

Bibliography

- [1] L. Bondan, M. F. Franco, L. Marcuzzo, G. Venancio, R. L. Santos, R. J. Pfitscher, E. J. Scheid, B. Stiller, F. De Turck, E. P. Duarte, A. E. Schaeffer-Filho, C. R. P. Santos, and L. Z. Granville; FENDE: Marketplace-Based Distribution, Execution, and Life Cycle Management of VNFs, *IEEE Communications Magazine*, 57(1): January 2019, pp 1389-1406.
- [2] P. Mell, T. Grance; The NIST definition of cloud computing, National Institute of Standards and Technology, 2011, Special Publication 800-145.
- [3] M. Raza; Reprinted from SaaS vs PaaS vs IaaS: What's The Difference and How To Choose by Stephen Watts, Retrieved from <https://www.bmc.com/blogs/saas-vs-paas-vs-iaas-whats-the-difference-and-how-to-choose/>, Last visit 1 November 2019.
- [4] J. F. Kurose, K. W. Ross; *Computer Networking: A Top Down Approach*, 2012, 6th Edition, Pearson, Book.
- [5] Xilouris, G., Trouva, E., Lobillo, F., Soares, J. M., Carapinha, J., McGrath, M. J. Rebahi, Y.; T-Nova: A Marketplace for Virtualized Network functions, *European Conference on Networks and Communications (EuCNC)*, Bologna, Italy, June 2014, pp. 1-5.
- [6] Xilouris, G., Kourtis, M. A., McGrath, M. J., Riccobene, V., Petralia, G., Markakis, E. & Ramos, A.; T-nova: Network functions as-a-service over virtualised infrastructures, In *2015 IEEE Conference on Network Function Virtualization and Software Defined Network (NFV-SDN)*, 2015, pp. 13-14.
- [7] Z. Ghadialy; Bringing Network Function Virtualization (NFV) to LTE, <https://blog.3g4g.co.uk/2014/12/bringing-network-function.html>, last visit 5 December 2019.
- [8] E. Scheid, M. Keller, M. Franco, B. Stiller; BUNKER: a Blockchain-based trUsted VNF pacKagE Repository, *16th International Conference on Economics of Grids, Clouds, Systems, and Services (GECON 2019)*, Leeds, UK, September 17-19, pp. 1-6.
- [9] K. Benzekki, A. E. Fergougou, E. B. Abdelbak; Software-defined networking (SDN), *A Survey, Security and Communication Networks*, Vol 9, No. 18, pp. 5803-5833.
- [10] Amazon; AWS Marketplace, Retrieved from <https://aws.amazon.com/marketplace>, Last visit 5 November, 2019.
- [11] Amazon; Was ist AWS Marketplace?, Retrieved from https://docs.aws.amazon.com/de_de/marketplace/latest/userguide/what-is-marketplace.html, Last visit November 8, 2019.

- [12] Cisco; Cisco Marketplace, Retrieved from <http://marketplace.cisco.com/home>, Last visit 1 November, 2019.
- [13] Cisco; Cisco NFV Infrastructure, Retrieved from <https://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/network-functions-virtualization-nfv-infrastructure/cisco-nfv-infrastructure-aag.pdf>, Last visit 1 November, 2019.
- [14] N. Serrano, G. Gallardo and J. Hernantes; Infrastructure as a Service and Cloud Technologies, in *IEEE Software*, Vol. 32, No. 2, 2015, pp. 30-36.
- [15] IBM; The enterprise outlook on cloud-native development, Retrieved from <https://www.ibm.com/cloud/cloud-native-research/>, Last visit 14 November, 2019
- [16] J. F. Gantz; The Salesforce Economy: Enabling 1.9 Million New Jobs and \$389 Billion in New Revenue Over the Next Five Years, 2016, White Paper.
- [17] W. Martin, F. Sarro, Y. Jia, Y. Zhang and M. Harman; A Survey of App Store Analysis for Software Engineering, in *IEEE Transactions on Software Engineering*, Vol. 43, No. 9, 2017, pp. 817-847.
- [18] Mijumbi R., Serrat J., Gorricho J., Bouten N., De Turck F. and Boutaba R.; Network Function Virtualization: State-of-the-Art and Research Challenges, in *IEEE Communications Surveys and Tutorials*, Vol. 18, No. 1, 2016, pp. 236-262.
- [19] Techopedia; What does Network Functions Virtualization (NFV) mean?, Retrieved from <https://www.techopedia.com/definition/5537/network#:~:targetText=A%20network%2C%20in%20computing%2C%20is,physical%20and%20For%20wireless%20connections.>, Last visit 10 November, 2019.
- [20] SDxCentral Staff; What is a Virtual Network Function or VNF?, 2014, <https://www.sdxcentral.com/networking/nfv/definitions/virtual-network-function/>, Last visit November 1, 2019.
- [21] ETSI NFV ISG; Network Functions Virtualization, Retrieved from https://portal.etsi.org/nfv/nfv_white_paper.pdf, Last visit November 8, 2019.
- [22] G. Gumpert, S. J. Drucker; From the agora to the electronic shopping mall, *Critical Studies in Mass Communication*, Vol. 9, No. 2, 1992, 186-200.
- [23] S. Goyal; Public vs Private vs Hybrid vs Community - Cloud Computing: A Critical Review, *IJCNIS*, Vol. 6, No. 3, 2014, pp.20-29.
- [24] Y. Jadeja, K. Modi; Cloud computing - concepts, architecture and challenges, 2012 International Conference on Computing, Electronics and Electrical Technologies (IC-CEET), Kumaracoil, 2012, pp. 877-880.
- [25] ETSI Group; Network Function Virtualization; Use Cases, 2013, Retrieved from https://www.etsi.org/deliver/etsi_gs/NFV/001_099/001/01.01.01_60/gs_NFV001v010101p.pdf, Last visit 12 November, 2019.
- [26] ETSI Group; NFv Whitepaper ETSI, 2012, Retrieved from https://portal.etsi.org/NFV/NFV_White_Paper.pdf, Last visit 2 November, 2019.
- [27] Open Networking Foundation; SDN Architecture, 2016, Issue 1.1, Retrieved from https://www.opennetworking.org/wp-content/uploads/2014/10/TR-521_SDN_Architecture_issue_1.1.pdf, Last visit 13 November, 2019.

- [28] Calsoft; NFV vs. VNF: What's the difference?, 2016, Retrieved from https://blog.calsoftinc.com/industry_insights/nfv-vs-vnf-whats-difference, Last visit 12 November, 2019.
- [29] T-NOVA; Project Overview, Retrieved from <http://www.T-NOVA.eu/overview/>, Last visit 12 November, 2019.
- [30] T-NOVA; Overview: Network Functions as-a-Service over Virtualised Infrastructures, 2013, Retrieved from <https://blog.zhaw.ch/icclab/category/projects/t-nova/>, Last visit 12 November, 2019.
- [31] Ramos, A., Melian, J., Markakis, E., Alexiou, G., Harsh, P., & Perez, M.; T-NOVA SLA and Billing. Network Functions as-a-Service over Virtualised Infrastructures, Deliverable D6.4., 2015, pp. 38-47.
- [32] Keller, M.; Design and Implementation of a Blockchain-based Trusted VNF Package Repository, Communication Systems Group CSG, Department of Informatics Ifi, University of Zurich, 2019, pp. 1-69.
- [33] S. Nakamoto; Bitcoin: A Peer-to-Peer Electronic Cash System, 2009, Retrieved from <https://bitcoin.org/bitcoin.pdf>, Last visit 10 November, 2019.
- [34] M. Alharby, A. van Moorsel; Blockchain-based Smart Contracts: A Systematic Mapping Study, International Conference on Cloud Computing, Big Data and Blockchain (ICCB), Fuzhou, China, 2018, pp. 1-6.
- [35] Ethereum Foundation; Solidity - Solidity 0.58.0 Documentation, Retrieved from <https://solidity.readthedocs.io/>, Last visit 10 November, 2019.

Chapter 2

An Economic Analysis of the Migration of Geographical Information Systems (GIS) to the Cloud.

Stéphanie Wismer, Jan Weber, Silvan Caduff, Sophie Sturzenegger

Spatial data is almost omnipresent nowadays, and not only traditional systems such as satellites are generating geospatial data, but also non-traditional devices such as mobile phones, social media and lately also a tremendous pool of IoT-devices are contributing to this large volume of data. To deal with spatial data, geographical information systems (GIS) are a good option, as they do not only deliver visualization capabilities but also tools to analyze spatial data. Considering the ever growing pool of spatial data which is generated, and which comes in a large variety of forms and formats, the capabilities of traditional geographical information systems are surpassed [1]. To deal with such a large amount of data, tremendous computing resources are required. Cloud computing represents a possible solution to this problem, since it enables the outsourcing of resources, and the on-demand usage of those computational resources [2]. The connection between those two concepts therefore is a logical step in the further advancement of GIS technology and the processing of data with spatial reference and it is one, that has already been exploited by many companies. However, this migration of large amounts of already existing datasets to cloud-based systems also raises security concerns and questions of cost-effectiveness, especially when it comes to choosing the right service providers. Amazon, Microsoft Azure and Google offer storage space solutions, which have often been used for cloud-based GIS use cases, while Esri, Hexagon Geospatial, Carto and Mapbox offer GIS specific SaaS solutions. The various offers of the different vendors vary in different parameters, which is why the suitable solution for each project should only be chosen after having conducted a requirements analysis. In the future, concepts such as big geospatial data and smart city concepts will most likely further strengthen the demand for cloud GIS.

Contents

2.1	Introduction	31
2.2	Geographical Information Systems (GIS)	31
2.2.1	Role and Use of GIS in different Sectors	32
2.3	Cloud Computing	33
2.3.1	Role of Cloud Computing in the Economy	36
2.4	GIS and Cloud Computing Integrated Architecture	37
2.4.1	Cloud Computing Service Models	38
2.4.2	Cloud Computing Deployment Models	39
2.4.3	Challenges of GIS to Cloud	40
2.4.4	Action Plan for the Migration	41
2.4.5	Advantages and Disadvantages of GIS in the Cloud	41
2.4.6	Discussion of GIS in the Cloud	42
2.5	Cloud Providers for GIS	43
2.5.1	Proprietary Web GIS Providers	43
2.5.2	Open Standards and Spatial Software for Web GIS	49
2.6	Case Studies	51
2.6.1	Development of a Cloud-based Web Geospatial Information System for Agricultural Monitoring Using Sentinel-2	51
2.6.2	Towards an Intelligent Integrated System for Urban Planning Using GIS and Cloud Computing	53
2.6.3	The Design of an IoT-GIS platform for performing automated analytical tasks	54
2.7	Future of Cloud GIS	56
2.8	Conclusion	57

2.1 Introduction

The efficient and proper management, processing and analysis of data is one of the biggest challenges in today's society. The amount of data that is being collected and stored is unimaginably high and an overwhelming part of the data structure is not even standardized. One type of data that is being stored, managed, processed and analyzed is spatially dependent data, which refers to coordinate-referenced information. This kind of geospatial data is collected through numerous approaches such as the traditional satellite systems, state-of-the-art IoT-devices or ground based laser scanning [3]. When it comes to analyzing geospatial data, usually some form of geographic information system (GIS) is involved. The increasingly large amount of spatially dependent data that is handled through a broad band of users projects major challenges for the traditional on-premise GIS. The requirements of present-day requests and tasks surpass the processing power and storage capabilities of on-premise GIS solutions. Cloud computing might be the answer to some of the current challenges in the GIS realm. Outsourcing the required resources by migrating the geographic information systems to the cloud opens up a new dimension of possibilities when it comes to processing and storing spatial data. The resulting possibilities through flexible usage of IT resources require a suitable billing model for the user's demands [4]. To understand the economical aspects of the migration of a traditional on-premise GIS to the cloud, the different GIS cloud providers and their pricing lists need to be compared.

The aim of this paper is to illustrate the economical aspect of the migration of GIS to the cloud. To this end, the concept of geographic information systems will be introduced and their applications and capabilities will be showcased to deliver a better understanding of what kind of computational system is migrated. This is followed by the explanation of the cloud computing concept, whereby the defining key factors of the cloud will be highlighted, since these factors are the main criteria for a migration in general. Furthermore, the cloud architecture and its components such as the different deployment models will be investigated. To start into the more economic side of this report, different cloud providers (e.g., Amazon AWS, Google Cloud, and Microsoft Azure) are investigated, taking into account their different services models. Next, the most well-known GIS cloud providers are introduced in detail. In addition, a discussion of such models is provided in order to highlight the advantages and drawbacks of the migration of GIS to the cloud. Finally, different case studies are conducted and discussed to show different real-world applications of cloud GIS, thus providing an outlook on the future trends of cloud GIS.

2.2 Geographical Information Systems (GIS)

A geographic information system is a tool to fulfil various operations such as capturing, storing, manipulating, analysing, managing and presenting geospatial data. The key word here is geospatial data, what refers to data which is coupled to a coordinate system and therefore is spatially dependent. This indicates that the data is normally referencing locations on the globe. Another thing that needs to be outlined is that GIS is also referred to GIScience. GIScience is the science behind the study field of geographic information systems and is summarizing the geographic concepts, applications and systems. The two definition of GIS go hand in hand, since one is defining the other. Consequently, GIS does generally not need to be specified, because most of the statements about GIS apply to both definitions [5, 6].

What makes GIS extremely powerful, is that the spatial data is connected to attribute data, which is usually stored in a tabular fashion. This allows the system to use the two different data types combined to excel at spatial analysis. For example, Spatial data is a

location in a defined space, such as the location of a tree in a forest, and attribute data is the corresponding attributes of the tree, such as the age, the type and the height of the tree.

As mentioned, GIS has a wide range of applications and is consequently a rather great decision making and problem-solving tool. The operations of GIS can be categorized into following six approaches [7].

- Mapping where things are
- Mapping quantities
- Mapping densities
- Finding what is inside
- Finding what is nearby
- Mapping change

The analysed geospatial data of the different operations can then be displayed in an illustrative way. Visualizing such data can be the key point to successfully provide a better understanding and allow deeper insights into the field of interest [8]. Because visualizations and examples are crucial in geography, the six categories will be exemplified to give the reader a better idea of what these operations consists of and what kind of data can be used.

2.2.1 Role and Use of GIS in different Sectors

This section provides an overview of the different operations that can be done with a GIS. Table 2.1 introduces examples for each category. These examples are important to show that a wide variety of industries are already using GIS in some form to analyze and visualize their data. These six different operation types can be combined to get even more information about the data, what leads to almost an indefinite amount of possibilities to analyze your data.

1. **Mapping where things are** (see Figure 2.1 (a)) We can map the spatial location of real-world features and visualize the spatial relationships among them.

Example: WiFi Hotspots in Mahattan, in this example you can see the location of the different WiFi Hotspots in an urban region. In the Map below, The relationships of the hotspots are highlighted for illustrative reasons (Green/yellow/orange areas).

Industry: Urban Planning/Social Behaviour.

2. **Mapping quantities** (see Figure 2.1 (b)) Quantities are mapped to show relationships between places, to see if certain criteria are met and to see where there is a maximum or minimum in the area of interest.

Example: The maps shows us the geothermal activity in each state of the United States of America. The predefined function determines the value which then corresponds to the representing color in each state. This kind of map is also known as a choropleth map.

Industry: Geology/Geothermal Energy.

3. **Mapping densities** (see Figure 2.1 (c)) Densities do represent concentrations of requested features. The density is usually a quantity normalized by the area of interest or a given sample set of features.

Example: The example shows a visualization of a drainage density map of a specified catchment area. It will allow the user to make assumptions about drainage through the illustrative map.

Industry: Hydrology

4. **Finding what is inside** (see Figure 2.1 (d)) GIS makes it possible to see what is inside of an area of interest. By creating a mask (area of interest), you will be able to determine the characteristics of the mask.

Example: The map shows us the range of long-distance missiles that could be fired from North Korea. The highlighted area represents the area which can be targeted from North Korea. Therefore, this kind of maps can be crucial for geopolitics.

Industry: Military / Geopolitics

5. **Finding what is nearby** (see Figure 2.1 (e)) There is the possibility to find out what lays within a defined distance of an area of interest. Those distances can be visualized in various forms using geoprocessing tools.

Example: The map shows an extract of Reno, Nevada, and it analysed how far someone can get from a defined location in any direction by car. The colors indicate the time stamps and show how far someone can get for each time interval.

Industry: Urban Planning / Traffic planning

6. **Mapping change** (see Figure 2.1 (f)) One of the more powerful operation is change detection over certain period of time. This will allow the user to investigate the change in a location for a sample set and then use it to make predictions for future trends or further evaluation along the line.

Example: The map shows us an extract of North Carolina. In the map we can see the temporal changes of forest areas and how they have been affected by growth and deforestation.

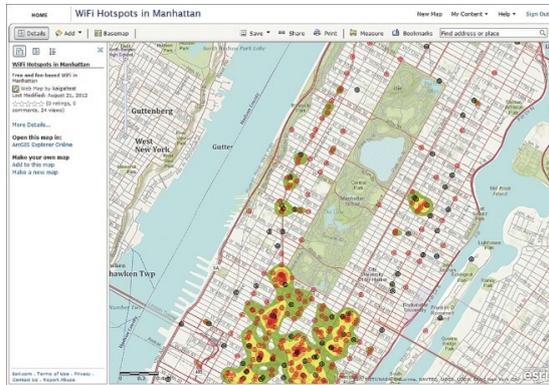
Industry: Agriculture

As those examples display, GIS can provide a broad range of possibilities for many different industry sectors. There are almost countless opportunities for different sectors to utilize GIS for further research. With its wide range of operation options, it will be able to assist almost any industry that handle and analyze any kind of geospatial data. These examples make it also clear, that a lot of industries use spatially related data in some form. This implies that GIS plays already a fundamental role and will become a potential key factor for various industries, such as insurances [15], real estate [16] and public health sector [17]. With industries as these, there will be enough financial support for further research and better understandings of spatial data and their patterns and relations.

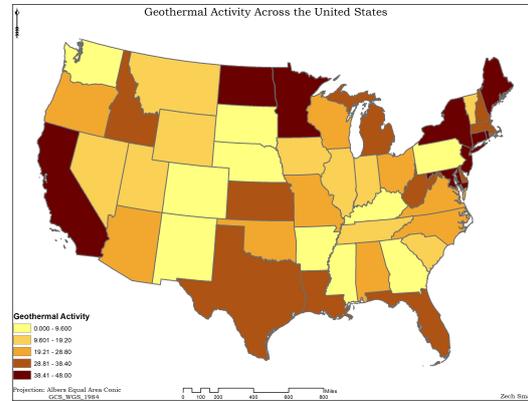
2.3 Cloud Computing

Cloud Computing are computer system resources which are available on-demand without the active management of the stakeholders. Those resources are ubiquitously accessible through networks and include computing components such as servers, storage, applications, services, networks and databases. The convenient access provides more flexible resources, faster innovations and economies of scale [18]. Five key factors define cloud computing. Such factors are described below:

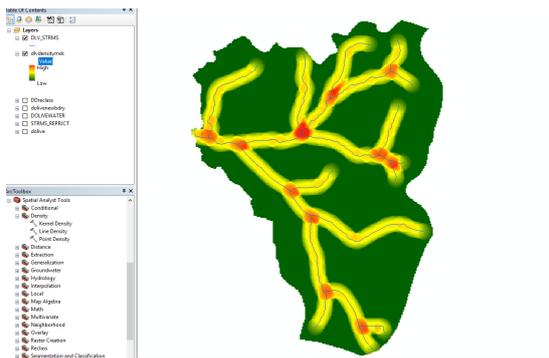
On-Demand Services Users of cloud services are able to provision server time and network storage as demanded. The whole process is completed with virtual machines such that no human interaction is required.



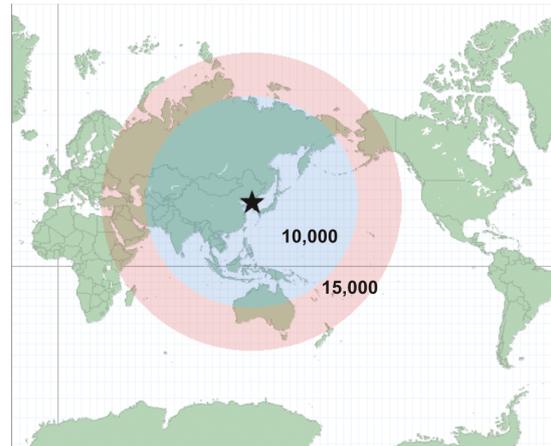
((a)) WiFi Hotspots in Mahattan [9]



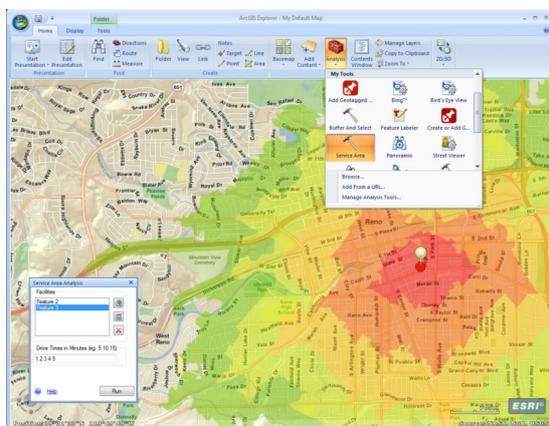
((b)) Geothermal Activity in the United States [10]



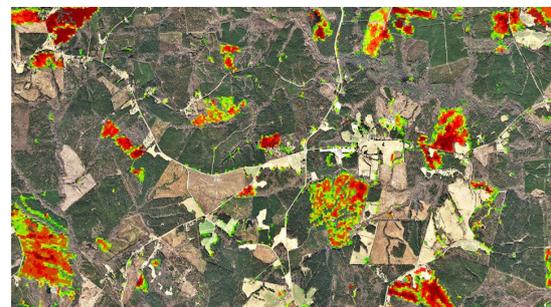
((c)) Drainage Densities of a Riversystem [11]



((d)) Geodesic Buffering [12]



((e)) Distance Raster for Travelling [13]



((f)) Forrest Change Detection [14]

Figure 2.1: Six Operation Types of GIS.

Broad Network Access BNA refers to resources hosted in a cloud network that are available for access from a wide range of client platforms such as tablets, PCs, Macs and smartphones. The access to these resources is also provided through various locations that offer online services.

Resource Pooling The computing resources of the provider are pooled for the purpose of sharing these resources with multiple consumers. This is possible with provisional and scalable services and these services can be reassigned accordingly to the consumers demand what makes them resource efficiently. These resources consist of storage, memory, network bandwidth and processing.

Rapid Elasticity Elasticity in this context refers to the degree of adaptability to the demanded workload of the consumer. Therefore, the scalability of the capabilities allows the user to manoeuvre accordingly.

Measured Service Providers of cloud services monitor and measure for transparency for both parties, the consumer and the provider. Monitoring allows different applications such as billing, optimizing resources and improve decisive planning.

On Demand stands for the self-service concept. The consumer has the option of selecting a cloud service, adapting it within a given framework and having it provided at any time. The consumer can cancel the service just as quickly and directly and return the resources used if not needed anymore.

The resulting possibility through flexible usage of IT resources requires a suitable billing model. Pay-per-use takes this into account, as costs are only incurred for the period in which the cloud service is used. This implicitly presupposes a calculation model that is based on how the cloud service is used and at the same time is economical for the provider [4]. Three aspects are noteworthy here: There is no limitation regarding the maximum amount of available resources. The provision of any quantity takes place on demand. This implies that the provision is independently and without having direct interaction with the provider. A minimum useful life is not explicitly required, but is determined by the calculation period of the pay-per-use model.

In summary, it can be stated that the five essential characteristics are mutually dependent and must be coordinated. The respective characteristics of on demand, pay-per-use and elasticity depend on how the cloud service is used by the user. For example, a service with a pay-per-use model based on a monthly calculation contradicts the principle of cloud computing of being usage- and demand-oriented if the user only uses the service for a few days within a month. He can expect at least a calculation based on days here, because otherwise it must be doubted that there is still an added value.

Cloud computing is conventionally categorized into three different service models. These service models consist of **Software-as-a-Service** (SaaS), **Platform-as-a-Service** (PaaS) and **Infrastructure-as-a-Service** (IaaS), whereas SaaS is the most prominent service model [18]. SaaS is the model where providers enable the client to have pay per use access to all computing resources. Platform as a Service provides the access to an environment which is cloud-based to the consumer. The environment offers the user the possibility of building applications and deliver the demanded output. The infrastructure is maintained by the provider itself. Infrastructure as a service is a model where the operations within a software and its application is via the web. The consumer accesses the software through the providers API or via the Internet. These different models will be extensively examined in Section 2.4.2.

2.3.1 Role of Cloud Computing in the Economy

Cloud computing is considered one of the leading IT trends for the future and this even long before it became a mainstream interest for companies, institutions and government. Researchers predict that cloud computing will have as big of an impact on the economic landscape of information and communication technologies as the world wide web once had and still has. It will restructure the way that companies and its employees engage in solving its tasks by allowing digital technologies to infiltrate all levels of economy and society. Through its ubiquitous and convenient point of access and usage it will allow individuals and small enterprises to be part of the information economy on a level matched by large companies. It is of interest to rationalize the impact of cloud computing in subjects such as security challenges, standardization of interfaces, reliability of infrastructures, software engineering processes and performance characteristics, to further understand its power and effects. The services provided by cloud computing requires a new infrastructure which facilitates the direct connectivity to the compute resources. This Internet of Services will have a huge impact on how services are being used and traded. It is said to outperform the conventional way of trading goods which will transform the economy on a national and global level by introducing a new multitude of services. This can already be seen in the consumer space, where the internet of services is already quite successfully implemented. This foreshadows on how it could be realized in the enterprise world [19]. Service providers already introduced different deployment models so far and those service models will potentially fuel the globalization process by bypassing national borders with the technology of internet. Since there are not many legislative barriers, the services being traded and hosted are being bundled and concentrated by only a hand of international companies. These companies will then use their advantageous position to reassemble the production chain by utilizing the available state of the art components to further manifest their position. Smaller companies, especially on national level, will have a limited opportunity to participate in this market and become a stable unit [20].

2.4 GIS and Cloud Computing Integrated Architecture

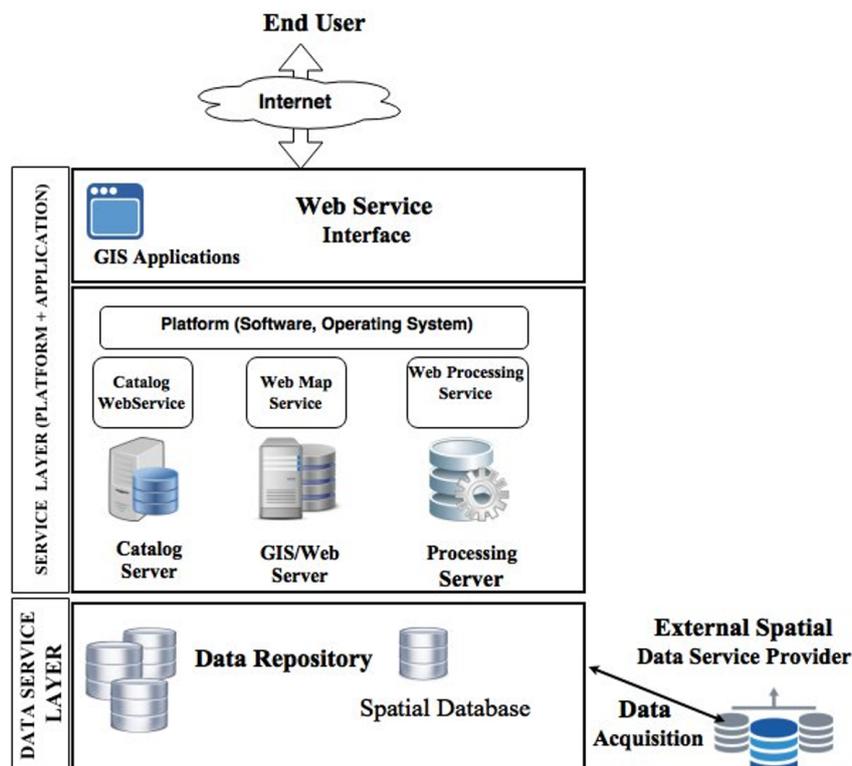


Figure 2.2: GIS Cloud Architecture [21]

The cloud computing technology has revolutionized the world of GIS in a sense, that you can combine GIS programs, servers, apps, storage and services, which has many advantages [21]. The GIS services can be run in the cloud but you can be accessed in a web browser. The combination of GIS and the cloud provides an advanced solution with broad capabilities for analyzing, collecting and publishing geospatial data. This section shows an overview, how the GIS cloud architecture looks like, what kind of service models can be used, as well as the different deployment models will be shown. Further all the benefits and drawbacks will be compared, to see, if it is more worth to do the migration in-house or have a partner by their side to implement. In the end, a simple action plan for a GIS migration to the cloud will be presented with their corresponding benefits and disadvantages.

Figure 2.2 introduces the GIS cloud architecture and its main components [21]. On the bottom, the data service layer consists of a data repository, which includes a spatial database and serves as data as a service to the other GIS services such as catalog web service, web map service and web processing service. This layer serves the data as data as a service and resides inside the cloud as an infrastructure as a service. Further this layer is connected to an external spatial data service provider for the data acquisition. The service layer, which includes the platform and the application, represents the core of the cloud infrastructure. On the one hand, it provides a web service interface to connect the end user with the data services. The GIS application in the topmost layer can be accessed by the end user using a web browser and the network infrastructure and protocols, then it consists of software as a service. On the other hand, it includes the platform with the software and the operating system. The Open Geospatial Consortium (OGC) standard GIS services such as Computer Web Service (CWS), Web Map Service (WMS), and Web Processing Service (WPS) are implemented as service over the cloud platform, where it

provides infrastructure to the GIS services. like the catalog server, GIS and web server and the processing server.

2.4.1 Cloud Computing Service Models

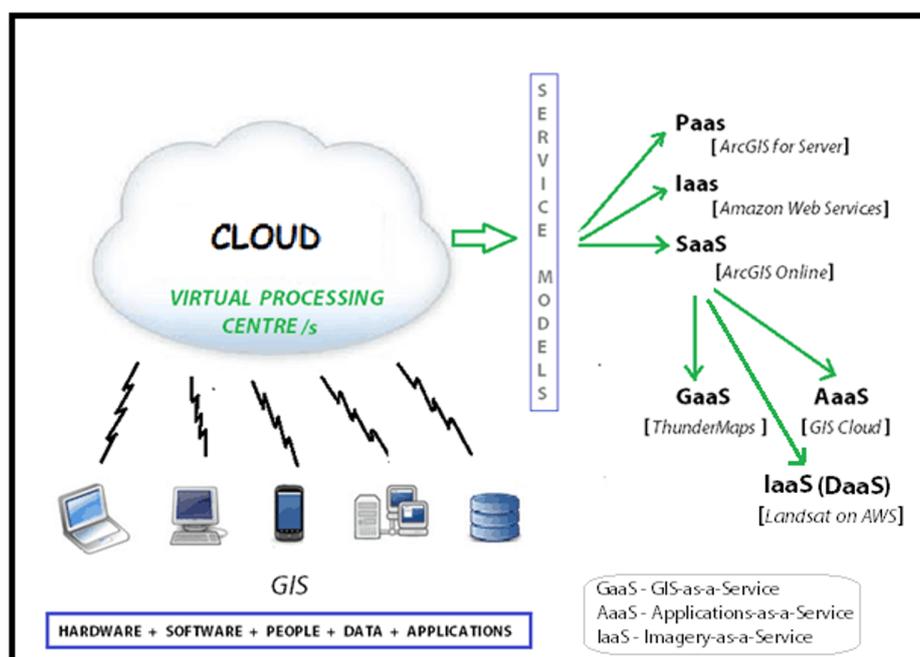


Figure 2.3: GIS Cloud Service Model [21]

In the above section we have seen the cloud architecture and the virtual processing center, which is connected by a web browser as well as network infrastructure and protocols to the end users such as computer, servers, mobile phones. In general, data can be stored in clouds but this data can have very different facets. Therefore, different service models can be chosen.

2.4.1.1 Platform-as-a-Service (PaaS)

PaaS is a type of service that provides a programming model and developer tools to build and run cloud-based applications [22]. A PaaS provider should automatically provide all the required resources such as computing power, memory, network, middleware such as message queuing or load balancing and databases when deploying the application and scale them according to the requirements. Built-in monitoring functions are also expected, with which the runtime behavior of the applications can be monitored.

2.4.1.2 Infrastructure-as-a-Service (IaaS)

In the IaaS model, service providers make their hardware equipment such as server and storage systems, operating systems and other software available to customers in a highly automated delivery model [22]. In some cases, the providers also perform tasks such as system maintenance, data backup, and emergency management (e.g., disaster recovery). Organizations can access infrastructure services on their own and usually pay according to their useful life. Users can thus put their own IT platforms into operation outside their own data center as required.

Some main characteristics of the IaaS are, that the workloads can be shifted between public and private clouds on demand thanks to a unified architecture [23]. Although all providers promise high security standards, the level of control for customers varies

widely. As a rule, all providers offer services that meet legal or other requirements. For example, data centers are often audited according to the SSAE 16 standard. [24] Further, all providers offer at least one hourly billing of virtual machines (VMs), whereby some providers also define shorter measurement steps, which can be more cost-effective for short-term batch jobs. Most providers calculate their fees accordingly on the basis of the virtual machines used. However, there are other price models on the market, some of which are more flexible. Last but not least, the monthly availability of IT services is typically 99.95 percent or higher. In managed hosting, the values are usually lower. Many providers have additional Service Level Agreements (SLAs) that cover network availability and performance as well as responsiveness to customer requests.

2.4.1.3 Software-as-a-Service (SaaS)

SaaS represents the top layer in the cloud model, where the provider provides its own applications to users [21]. This is where SaaS differs from its predecessor Application Service Provider (ASP), where service providers offered applications from other vendors for rent. These were usually not multi-client capable, while support for multiple clients is the rule with SaaS. Further software as a service contains three subgroups in the GIS cloud computing model.

2.4.1.4 GIS-as-a-Service (GaaS)

GaaS is an extension of the SaaS model, which provides GIS solutions in the cloud [22]. Mostly they offer specialized software for reporting incidents, disaster and transport management. These products and apps were often developed by many companies, where you can add different tools and modules to your software. This makes it very flexible and powerful and therefore it is expected to become a dominant method for geospatial data.

2.4.1.5 Application-as-a-Service (AaaS)

Applications as a service includes the delivery of software applications as a service over the internet [21]. This on-demand software has been gaining an increasing share on the market because the end user saves a lot of costs and also the efficiency, regardless of their size.

2.4.1.6 Imagery-as-a-Service or Data-as-a-Service (Daas)

Because the National Aeronautics and Space Administration (NASA) and European Space Agency (ESA) are very much funded by the government and therefore by the tax payer, a lot of the taken picture and dataset from space will be distributed for free over the internet as imagery as a service. Moreover, the data amount of volunteering GIS data and GIS data from the canton can be downloaded for free from webpages as data-as-a-service.

2.4.2 Cloud Computing Deployment Models

The user requirements for security, availability and possibilities of services from the cloud can differ considerably. Several deployment models meet the different requirements. Companies that want to reduce their IT costs, for example, use the Public Cloud deployment model. They work with standard applications and can do without extensive customizing. The public cloud is also suitable for companies with high demands on collaboration and mobile data access (field service, branch offices, branches, home offices).

2.4.2.1 Public Cloud

A public cloud is defined as a service that can be used by the general public or a large group and that is provided by a provider and used by the general public or a large group [25]. Private users, for example, use the public cloud like an additional hard drive to store music, video and image files. In the public cloud, companies use IT services such as e-mail systems, computer capacity, storage space, software, and communication solutions.

2.4.2.2 Private Cloud

In a private cloud, the cloud infrastructure is only operated for one institution [25]. It can be organized and managed by the institution itself or by a third party. The required infrastructure is located in the data center of the institution itself or another institution. As a rule, the IT services are only available to authorized users such as employees or business partners who access them via the Internet or Virtual Private Network (VPN).

2.4.2.3 Community Cloud

In a community cloud, companies or organizations from the same industry join together to form the community cloud from their private clouds, which are then only accessible to members of the community. Such clouds can be used wherever companies or organizations have the same requirements and tasks and want to share the existing infrastructure. The advantage of such clouds lies in reducing capacity requirements by sharing resources, reducing overhead and costs, and preparing to use other cloud services. In addition to infrastructure, many application programs can be shared by community members.

2.4.2.4 Hybrid Cloud

The Hybrid Cloud is a combination of public and private cloud and traditional IT environment [25]. For example, certain services provided by public providers run over the Internet, while critical applications and data are operated and processed within the company. The challenge here lies in separating business processes into critical and non-critical data. The prerequisite is a clean and consistent classification of the data available and processed in the company.

2.4.3 Challenges of GIS to Cloud

It requires a lot of attention from your enterprise to do a successful migration to the cloud. In case your business is not prepared to deal with the challenges of cloud migration, then it could get dangerous with your data and as well very expensive. However, you should know these challenges and how to overcome them, before you start with the migration process[26].

First of all, especially in GIS, the size of data could be very large, because in the GI-Science you often litigate images from optical satellite sensors or data clouds from Light detection and ranging (Lidar). These data also depend on the spatial, spectral and temporal resolution of the picture. As a conclusion of the large size of the data files, it is one of the main challenges in GIS how to perform compute-intensive operations such as data overlay, raster analysis and data imagery processing [22]. Additionally, it is concurrent intensive, what means the concurrent use or concurrent processing of the same data. When GIS server is faced with a large number of concurrent accesses, the application needs to use a WebGIS server cluster and a corresponding load balancing to leverage the burden of requests. The communication intensity needs to be considered as well, then

the business data is stored in a database and it is difficult to comprehend locational relationships among this data. Therefore, a visualization of the data on a map helps to show and keep track of the relations of the dataset. It helps to make better decisions based on the information you have and make efficient plans as well as improve the communication with your team members. A very important challenge is the security risk [23]. Then nowadays a lot of cyber attacks take place and hacker are placing worms and viruses on your computer, where you can get traced without even realizing it. Your data can be accessed from anyone on the internet, which can have dramatical ensues and could costs you a lot of money. Further you do not want to store your data on a foreign server because of the security risks. In case there are multiple GIS systems working individually in your company and there is no integration of the programs, it requires high maintenance costs. The last challenge will be the tools because when you develop a new product for a customer, you do not know which platform they demand.

2.4.4 Action Plan for the Migration

If you want to migrate your data to the cloud, you should follow a strict action plan to do not lose or miss any kind of your data [27]. Firstly, you need to analyze your existing infrastructure to see, if it is even worth to spend a lot of money for a migration, when an in-house hosting would be sufficient as well. So, you need to determine the issues with your current deployment and problems that may emerge in future. In a second step, you need to select the right cloud platform according to your needs and requirements. Because there are many different cloud providers with different options, should do a good market analysis and to examine the market of existing cloud architectures and see how they can fit your applications and infrastructures. You should have list of your favorites and the best would be to talk to GIS system providers directly to have a better comparison. If you had chosen a cloud provider the main step comes into effect and you can start building up your cloud space according to your needs. This involves installing, uploading data, connect to existing applications, training and testing. All this can be done internally in a company but you need to have the knowledge and the resources. Otherwise to create a solid migration strategy and moving the data to the cloud will be seamless with a reliable implementation partner by your side. The have a lot of experience in this field and on the one hand, they can do it fast and easy, but it can get expensive on the other hand. In case of thinking about to have a reliable partner by your side, which is doing the migration for you, you need to consider following catalogue of criticism [28]. The most important aspect for a company is the costs, which includes the procurement costs of different patents, materials, salaries and also the maintenance work needs to be considered. A migration of your data intercepts your current daily business because your data you are working with is not usable during the migration process and therefore quality and time is very important to have a shortest possible time-to-market procedure [29]. A further question would be, if you have internal resources available and are they good enough to do it properly. Then a worst case would come up, if you would spend a lot of time, money and resources on the migration project and in the end, it would even not be possible to transfer the data or some data will be lost. If you had chosen one GIS cloud provider you need to take his position in the market into account with his pricing policy. Moreover, the know-how into this field must be given and possible training, maintenance and support offers should be provided the supplier.

2.4.5 Advantages and Disadvantages of GIS in the Cloud

The advantages of using cloud services for companies are, among other things, the saving of considerable investments for services, that can simply be rented from external service

providers in the cloud. This saves costs in terms of both personnel and hardware [27]. This also means that no long term capital commitment is required. Costs for all the services, such as the use of hardware resources or special application software, are usually billed monthly and can therefore be easily calculated. In addition, the costs for renting a cloud solution are generally much lower, than the costs for purchasing your own hardware and software, which correspond to the desired service. The fact that hardware in particular requires regular upgrades in order to remain up-to-date with the latest technology is also a significant cost factor, that is not incurred when using a cloud service and is transferred to the respective provider. This also applies to cost intensive preventive measures for failure protection [30]. A further advantage is the saving of costs for IT specialists, who must always maintain and repair the infrastructure, since this is done by the cloud provider. Another advantage for companies is the good adaptability of the respective services. The amount of storage space, computing power and software package can thus be precisely selected and adapted to actual requirements at any time. In addition, several branches or company locations can be easily connected to the IT infrastructure used at low cost. Ultimately, the security factor can also be mentioned as an advantage. Access to resources by a company's employees is controlled by the cloud provider's administration. The same applies to access protection to computer systems.

Disadvantages for companies are, for example, dependency on the supplier, who may not take sufficient care of customers, may not be able to provide sufficient capacity or may become unable to act due to insolvency [30]. This would of course have a negative impact on all services, that a company has booked with the cloud service provider concerned and possibly already paid for. When a company uses a cloud provider. Furthermore, the question of long term storage and handling sensitive and company-related data arises. If you leave this to the care of the service provider, you inevitably enter into a certain dependency. The same applies to one's own IT competence. If one has no or only a few experts in the company, who are familiar with the technology, hardware and software, one is largely dependent on the performance of the cloud provider. The quality of the Internet connection can also be critical. Especially in rural areas, high-speed Internet has often not yet arrived. And even with a 400 Mbit line, the high bandwidth does not guarantee that the connection will always function reliably and an Internet connection is an indispensable prerequisite for cloud computing. Another danger is the use of data storage devices in other EU countries. Data storage on a server in the USA, such as Google or Amazon, is not subject to German or European data protection directives. A further disadvantage can be that existing work processes in the company have to be adapted to the software solution provided by the cloud provider, if the own or previously used software is not offered.

2.4.6 Discussion of GIS in the Cloud

All in all, cloud computing offers you a wide range of options for accessing IT infrastructure and web applications cost-effectively and efficiently. This relieves the strain on your IT department, allowing you to concentrate on your core business and also save money. This gives you an important competitive advantage. However, you must have fast Internet access and accept that your data may not be completely secure. If you can strike an acceptable balance between these advantages and disadvantages as a private individual or as the person responsible for a company, I think cloud computing is a highly recommendable solution. Some drawbacks may be negligible, especially when compared to the weight of some benefits and what they bring. Ultimately, of course, everyone has to decide for themselves whether and to what extent it makes sense to go for a cloud solution and with which deployment and service model they want to work with.

2.5 Cloud Providers for GIS

The US national Institute of Standards and Technologies defines cloud computing as a model for the enablement of a convenient, on-demand network access to a shared pool of configurable computing resources, which can include networks, servers, storage, applications or services that can be rapidly provisioned and released with minimal management effort or service provider interaction [18]. This definition allows roughly for at least three of business models already defined in this report: IaaS, PaaS, and SaaS.

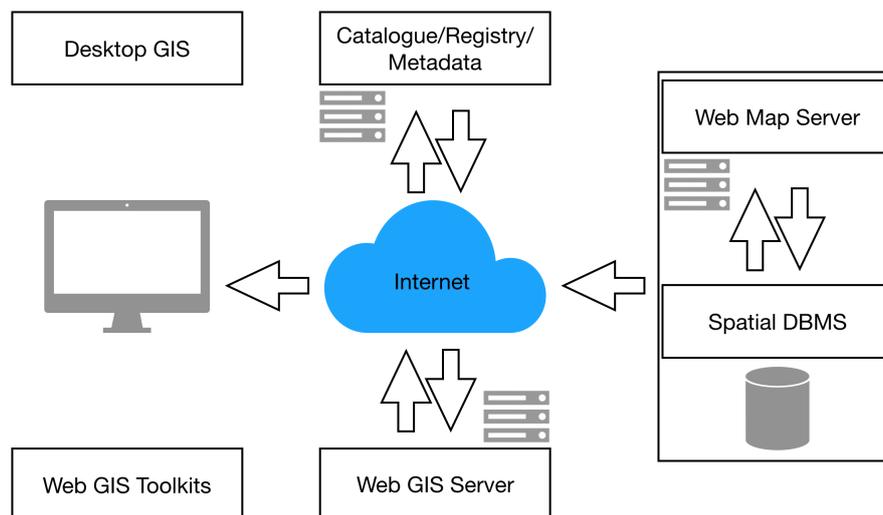


Figure 2.4: Software needs of a web-based spatial data infrastructure, based on the model of Steiniger and Hunter [39]

Looking at typical GIS cloud architectures proposed by various studies, such as the one in Figure 2.4, the different components that are important for implementing a complete Web GIS become clear. In order to get a GIS on the cloud and running, data storage has to be taken into consideration, since spatial data typically has a tendency of requiring a lot of storage space, especially raster data. But a GIS does not solely exist of geospatial data, much more it is the combination of this data with a platform, services and a collection of tools to capture, analyze, manage and visualize data. Thus, there are different implementation possibilities, where a company that decides to migrate their GIS infrastructure to the cloud can make use of one or several X as a service, depending on their own skills and requirements. In addition, open source and free software has always been an important topic in the GIS field, so companies also have to choose between proprietary and free and open source solutions that are available on the market, especially when it comes to SaaS.

2.5.1 Proprietary Web GIS Providers

Proprietary web GIS providers offer their customers different cloud services against payment. From the customer's point of view, they may not always be the cheapest solution, but often the simplest and fastest. The additional costs incurred by the use of these services can therefore often be compensated by the reduced costs for own infrastructure as well as for maintenance and management of the system. In addition, a certain part of the responsibility for the permanent operability of all systems can be transferred to the providers of these X-as-a-service solutions, for example through service level agreements. Many companies therefore decide to use offers from proprietary web GIS providers. This section discusses the different types of X-as-a-service in the context of Cloud GIS and the major associated vendors and their offerings from an economic perspective.

2.5.1.1 IaaS and PaaS

The most important cloud infrastructure providers in the domain of cloud GIS are clearly Amazon AWS [32], Microsoft Azure [33] and Google Cloud [34]. Ziani and Medouri [31] discussed the use of cloud computing for geographic information systems in their study and compared those three provider's offers, based on the computing powers of their virtual machines. The result of this comparison is shown in Table 2.1. It can be seen that Amazon, Microsoft and Google provide customers with very similar offers, when it comes to number of instance templates, CPU, memory limits, and number of supported operating systems. All of them offer autoscaling and dynamic size change. Additionally, they all offer service level agreements, which are very much alike, and they all use a pay-per-use pricing model. It is possible to integrate virtual machines of the Google, Azure and Amazon cloud systems into a Hybrid Cloud model and each of the three providers offers a specific solution in collaboration with different business partners to ensure an optimal cooperation between private and public cloud of the customers. All of the three providers further offer platforms as a service, where potential GIS applications could be built upon. Therefore, there seems to be little difference in the parameters that could lead a customer to choose one of the three suppliers and not the other. One of those differences could be the number of predefined instance templates, of which Google provides much fewer than both Amazon and Microsoft. Google, on the other hand, is the only company that chooses to enable custom instance creation, which can potentially be very valuable in getting exactly the services needed for a particular project, and nothing else that might be unnecessary and too expensive. Additionally, according to Table 2.1, Microsoft Azure also offers higher memory limits for projects with large datasets and therefore very high requirements concerning the storage space needs, but Amazon on the other hand is much stronger in the allocation of temporary storage. However, temporary storage is arguably not very useful for most of the standard GIS projects, since it is volatile and data that is saved in the temporary storage is lost every time the data is moved to a different VM due to different reasons. Nevertheless, temporary storage often comes for free and in the context of GIS applications could possibly be used to store intermediate products from analyses, that are only needed for one calculation anyway.

Table 2.1: Comparison of the computing powers of the virtual machines of Amazon, Google and Microsoft, assessed by Ziani and Medouri [31], extended in the scope of this report

Challenge	Amazon EC2	Google CE	Microsoft Azure VM
Number of instance templates available	39	18	40
GPU acceleration	Yes	No	No
Custom instance creation feature	No	Yes	No
CPU limits	1-40 CPU	1 Shared - 32 dedicated CPU	1 - 32 CPU
Memory limits	0.5-244 GB	0.6-208 GB	0.75-448 GB
Temporary storage limits	Up to 48 TB (Multiple Disks)	3 TB	2 TB
Number of OS supported	11	9	9
Number of databases supported	5+	3	3
Autoscaling	Yes, clone building	Yes, clone building	Yes, presettable group
Size change	Available	Available	Available
Service Level Agreement (SLA Terms)	Credit for 1+ minutes downtime, max monthly credit: 30%, uptime SLA: 99.95%	Credit for 5+ consecutive minutes downtime, max monthly credit: 50%, uptime SLA: 99.95%	Credit for 1+ minutes downtime, max monthly credit: 25%, uptime SLA: 99.95%
Hybrid Cloud integration	Yes	Yes	Yes

Since their offers and business models are very similar, those three cloud providers are in a constant battle for the first place in the cloud computing field, which is also expressed very openly on their webpages. For example, Microsoft Azure advertises at the top of its pricing page that an equivalent offer from amazon web services is five times more expensive than their own offer. They also claim to adjust their prices at the same time as Amazon, so they remain lower. Actually, it is really difficult to find out if that is true, because there exist incredibly many different configuration possibilities for virtual machines and storage instances at all of the providers, concerning computing power, but also storage redundancy and network bandwidth, so that a direct comparison is very hard to make. Furthermore, the cloud computing field is evolving at a very high speed, and all of the three providers constantly have to adapt and broaden their offers in order to stay in competition. However, Microsoft Azure based their claims on the fact that their customers don't have to buy additional windows licenses for virtual machines, which clearly is a benefit if windows already is the operating system potential customers are working with. Therefore, when it comes to choosing the right cloud infrastructure provider for the migration of a GIS to the cloud, in general this largely depends on the types and amounts of data working with, since 3D and raster data clearly need a lot more storage than simple 2D vector data. The three IaaS providers Amazon, Microsoft and Google

however have very similar offers and performance parameters, which is why all of them seem to be suited to use in any GIS application that is moved to the cloud.

2.5.1.2 SaaS and GaaS

For a company to build their own GIS applications on top of an infrastructure and possibly a platform as a service requires a lot of computer science knowledge and several employees who only concentrate on building and maintaining these applications. However, most of the big companies that use GIS in their daily business are administrations, engineering offices, banks, insurances, NGOs or retail chains, that are often not specialized in these topics. Additionally, employees trained to work with GI systems don't usually have this kind of computer science background. For this reasoning, many companies have been using desktop GIS solutions provided by different GIS software providers, many of which have now started to also offer software - or in this case GIS - as a service solutions. This typically includes the software and applications, but also a platform and a data storage and service hosting solution and sometimes also the provisioning of data as a service. GIS as a service is therefore often delivered as a whole package providing all of the components and services that Steiniger and Hunter [39] included in their model for a spatial data infrastructure. It is convenient for companies, as they don't need to worry about interoperability of different components and they get everything needed for their GIS applications from only one provider, which is why this is a solution that many companies choose.

Four of the most important GIS as a service providers are Esri [35], Hexagon Geospatial [36], Carto [37], and Mapbox [38]. Starbucks, Facebook and Snapchat are only a selection of the countless companies that have decided to use the services of one of these four companies. Table 2.2 shows some specifications and differences in their respective business models. It can be seen that three out of the four top players in the cloud GIS field presented here already offer a desktop GIS solution. Esri and Hexagon, whose software is not open source, are committed to continuing the development of their desktop solutions, also offering support plans.

Table 2.2: Comparison of the Cloud GIS SaaS solutions of three of the most important players in the Cloud GIS field

Provider	Desktop Solution	Pricing	Pricing model	User roles	Data Storage
Esri	Yes	Yearly fixed price for individuals or teams	License model: <i>Individual:</i> Creator <i>Teams:</i> Creator, Viewer, Field Worker, Editor, GIS Professional	Different roles: As many as you want, rights for a user role can be precisely defined by an administrator	Own Server, offers possibility to connect to others, flexible storage, which consumes credits, but in theory no limits
Hexagon	Yes	Not openly available, based on the number of products used	Only solutions and apps that you chose	-	Own Server, offers the possibility to connect to others
Carto	No	Monthly fixed price: <i>Individual:</i> 199\$ if billed annually, 299\$ if billed monthly <i>Enterprise:</i> Depending on size of the company and their requirements	Fixed possibilities that come with the individual or enterprise solution	Different roles: Owner, Admin, User	Own Server, offers possibility to connect to others, fixed storage, thus if full you have to delete elements or upgrade your account
Mapbox	Available, but no longer in active development	Mostly user-based, depending on number of requests and loads of maps, data and services	Very flexible, almost solely based on monthly users and products	Different roles: Root, Admin	On Amazon AWS

Mapbox recently decided to stop the active development and support of their desktop GIS solution, however it can still be installed from source, as the code is openly available. Their decision to proceed their business only supporting their web-based GIS seems to be a clear sign of the direction in which the entire GIS field is currently moving. Carto does not provide a desktop solution at all and instead advises on their website to make use of an open source desktop GIS for operations that cannot be done in their cloud solution. Only Mapbox and Carto display the prices for their offers openly on their web pages, whereas in the case of Hexagon, you have to fill out a form declaring what specific products you are interested in, in order to get pricing information. Esri links its pricing information directly to the creation of a user account because it works with multiple distributors around the world and customers must first be assigned to one of them. However, their

pricing model is available for the public. Like Carto, Esri offers fixed prices for solutions for individuals or teams. Their pricing model is based on licenses for the different user types Creator, Viewer, Field Worker, Editor or GIS professional, which allow for the use of different applications and services. For a subscription of an individual this is per default the license type 'Creator' that can generally make use of all functionalities Esri's cloud solution ArcGIS Online offers and can publish as many web maps and apps as they want. Carto uses a model that comes with fixed prices and permissions depending on whether the subscription is used by an individual or enterprise. They specify limits for individuals when it comes to data storage and publication of public web map and apps, so they are only allowed to create up to 10 public maps. For enterprises there are no such limits. Hexagon charges specifically for the apps and solutions that you choose and does not seem to generally make use of a bundle-based pricing model. As in almost all aspects discussed, Mapbox differs from the others in the pricing model they use. They charge for sessions and users, support plans and APIs rather than for clearly defined applications, storage space or functionalities. Monthly prices their customers have to pay can thus vary greatly depending on the number of users who used any Mapbox services within apps that the user created.

Since all of the four cloud-based GIS solutions discussed here also offer data management, it is also possible to define the rights a specific user has on the platform. In the case of Esri it is possible to define as many different user roles as you like with very specific rights, since administrators are allowed to create their own user roles and don't just have to use predefined roles. Carto on the other hand provides three different user roles - owner, administrator and user, whereas Mapbox only works with two roles, root and administrator.

Three of the four providers offer storage space on their own servers, however, all of them also come with the necessary prerequisites to connect their cloud GIS software to another infrastructure provider, like Amazon, Microsoft Azure or Google. However, it is often not possible to select an IaaS vendor that the customer wants, but SaaS vendors have specific contracts and partnerships with some of them that enable optimal interoperability. If the storage space of the GIS software provider is used, there is a fixed, non-dynamic limit of data storage that can be used in the case of Carto. If this storage space is full, data has to be deleted, or another plan must be chosen in order to be able to proceed with the work. Esri uses credits as a kind of virtual money which come in a fixed amount with the purchase of a named user and they are renewed every year, but there is the possibility to buy more if needed. Based on the amount of data that you host on the servers of Esri more or less credits are subtracted from your account daily. This principle therefore allows for a dynamic storage space allocation. Unlike the others, Mapbox stores all its customer's data directly on Amazon's cloud solution and does not provide storage space on their own servers. Capacity planning is not necessary while working with their solution, since dynamic storage space allocation is performed automatically and pricing is regulated through the amount of users and the use of specific services rather than the amount of data and maps stored in the database.

When it comes to choosing the cloud GIS solution that is best fitted for a project there are different factors that need to be considered. First of course a company has to perform some kind of requirements analysis to get an overview of what is actually needed for their projects, since this greatly influences what provider is best suited for their purposes. According to these outcomes, a SaaS provider can be chosen depending on their offers and prices. Carto is powerful and provides good solutions for general cloud GIS computing, which is why it is perfectly suitable for a lot of use cases. Their billing system is very clear and fixed, so you always know what functionalities are included and how high the bill at the end of the month is going to be. However, Carto does not provide their customers with the possibility to allocate storage space on-demand which can be a problem when

large datasets like raster, 3D or big geospatial data and rapidly changing amounts of data involved in a project. Furthermore, Carto does not provide a desktop GIS software, which is why open source software needs to be used for calculations and operations that are not supported in Carto's web-based GIS. This could potentially cause problems of interoperability between the two systems, which could slow down daily workflows. The offer of Carto thus seems to be especially useful for small or medium sized projects with a clear scope and where internet connectivity can always be assured while working. Esri and Hexagon both offer a desktop GIS software and are known to offer very similar applications and services in general. A downside to Hexagon for sure is that it really is incredibly difficult to get information about the pricing, licensing and storage model they use. This can really slow down the process of getting started with a project, since unlike Carto, it is not possible to buy anything directly on their website and it is necessary to contact Hexagon's sales. However, contrary to both Esri and Carto, Hexagon Geospatial offers the possibility to buy exactly the products needed and not necessarily a whole bundle, which could make this offer the best option if a cloud GIS is planned for a very specific field or application. Esri makes it a little bit easier to start with a cloud GIS, by offering their subscriptions online. However, customers have to create an account before they get any pricing information. They offer their cloud products as a bundle by using different licensing types. User roles can be defined, and dynamic storage space allocation is possible using credits, which makes it possible to handle varying amounts of data. Additionally, there is a desktop solution available, which makes data transfers between local computers and the cloud very easy. The solution thus seems to be able to handle almost all kinds of projects and data. However, the bundle based pricing model that is used, might make this solution in general more expensive than others and customers might pay for more than they actually need. Therefore, this offer is probably the best if also the desktop solution and many of the provided functionalities and apps are needed for a specific project. Mapbox works with a very different approach than the previous three providers and probably their target customers are not the same either. Like Carto, also Mapbox does not really provide a desktop solution anymore. Even if there is the possibility to download the source code to their traditional desktop GIS and use it in projects, it is not in active development and Mapbox does not provide any support on it, which is why this would most likely not be a very reliable solution to use for important projects. The strengths of Mapbox can rather be found in their web services, which are perfectly suited for very specific applications where a lot of users interact with maps and data which is why also their pricing model is user-based. Possible use cases where this might be the best Cloud GIS solution are therefore more like those of Facebook or Snapchat, both of which are Mapbox customers.

2.5.2 Open Standards and Spatial Software for Web GIS

This essay does not explore into detail the different free and open source spatial software (FOSS) solutions there are. However, the use of this kind of software is very common in the field of GIS, especially when it comes to projects that are conducted on a very tight budget. This could for example be the case in the scientific field, but also in the field of NGOs and similar organizations that for instance want to make the advantages of an openly available web-based GIS accessible to the general public in developing countries.

Table 2.3: Different free and open source software components needed for a complete SDI, collected by Steiniger and Hunter [39]

Desktop GIS	Web GIS Toolkits	Catalogue/Registry/Metadata	WebGIS Server	Web Map Server	Spatial DBMS
gvSIG uDig OpenJUMP Quantum GIS GRASS MapWindow	MapBender GeoExt MapFish GeoMoose GeoMajas SharpMap	GeoNetwork MDWeb deegree	52 North WPS PyWPS deegree 3	MapServer GeoServer MapGuide OS deegree QGIS server FeatureServer GeoRest	PostGIS SpatialLite

Since a database in the cloud must be used anyway, existing FOSS software components can be used in conjunction with an IaaS vendor solution as described in Section 2.6.1.1. Steiniger and Hunter [39] defined six components of a complete spatial data infrastructure (SDI), which can be seen in figure 2.4. In their study, they also explored the corresponding different free and open source software components available to accomplish these tasks. A list of those services is given in table 2.3.

The reason why it is possible to use so many different software products for datasets as complex as spatial data, which include not only data, but spatial indexes, very different data structures and implicitly stored visualization attributes, is because there is a large set of standards for those kind of data, which was defined by many different stakeholders in the Open Geospatial Consortium (OGC). This allows a relatively easy interaction between not only different free and open source software products, but also between those products and proprietary SaaS providers for GIS software, as many of them have also integrated the common standards in addition to their own standards.

2.5.2.1 Open Geospatial Consortium (OGC)

The OGC is an international consortium of more than 530 stakeholders from very different industries. Their goal is to make geospatial information and services FAIR (Findable, Accessible, Interoperable and Reusable). For this purpose, many standards of varying nature have already been defined, including formats for storage and transfer of geospatial data. As a result of the continuing shift of GIS environments to the cloud, a subgroup called the Open Cloud Consortium was formed in 2010 [40]. OCC is specifically dedicated to a focus on geospatial content sharing, the integration of geospatial services into different infrastructures and other specifically GIS cloud related topics. McKee et al. argue that standards play an even more important role in the cloud GIS environment, than in traditional systems, since they help enabling links through interoperability and choice. They provide customers with more flexibility, since cloud providers can be changed without great efforts. On the other hand, the cloud providers themselves can also benefit from common standards, as they are able to meet the diverse needs of their customers and have the ability to work easily with their competitors, for example by shifting loads to their systems when problems arise in their own systems [40]. The standards defined by the OGC are therefore in fact used by many.

2.6 Case Studies

In this section case studies will be presented, which relate to the topic of GIS and cloud computing. The aim is to give an overview in which domains GIS-Clouds are applied, and how they are implemented. The topics which are discussed in the three papers are agricultural monitoring, urban planning, and smart cities. The latter will give a good transition to the next section, where future trends are elaborated.

2.6.1 Development of a Cloud-based Web Geospatial Information System for Agricultural Monitoring Using Sentinel-2

The paper 'Development of a cloud-based web geospatial information system for agricultural monitoring using Sentinel-2' was written in 2018 by Hnatushenko, Sierkova and Sierkov. The authors present a system architecture of a cloud-based web GIS, which is used to support agricultural monitoring by providing a web mapping application [41].

As agricultural monitoring is an important technique for the targeted use of security and for guaranteeing food security, systems supporting such tasks are very valuable. With remote sensing imagery, which in this case study is delivered from the satellite Sentinel-2, not only vegetation indices can be calculated, but also land cover classifications can be performed [41]. By using vegetation indices, disease and water stress can be detected, and the nutrient status of plants can be assessed. Such information is helpful for farmers to use resources like water, fertilizers and disease control agents only at those places, where they are needed by the plants. This helps in saving resources and detecting disease at early stages, which helps to prevent farmers from big harvest loss. Furthermore, biomass estimations and yield predictions can be created by using satellite images, which are important calculations for guaranteeing food security. Another useful application is the landcover classification of the different plant types [42].

Hnatushenko, Sierkova and Sierkov have implemented a system to support agricultural monitoring tasks, which is based on a dynamic web map tile service (WMTS) with a geospatial processing service (GPS) which result in a web GIS that is deployed on Amazon Web Service (AWS). They proposed a system architecture which is shown in Figure 2.5. The most important parts of this architecture are the input data, the WMTS and the GPS. As input data serve crop type maps to visualize and Sentinel-2 time series to analyze the data. The WMTS is responsible for rendering raster images. Due to the dynamic rendering, only the fractions needed have to be kept on disk, which reduces the computational resources which are required. The GPS delivers the geospatial processing functionality, as statistical calculations for a user-defined area of interest, which allow calculating e.g. vegetation indices.

Those components result in the web mapping application, which allows for interactive use of the map tile visualizations and the geoprocessing services. The web mapping application enables to search and view crop type maps and Sentinel-2 imagers, to draw an area of interest and make statistical calculations for this area and to visualize the derived calculations in a plot. This can be seen in figure 2.6, which shows the interface of the final web mapping application.

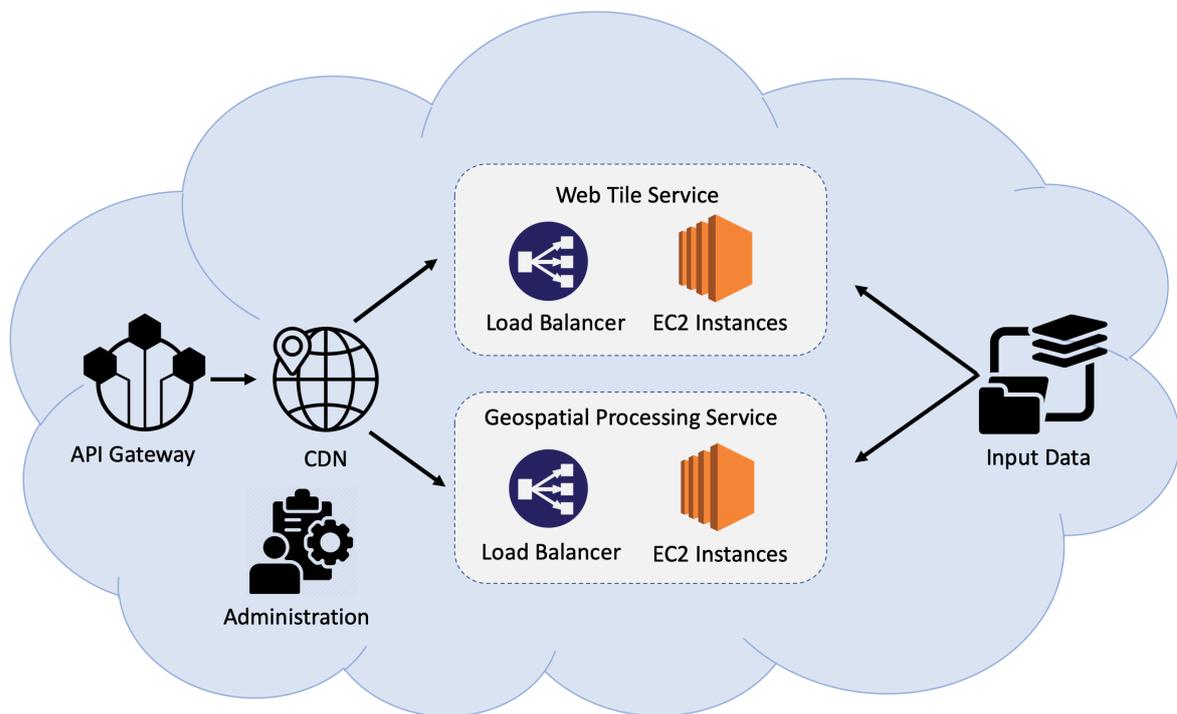


Figure 2.5: System architecture of cloud based web GIS [41]

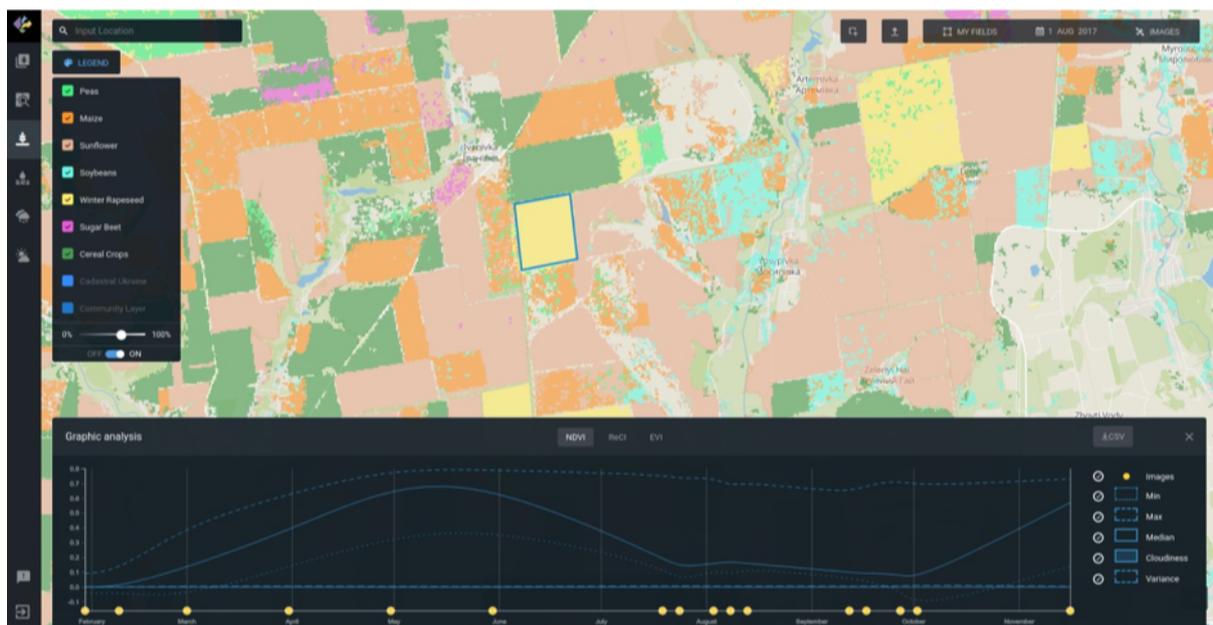


Figure 2.6: Interface of the web mapping application [41]

The system provided in this study is providing a potential solution to the problem of the large-scale data that is used for agricultural monitoring, for which calculations are complex and computationally intensive. By using a cloud-based solution, individual farmers are able to outsource the cost of maintaining professional software and hardware for processing such a large amount of data. Otherwise, it would not be possible for smaller farms to benefit from the results of agricultural monitoring data [41]

2.6.2 Towards an Intelligent Integrated System for Urban Planning Using GIS and Cloud Computing

The paper 'Towards an Intelligent Integrated System for Urban Planning Using GIS and Cloud Computing' was written in 2018 by Khelifa, Laoar and Eom. They present an intelligent integrated system, which supports planners and decision-makers of the urban planning domain. An intelligent decision support system should enable them to deal with different components of different urban planning fields that interact with and affect each other. As the process of decision making often includes conflicting objectives, and the amount of data that needs to be processed is huge, a combination of GIS with its analyzing capabilities and cloud computing with its high computational resources is a good approach to deal with these problems. They want to provide a system that enables them to easily extract urban needs by predicting requirements of urban development projects and to fit available projects to appropriate locations [2].

More than 50 % of the total population is living in cities, and by 2050 this percentage is predicted to increase up to 70 % This population growth in urban areas challenges sustainability and resilience of cities and introduces a bunch of problems [43]. Urban planners have to deal with conflicting land-use requirements (e.g. they have to provide enough living space, at the same time provide infrastructure and take ecological factors into account). They also have to ensure that every household is supplied with energy, food, and water, and must ensure the availability of transportation networks [2].

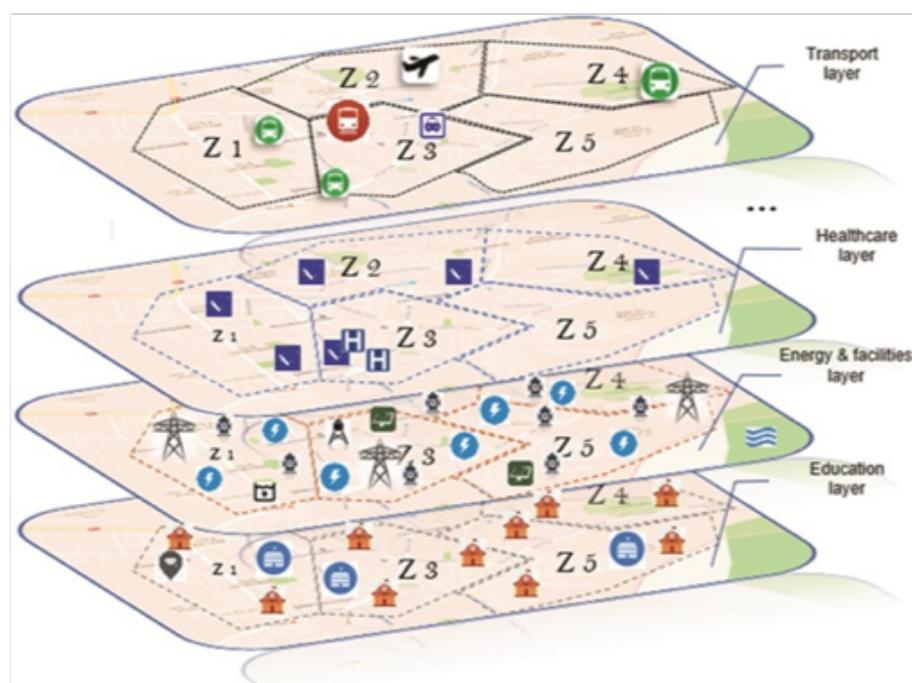


Figure 2.7: System architecture of cloud based web GIS [2]

To deal with the different domains of urban planning, each urban field is displayed on an individual layer. Furthermore, different regions of urban areas are divided into different zones, where each zone is defined by its own urban data. The zones are defined regarding geographic features, socio-economic factors, and administrative partition plans. To ensure the integrity of urban systems, the components of all contiguous areas are analyzed field by field (horizontally) as well as area by area (vertically). This results in high computational complexity, wherefore the use of cloud computing is recommended. On the other hand, GIS-Systems deliver tools to make computations zone by zone but also layer by layer [2]. The system proposed in this study consists of three main components: The Intelligent Decision Support System (IDSS), the Geographic Information System (GIS) and Cloud

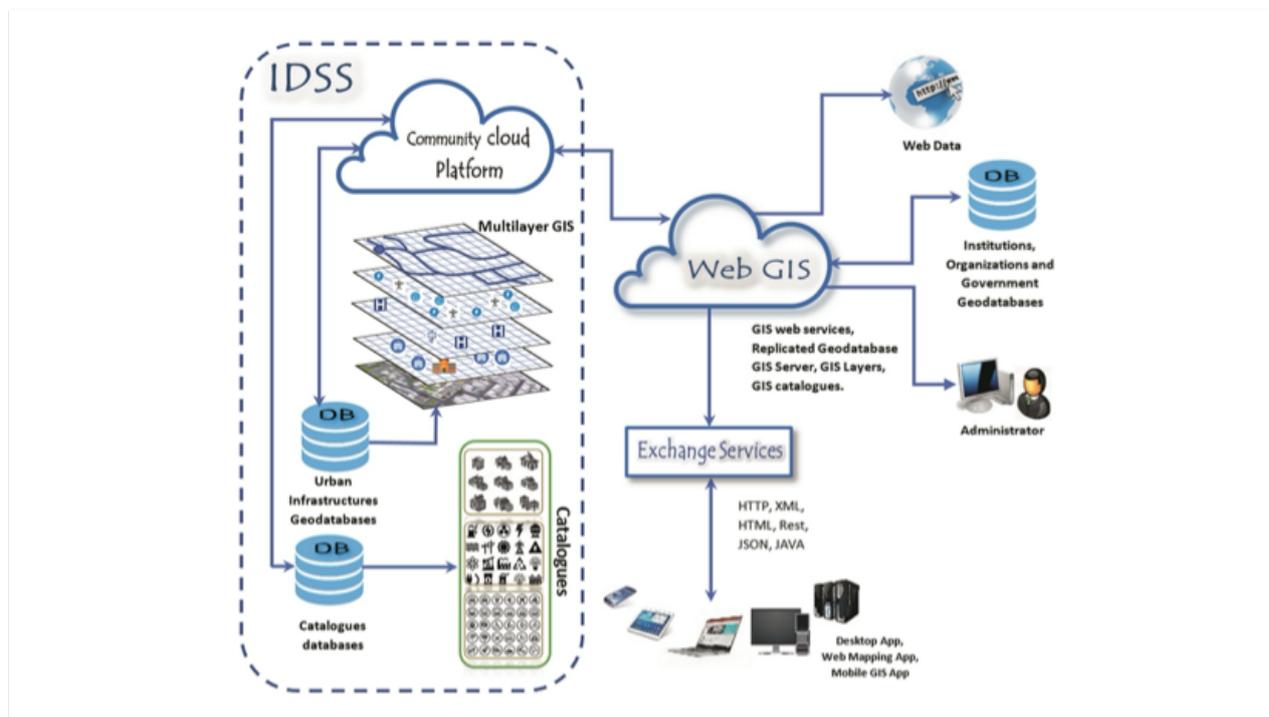


Figure 2.8: System architecture of the intelligent integrated system for urban planning [2]

Computing. The architecture of the system is shown in figure 2.8. The IDSS points out the best alternatives to reach suitable results. IT relies on a combination of artificial intelligence and a decision support system. The GIS part enables the visualization and analysis of data which allows to detect relationships, patterns, and trends of different urban fields. Cloud computing provides a pool of shared data and processing resources, which can be acquired on-demand by decision-makers. This enables urban planners of different fields to access data of several fields and to share their results on a web GIS [2].

2.6.3 The Design of an IoT-GIS platform for performing automated analytical tasks

The paper 'The design of an IoT-GIS platform for performing automated analytical tasks' has been written by Cao and Wachowicz. They present a first step towards the design of an IoT-GIS platform which should be able to perform analytical tasks, with the purpose of delivering valuable information to transit services. The goal is to design an IoT-GIS platform, which performs analytical tasks without human interaction and is capable to deal with the transportation of unbounded data streams, which requires a lot of processing power. A good way to achieve this is, by combining the IoT-GIS platform with cloud computing. The paper at the end delivers an IOT-GIS platform, which is able to perform data ingestion, data cleaning and data contextualization automatically [44].

IoT devices usually generate unbounded sequences of tuples, which often are out-of-order and have high data rates. A vast number of devices will be integrated into the fabric of smart cities, revolutionizing urban processes and their planning through e.g. the optimization of traffic flows, smart parking or digital health. To enable real-time functioning cities with routinely sensed data, a shift from traditional GIS platforms towards IoT-GIS platforms is necessary. This is the case because traditional GIS platforms are inefficient, as they require coordination of different tasks and use limited computing resources. Furthermore, most traditional platforms require the intervention of humans, which is time-consuming and leads to errors in the data. One of the biggest challenges for the IoT-GIS platform is, to include the context of mobility, which links location, date and time. There

is almost no research in the domain of moving IoT devices, but to understand traffic flows and transit network performance mobility is crucial.

Cao and Wachowicz focused on developing the IoT-GIS platform for transit agencies, which serve people that live in small urban areas. They have selected Codiatic Transit for their mobility context. They operate on 30 different bus routes, and each bus is equipped with a GPS sensor, but still, they need to go on the road to measure the pace of a route. With the IoT-GIS platform, the pace of a route could be computed accurately and automatically without the need for physically measuring it. Figure 2.10 shows how the data streams generated by the IoT devices attached to the buses are sent to the implemented cloud infrastructure, where automated tasks are executed for generating the context of mobility. Every trip is automatically created, containing information about the origin, stops, destination, street names, and its duration. This enables to display the behavior of the Codiatic Transit network [44].

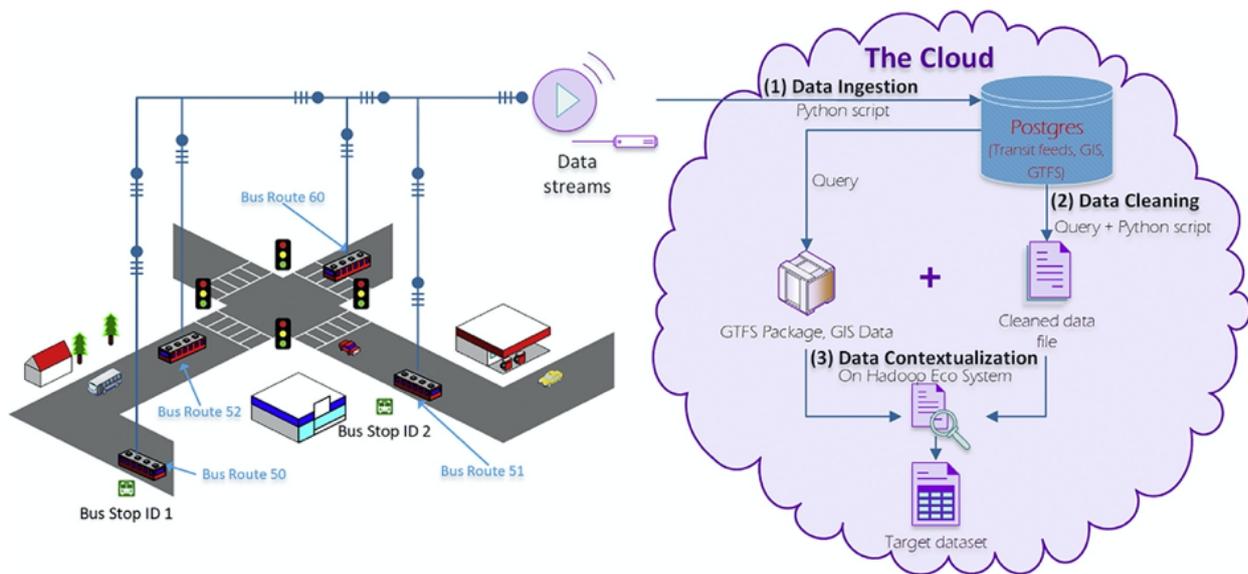


Figure 2.9: IoT-GIS platform [44]

There are three automated tasks executed by the IoT-GIS platform, namely data ingestion, data cleaning, and data contextualization. Data Ingestion pushes the incoming tuples to the PostgreSQL database. For this task, low latency is crucial, which could be minimized by minimizing the impact of disk I/O and using faster networking. Data Cleaning corrects data inconsistencies and redundancy in the tuples. The processing time of data cleaning is dependent on the bus route. Data Contextualization takes the cleaned tuples and orders them as a first step to provide contextualized tuples by using MapReduce. Then it determines if the bus was moving or if it had stopped. Then it classifies the stop or the move (e.g. running, passing, suspension or stopover). Then, the street name is annotated. An Already stored GIS layer was used for the contextualization. Afterward, geographical features are annotated e.g. tagging bus station IDs to corresponding tuples. Then street intersections, time and finally a trip is annotated. The data contextualization takes the longest processing time [44].

In the end, a summary statistic can be delivered for each route. The statistics give new insights to the traffic behavior for each route and show where improvements of routes would be appropriate. A high number of cases of movement suspensions is e.g. an indicator that for this route signal synchronization and bus priority would be an improvement. The paper has shown that it is possible to explore mobility semantics by contextualizing raw data. The IoT-GIS platform presented with its analytic tasks could potentially enable transit network services to adapt at real-time to current or expected mobility contexts, which is especially valuable for transit agencies with a limited number of personnel [44].

2.7 Future of Cloud GIS

Every day, a huge amount of geospatial data is created. Nowadays, not only satellites contribute to the pool of geospatial data, but also non-traditional devices as mobile phones, social media, and IoT devices deliver geospatial data even faster than satellites. In addition to the large volume of data, there is a tremendous variety of forms and formats of geospatial data, and the data accuracy is uncertain. This large amount of varying data is often referred to as Big Data, which on one hand is a big opportunity to extract information but on the other hand poses challenges, as it is difficult for traditional systems to deal with such large amounts of data [1]. As in the near future, the amount of data is going to increase further, solutions have to be provided, and systems have to be created that are able to deal with Big Data problems. One potential solution to deal with this large amount of data, which often is linked to spatial information, is a combination of cloud computing and GIS. While cloud computing delivers high computational resources on-demand, GIS-systems deliver the capability to analyze and visualize spatial data [2].

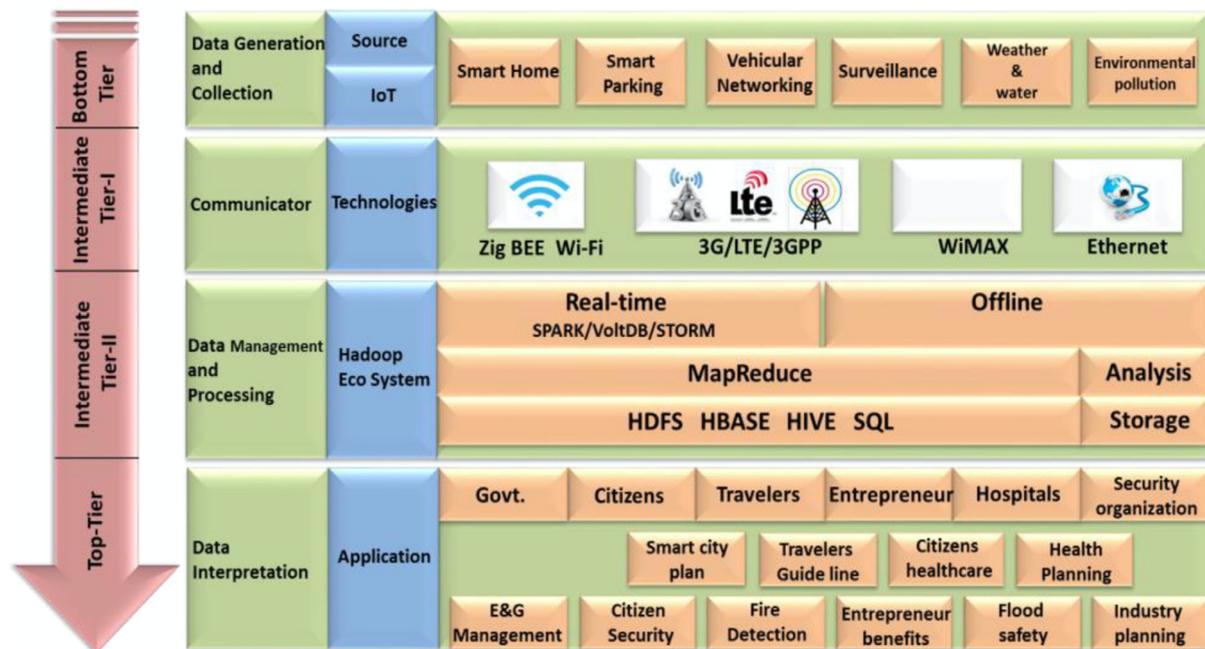


Figure 2.10: Architecture for IoT Big Data analysis for smart cities and urban planning [45]

Such combined GIS and cloud systems can be applied e.g. in smart cities, which are dominated by IoT devices and sensors which collect tremendous amounts of data to support various domains as e.g. traffic planning or public safety. In the paper of Rathore et al. a complete architecture for developing smart cities and conducting urban planning by using IoT-based Big Data analytics is presented. An overview of the architecture they presented can be seen in figure 2.10. The architecture consists of four tiers with different functionalities as e.g. the collection of data and their aggregation, communication, and processing of the data but also data analysis and interpretation. The data set consists of vehicular networks, smart parking, weather, pollution, and surveillance but also includes data from smart homes. By analyzing and interpreting the collected data, smart cities can be developed and urban planning decisions can be supported [45].

2.8 Conclusion

The large amount of geospatial data generated by millions of different devices and the associated new demands on geographic information systems are leading to a change from traditional local to cloud-based solutions. In the first section of this essay, an overview of the concepts and roles of GIS and cloud computing were presented. Then, a closer look was taken at a representative GIS cloud architecture and its corresponding service and deployment models. Furthermore, an action plan for a migration of a GIS to the cloud was proposed followed by a discussion of advantages and challenges of such a process. Depending on the amount of data and the structure of a company a cloud solution could be an economic efficient solution. Obviously there are some minor drawbacks like the data storage in foreign country or if a cloud provider does not take caution. Overall, if an internet connection is fast enough, a movement to the cloud should be considered of any GIS company, as the advantages strongly outweigh. To give an overview over different possible cloud service companies that work with spatial data could make use of, several currently dominant service providers of the cloud GIS field have been subjected to an economic analysis. To this end, the business models and offers of those providers were compared and the best suited use case for each technology determined. Currently, the most prominent IaaS providers in the cloud GIS field are found to be Amazon, Microsoft Azure and Google, whose offers appear to be very similar. Apart from the IaaS providers, there are also specialized SaaS providers who are more dedicated to the GIS field and not only provide storage space, but also offer analysis and visualization tools to process spatial data. In this field, some of the most successful players are Esri, Hexagon Geospatial, Carto and Mapbox. Project requirements play an important role in the selection of one of these vendors, as the offerings differ in several parameters, such as location, payment model, and desktop solution deployment. At the end of this section, an additional short introduction of free and open source solutions was provided and the standards defined by the OGC, which are very important for the interoperability between different systems, are addressed. To give an overview of the state of the art, three case studies were presented, which focused on different fields where GIS-Clouds can be applied. While the first study presented a web-mapping application for the agricultural domain, the second case study provided an example for an intelligent integrated system to support urban planning. The last case study focuses on the topic of traffic monitoring by using IoT devices, which use a IoT-GIS platform based on cloud computing to analyze the collected data. In all case study the benefit of cloud computing and GIS is stated, as GIS systems deliver the capability to analyze and visualize spatial data, while cloud computing provides a pool of computational resources which can be accessed on-demand.

Bibliography

- [1] Chaowei Yang, Yu Manzhua, Fei Hu, and Yongyao Jiang Yun Li: *Utilizing Cloud Computing to address big geospatial data challenges*, Computers, Environment and Urban Systems, vol. 61, 2016, pp. 120-128.
- [2] Boudejema Khelifa, Mohamed Ridda Laouar and Sean Eom: *Towards an Intelligent Integrated System for Urban Planning Using GIS and Cloud Computing*, International Conference on Decision Support System Technology, Cham, Springer 2018, pp. 26-37.
- [3] Deepak Puri: *Monitoring the Amazon wildfires with satellites, IoT sensors and GIS*; Network World, <https://www.networkworld.com/article/3434517/monitoring-the-amazon-wildfires-with-satellites-iot-sensors-and-gis.html>, last visit November 14, 2019.
- [4] Goran Novkovic: *Five Characteristics of Cloud Computing*; Control engineering, <https://www.controleng.com/articles/five-characteristics-of-cloud-computing>, last visit November 14, 2019.
- [5] Nicholas Chrisman: *What Does 'GIS' Mean?*; Wiley Online Library, vol. 3, no. 2 , March, 1999, pp.175-186.
- [6] Caitlin Dempsey: *GISLounge - What is GIS?*; <https://www.gislounge.com/what-is-gis/>, last visit November 14, 2019.
- [7] Caitlin Dempsey: *Mapping and Geographic Information Systems (GIS) : What is GIS?*; <https://researchguides.library.wisc.edu/GIS>, last visit November 14, 2019.
- [8] Hardy Pundt, Klaus Brinkkoetter-Runde: *Visualization of Spatial Data for Field Based GIS*; Computers and Geosciences, vol. 26, no. 2, 2000, pp. 51-56.
- [9] Esri: *A Workflow for Creating and Sharing Maps*; <https://www.esri.com/news/arcuser/1012/a-workflow-for-creating-and-sharing-maps.html>, last visit December 12, 2019.
- [10] University of Oregon: *Maps of Zech*; <https://blogs.uoregon.edu/zechmap/lab-1/>, last visit December 12, 2019.
- [11] GIS Stackexchange: *Understanding Drainage Density*; <https://gis.stackexchange.com/questions/248050/understanding-drainage-density/248084>, last visit December 12, 2019.
- [12] ESRI: *Understanding Geodesic Buffering*; <https://www.esri.com/news/arcuser/0111/geodesic.html>, last visit December 12, 2019.
- [13] GIS Stackexchange: *Creating Distance Raster Using Road Network ArcGIS*; <https://gis.stackexchange.com/questions/211732/>

- creating-distance-raster-using-road-network-arcgis?rq=1, last visit December 12, 2019.
- [14] University of North Carolina: Forest Change Detection Tool; <https://nemas.unca.edu/forest-change-detection-tool-southfact>, last visit December 12, 2019.
- [15] Earl Epstein: *Liability Insurance and the Use of Geographical Information*; International Journal of Geographical Information Science, 1998, vol. 12, no. 3, pp. 203-214.
- [16] Richard Reed, Chris Pettit: *Real Estate and GIS*; Book, Routledge, Abingdon, 2018.
- [17] Gerard Rushton: *Public Health, GIS, and Spatial Analytic Tools*; Annual Review of Public Health, 2003, vol. 24, no. 1, pp. 43-56.
- [18] Peter Mell, Tim Grance: *The NIST Definition of Cloud Computing*, Recommendations of the National Institute of Standards and Technology, September 2011. <http://faculty.winthrop.edu/domanm/csci411/Handouts/NIST.pdf>, last visit November 14, 2019.
- [19] Federico Etro: *The Economics of Cloud Computing*; In I. Management Association, Cloud Technology: Concepts, Methodologies, Tools, and Applications, 2015.
- [20] Ergin Bayrak, John Conley, Simon Wilkie: *The Economics of Cloud Computing*; Korean Economic Review, 2011, vol. 27, pp. 203-230.
- [21] Reem Adnan Al Kharouf, Abdel Rahman Alzoubaidi, and Mazher Jweihan: *An integrated architectural framework for geoprocessing in cloud environment*, Spat. Inf. Res., vol. 25, 2017, pp. 89-97.
- [22] Taha Alfaqih: *GIS Cloud: Integration between cloud things and geographic information systems (GIS) opportunities and challenges*, ResearchGate, 2013, pp. 1-7.
- [23] Virendra Singh Kushwah and Aradhana Saxena: *A security approach for data migration in cloud computing based on human genetics*, Lect. Notes Bus. Inf. Process., vol. 3, no.5, 2013, pp. 1-8.
- [24] Iryna Windhorst and Ali Sunyaev: *Dynamic Certification of Cloud Services*, 2013 International Conference on Availability, Reliability and Security, Regensburg, 2013, pp. 412-417.
- [25] Shivaji Pandurangrao Mirashe and Namdeo V. Kalyankar: *Cloud Computing*, J. Comput., vol. 2, no. 3, 2010, pp. 78-82.
- [26] Raouf Boutaba, Qi Zhang and Mohamed Faten Zhani: *Virtual machine migration in cloud computing environments: Benefits, challenges, and approaches*, Communication Infrastructures for Cloud Computing. IGI Global, 2014, pp. 383-408.
- [27] Artem Barsukov: *Steps to Consider Before Moving Your GIS to the Cloud*, Intellas, 2019. <https://www.intellias.com/steps-to-consider-before-moving-your-gis-to-the-cloud/>, last visit November 12, 2019.
- [28] Shareeful Islam, Stefan Fenz, Edgar Weippl and Haralambos Mouratidis: *A Risk Management Framework for Cloud Migration Decision Support*, J. Risk Financ. Manag., vol. 10, no. 4, 2017, pp. 1-10.

- [29] Rashmi, Shabana Mehruz, G.Sahoo: *A five-phased approach for the cloud migration*, International Journal of Emerging Technology and Advanced Engineering, vol. 2, issue 4, 2012, pp. 1-6.
- [30] Azam Abdollahzadehgan, Ab Razak Che Hussin, Marjan Moshfegh Gohary, and Mahyar Amini: *The Organizational Critical Success Factors for Adopting Cloud Computing in SMEs*, J. Inf. Syst. Res. Innov., vol. 4, no. 1, 2013, pp. 67-74.
- [31] Ahmed Ziani and Abdellatif Medouri: *Use of Cloud Computing Technologies for Geographic Information Systems*, In: Mostafa Ezziyyani, Mohamed Bahaj, Faddoul Khoukhi (eds), Advanced Information Technology, Services and Systems, AIT2S 2017, Lecture Notes in Networks and Systems, Cham, Vol. 25, Springer, 2018.
- [32] General information about AWS virtual machine offerings, AWS, Amazon Web Services, Inc. or its affiliates. <https://aws.amazon.com>
- [33] General information about the Microsoft Azure virtual machine offerings, Microsoft Azure, Microsoft. <https://azure.microsoft.com/de-de/pricing/>
- [34] General information about the Google Cloud virtual machine offerings, Google Cloud, Google. <https://cloud.google.com>
- [35] General information about the Esri GIS cloud offerings, Esri Suisse. <https://www.esri.ch>
- [36] General information about the Hexagon Geospatial GIS cloud offerings, Hexagon, Hexagon AB and/or its subsidiaries and affiliates.. <https://www.hexagongeospatial.com/>
- [37] General information about the Carto GIS cloud offerings, Carto. <https://carto.com>
- [38] General information about the Mapbox GIS cloud offerings, Mapbox. <https://www.mapbox.com>
- [39] Stefan Steiniger and Andrew JS Hunter: *Free and Open Source GIS Software for building a Spatial Data Infrastructure*, Geospatial free and open source software in the 21st century, Springer, Berlin, Heidelberg, 2012, pp. 247-261.
- [40] Lance Mckee, Carl Reed and Steven Ramage: *OGC Standards and Cloud Computing*, Open Geospatial Consortium, April, 2011.
- [41] Volodymyr V. Hnatushenko, Kateryna Yu Sierikova, and Ivan Yu Sierikov: *Development of a Cloud-Based Web Geospatial Information System for Agricultural Monitoring Using Sentinel-2 Data*, IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), vol. 1, 2018, pp. 11-14.
- [42] Wouter Maes and Kathy Steppe *Perspectives for Remote Sensing with Unmanned Aerial Vehicles in Precision Agriculture*, Trends in Plant Science, pp. 1-13.
- [43] Jack Ahern: *Urban landscape sustainability and resilience : the promise and challenges of integrating ecology with urban planning and design*, Landscape Ecology, vol. 28, no. 6, 2013, pp. 1203-1212.
- [44] Huang Cao and Monica Wachowicz: *The design of an IoT-GIS platform for performing automated analytical tasks*, Computers, Environment and Urban Systems, vol. 74, 2019, pp. 23-40.

- [45] M. Mazhar Rathore, Awais Ahmad, Anand Paul, and Seungmin Rho *Urban planning and building smart cities based on the Internet of Things using Big Data analytics*, Computer Networks, vol. 101, 2016, pp. 63-80.

Chapter 3

Economic Assessment of Distributed Denial-of-Service (DDoS) Attacks

Adrian Iten, Artemis Kardara, Vasiliki Arpatzoglou, Timo Schenk, Noah Berni

Distributed Denial of Service attacks, commonly called DDoS, are constantly evolving and impact the internet infrastructure. As DDoS attacks result in the unavailability of network resources for the intended user, they can lead to economic losses for businesses and individuals in various ways. A noteworthy example of DDoS attack, was Dyn cyberattack targeting systems operated by Domain Name System (DNS) provider Dyn, happened on October 2016. This kind of DDoS attack caused a considerable amount of services to be unavailable, even organizations which were not customers of Dyn provider. This report, aims to explore the economic impacts of such attacks, as well as identify and assess the risks and provide solutions for risk management. To this end, we study the risk management framework "SEconomy", developed at the University of Zurich, which provides a way to measure the economic impact of cybersecurity activities in a distributed ecosystem with several actors. Finally, we apply said framework on a case study, namely the Dyn cyberattack, and discuss the results.

Contents

3.1	Introduction	65
3.1.1	Motivation	65
3.1.2	History of attacks	65
3.1.3	Economic impact	68
3.2	DDoS Background	71
3.2.1	Definition of DoS and DDoS	71
3.2.2	Types of DDoS	72
3.3	Economic Background & Risk Analysis	75
3.3.1	Figures	75
3.3.2	Risk Identification	75
3.3.3	Risk Assessment	75
3.3.4	Risk management	76
3.3.5	Application on DDoS	77
3.4	SEconomy	78
3.4.1	Actors and Components	78
3.4.2	Modeling Risks, Impacts, and Prevention Measures	79
3.4.3	Modeling Costs and Attributes	79
3.4.4	Overall Economic Assessment	80
3.5	Case study	80
3.5.1	Motivation	80
3.5.2	The Mirai Botnet	81
3.5.3	Involved Actors	82
3.5.4	Overview, Components, Systems and Subsystems	83
3.5.5	Risks and Impacts	83
3.5.6	Incident Response and Prevention Measures	87
3.5.7	Costs and Overall Assessment	90
3.6	Outlook	91
3.6.1	5G	91
3.6.2	IoT	91
3.6.3	Open Source Malware	92
3.6.4	X-as-a-Service	92
3.6.5	Software Defined Networking (SDN)	92
3.7	Concluding Considerations	93

3.1 Introduction

3.1.1 Motivation

Motivated by the high number of cyber threats and especially many economic aspects involved in cybersecurity, it is imperative to understand the economics behind these malicious activities. Distributed Denial-of-Service (DDoS) attacks, leveraged by the increase in available bandwidth and the exponential growth of Internet of Things (IoT) devices, are currently a top threat to major Internet service providers. Based on the potential enabling of 5G networks, the DDoS attacking scenario is expected to gain even higher levels, offering greater connectivity to an equally larger bandwidth to a number of malicious devices. This report applies the use of Dyn DNS within the SEconomy framework, analyzing the involved actors, systems and sub-systems, their relations and impacts, as well as associated risks.

3.1.2 History of attacks

DDoS attacks have existed in some form since the beginning of the commercial web itself and the problem has gotten progressively worse. The simple yet effective nature of DDoS makes the subject more relatable to all new technologies. Especially with the growth of cloud adoption and IoT, their frequency and especially their magnitude has grown exponentially. In this section, we are mentioning some of the most recent and biggest DDoS attacks ever to be recorded.

3.1.2.1 Attack against KrebsOnSecurity

On September 2016, KrebsOnSecurity website, the blog of a renowned security journalist, Brian Krebs, with DDoS protection measures in place suffered an attack which was at the time, the largest one ever recorded and managed to keep the site offline for a few hours. Originating from a botnet, the biggest chunk of the attack came in the form of traffic designed to look like it was generic routing encapsulation (GRE) data packets, a communication protocol used to establish a direct, point-to-point connection between network nodes. GRE lets two peers share data they wouldnât be able to share over the public network itself. Akamai, that was at the time providing the website pro bono protection service, confirmed the unprecedented 620Gpbs attack rate. Akamai claimed that a sustained DDoS attack against Krebsâ site could have cost millions of dollars to mitigate. As a reference, the attack was almost twice as much traffic as Akamai had ever seen in a previous attack, and the level of protection that they offered would cost a website between \$150,000 to \$200,000 annually. It is evident, that this kind of cost is prohibitive for any small and medium business with even a simple website.

3.1.2.2 OHV Hosting Company

The attack, which happened on September 2016 just a week after the KrebsOnSecurity attack, and using the same botnet, had an attack rate of 100Tbps, the largest to have ever occurred at the time. To put a terabyte of traffic in perspective, using AT&T estimates, it is equal to streaming 170 hours of 4K video. But the significance of the attack can be better evaluated by its effect. Bringing huge clusters of websites down even for a few hours can have damaging results for many businesses at the same time, while attacking a single target. Hypothetically, if the attacks were to be timed to focus on high revenue seasons, like on Black Friday or around Christmas, the losses for the companies could be devastating.

3.1.2.5 Github and the memcached servers

One of the biggest DDoS attack ever recorded to date was in February 2018 against Github. The attackers took advantage of the caching system memcached. The servers running memcached are used to speed things up by sending any program's first attempt to retrieve data to its cache before querying the database which requires more time. This kind of attack falls under the reflection/amplification category and it works by exploiting a memcache server to spoof the IP address of an actual website, in this case Github. More specifically, as seen in Figure 1.2, the first steps of the attack involves creating a list of IP machines responding to udp on port 11211 and preloading items to memcached to be used as attack payloads. These items don't have to be big, just the server's default limit of 1Mb of data. The attack is launched by issuing commands to the attacker's list of preloaded servers, asking them for the 1Mb payload. The attacker then spoofs the source address with that of the victim, so that the data responses are sent to the victim not back to the attacker, which is the reflection component of this case. The reflection continues while the attack builds up and the memcached servers are being hit multiple times a second. This, combined with the fact that you can ask for many copies of the file at the same time, results in the amplification factor getting really big. The servers can therefore mistakenly send a flood of data to the victim website, overwhelming it with traffic and taking it offline. At its peak rate, the recorded traffic of the attack against Github reached 1.3 Tbps and lasted for about 20 minutes. Github managed to successfully defend against the it, as it has defence mechanisms in place. Traffic was directed to Akamai, a cloud computing company that provided protection from the flood, so the Scrubbing centers were able to filter the traffic. Unfortunately, there are no available data about the cost of memcached attacks, but if we revisit the case of KrebsOnSecurity, we can easily infer that the costs of protection against these attacks are definitely far from insignificant. Maybe the biggest, but less famous, DDoS attack to date, which we have even less information about, happened in May 2018. Researchers from Arbor Networks, a different DDoS mitigation service, reported a 1.7Tbps DDoS that also relied on the newly documented memcached amplification method. The attack targeted a customer of a US-based service provider that remains unnamed. Reportedly, the attack was successfully defended against.

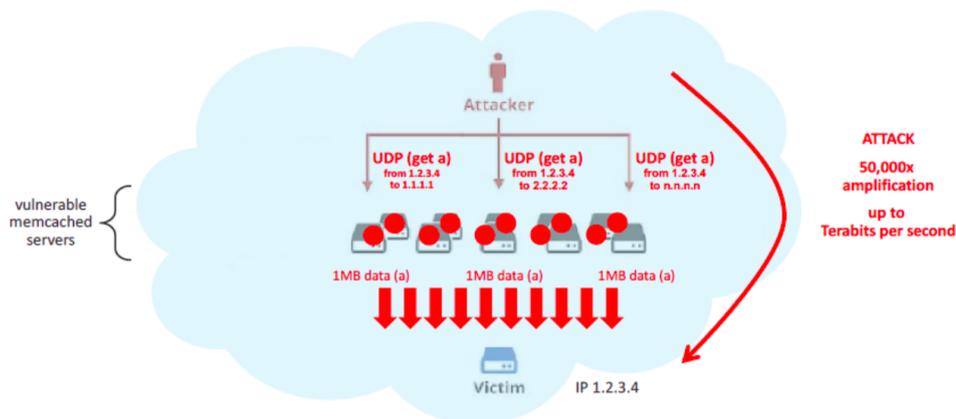


Figure 3.2: Reflection/Amplification DDoS with memcached

So far, we were measuring the attacks in terms of bandwidth, which is not the only way to measure the magnitude of an attack. Often times, it is measured in packets per second (PPS), which is the forwarding rate. Surprisingly, If we consider the second measurement, there is another contender for first place as the biggest DDoS attack. In January 2019, a client of Imperva (a cybersecurity company) faced an attack of 500 million packets

per second (Mpps). In terms of packets, it is four times larger than the ones against Github. One might argue, that for the mitigation of an attack, packets per second is more important than their size, because the defender has to filter through vastly more packets to mitigate their malicious effect. Certainly, it always comes down to the specifics and complexity of each case, but for the sake of consistency, it is worth mentioning.

3.1.2.6 Shifting the DDoS trends

Researchers estimate there are more than 90,000 memcached servers on the internet, of which more than 50,000 are currently vulnerable to reflection attacks. Figure 1.3 shows the distribution of servers still listening on TCP or UDP port 11211 as of 2018. Even though no recent numbers for the number of servers were found, the vast number of servers running memcached today, will most probably make this a lasting vulnerability that attackers will continue to target and try to exploit. This kind of attack clearly shifts the DDoS trends and one could say that it marks the beginning of a new era for DDoS. Therefore, it is crucial for companies to take measures to protect themselves. For instance, Network operators should take proactive measures to ensure they are prepared to detect, traceback, and mitigate these attacks, and at the same time ensure that any memcached installations in their networks cannot be exploited for reflection or amplification attacks. Internet Service Providers need to block spoofed packets from exiting their networks, and protocol developers need to better understand velocity checking and amplification attacks.

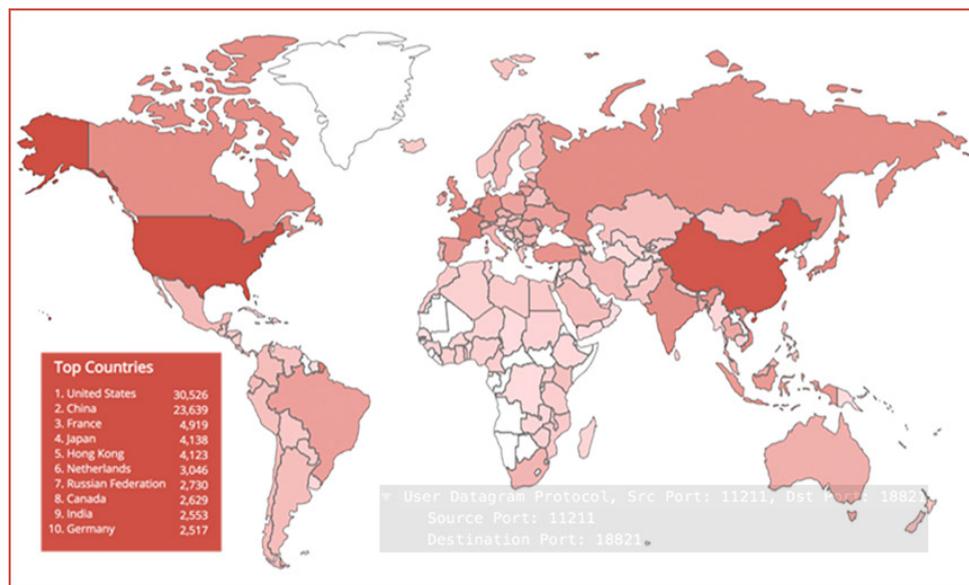


Figure 3.3: Servers listening on TCP or UDP port 11211 as of 2018

3.1.3 Economic impact

As the complexity and frequency of DDoS attacks has grown, so has the cost to businesses and even individuals. Especially for companies, the business model of which, relies on their service's continuous availability and their online presence, the stakes get higher. It is important to point out, that the lack of centralized data sources about cyberattacks makes it difficult to accurately measure their economic impact. Most statistics come from estimates based on cases and reports of surveys based on the limited data of clients of usually security companies. That said, and based on the openly available data, we can associate any given DDoS attack with both direct and indirect costs. The direct costs are generally easier to measure, while the indirect costs are often not measured (or fully understood) for months or even years after the successful execution of a DDoS attack.

The costs generalize to cyberattacks, and certainly differ depending on the incident, but they also stand in the case of DoS attacks.

Direct costs include:

- Loss of revenue and loss of traffic
- Loss of productivity, especially for the affected users
- Personnel costs (specifically within IT operations and security)
- Specialized consultants: This can include specialists that perform forensics to find details about the attack or professionals that a company may hire to help defend its systems from future attacks
- Consumer credits
- Legal and compliance fees
- Public relations

Indirect costs often include:

- Damage to the brand: usually referred to as reputational damage, is when a company faces a loss of profits as a consequence of loss of confidence in the company among its customers
- Theft of vital data
- Loss of valuable customers
- Opportunity loss.

The annual report from Verizon [2], showed that from the 41,686 incidents that they observed in 2019 almost 30% were DoS attacks and it was the most frequent incident after privilege misuse. Kaspersky Lab's research indicates that DDoS attacks were on the rise in Q1 2019: "Kaspersky DDoS Protection statistics show that all DDoS attack indicators increased last quarter. The total number of attacks climbed by 84%, and the number of sustained (over 60 minutes) DDoS sessions precisely doubled. The average duration increased by 4.21 times, while the segment of extremely long attacks posted a massive 487% growth." [3] The costs might be more prominent in cases of attacks resulting in data breach or in cases of bigger companies, but given that many malicious acts employ a combination of attacks, it still falls under the case we are examining. As an example, A 2014 study by the Ponemon Institute [5], measuring consumer sentiment based on responses to companies that were affected by data confidentiality breaches concluded that 44% of respondents would end or somehow reduce their business relationship with a company affected by a data confidentiality breach (where informed by media coverage). However, 71% of those that were actually affected by such an incident continued their business relationship with the affected company. That could be due to a lack of alternatives or the belief that most companies are, or can be, affected by such incidents. It is also notable that from the different victims in Verizon's dataset, over 99% of the times, the victims of DoS attacks are large organizations.

According to the latest report from Netscout [4], organizations reported increased per-minute and total costs associated with the outage of internet services. Almost half estimated a per-minute outage cost of between \$1,000 and \$10,000, instead of \$0 to \$1,000 as it was in 2017 (Figure 1.3). When it comes to overall attack cost, the stakes grew higher as well. In 2017, according to the same report, 55 percent estimated the average attack

cost to be less than \$10,000. In 2018, 53 percent saw the cost impact significantly higher, between \$10,000 and \$100,000.

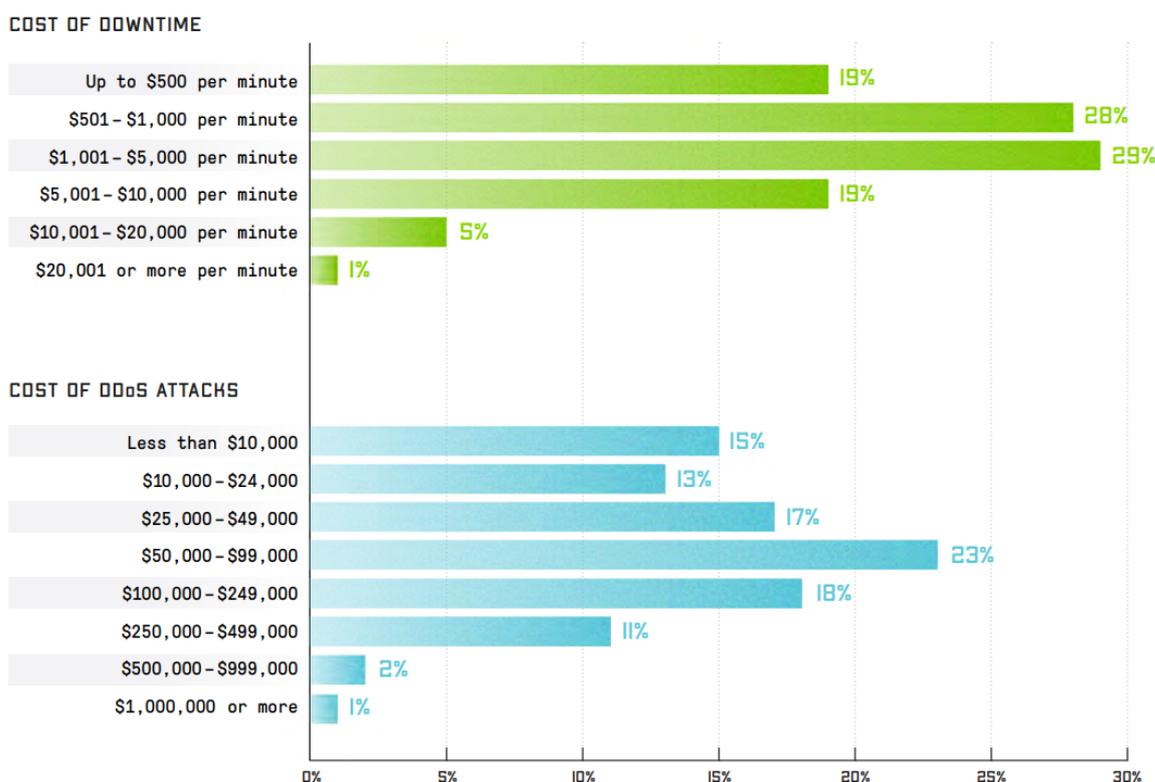


Figure 3.4: Cost-per-minute of website downtime and cost of DDoS

In the case of Lonestar, and the outage of the internet in Liberia, BBC quotes the written submissions to the court, of Babatunde Osho, Lonestar's former chief executive, who stated that the attack's consequences have been devastating. The submissions stated that "The DDOS perpetrated by Daniel Kaye seriously compromised Lonestar's ability to provide a reliable internet connection to its customers," "In turn, Mr Kaye's actions prevented Lonestar's customers from communicating with each other, obtaining access to essential services and carrying out their day-to-day business activities. In terms of financial losses he mentioned that "A substantial number of Lonestar's customers switched to competitors" and that "In the years preceding the DDOS attacks, Lonestar's annual revenue exceeded \$80m (£62.4m). Since the attacks, revenue has decreased by tens of millions and its current liabilities have increased by tens of millions.

Closely tied to the economic costs and risks of DDoS, is the motivation behind them, which lies in the most hard-to-stop cases, in economic incentives. In their paper, Zargar et al. [6], categorized the various incentives in five categories. Financial/Economic (gain motivated attacks which are usually the ones with the most experienced attackers and they are the most hard to stop), revenge, ideological belief, intellectual challenge, cyber-warfare. The 2018 report of the White House about the Cost of Malicious Cyber Activity, analysed the different motivations being either political, economical, technical, or having a military agenda and categorized the bad actors in five categories, namely corporate competitors, hacktivists, organized criminal groups, opportunists and company insiders. In the case of Liberia's Lonestar for example, the hacker was hired by Cellcom, one of the company's competitors to launch the attack. Another example comes from the case of Krebs on Security website and although the motivation and people behind the attack are not officially confirmed, Krebs had talked extensively on his website about takedown of the DDoS-for-hire service vDOS, which coincided with the arrests of two young men named in his original report as founders of the service. It has also been reported that

some of the POST request that were part of the attack included the string `freeapplej4ck`, a reference to the nickname used by one of the vDOS co-owners. So, as the two seem closely related, this attack could be categorized under criminal groups seeking revenge. Furthermore, Akamai's 2019 [7] research reveals that malicious bots, which are one of the most common threats, are constantly evolving, and the people developing them are actively looking for evasion techniques, going so far as to hire developers with unique brand- and vendor-specific expertise. This statement comes to support the words of Dyn's Chief strategy officer Kyle York, who stated that the attack contained "specific nuance to parts of our infrastructure". What is more, incidents that are exploiting the many still open to public memcached servers, are reported to have messages for demands for payment, as ransom for the attacks to cease.

Given the adverse impacts major DDoS attacks can have, it is interesting to point out that DDoS attacks are not universally punishable by law. In the US, such attacks constitute a federal crime and can be punished with fines or imprisonment. They are also illegal in the UK and are penalized with up to a ten year imprisonment. Referencing again the case of the attack against Lonestar, the attacker, who has been since then arrested and went through trial in the UK, was charged with 32 months of imprisonment. In Switzerland, DDOS attacks are criminally punishable with sanctions that similarly range from fines to imprisonment of up to three years.

3.2 DDoS Background

3.2.1 Definition of DoS and DDoS

A denial-of-service (DoS) attack is a type of cyber attack in which a perpetrator aims to cause a computer or device to be unavailable to its users by interrupting its normal functionality. The unavailability of the computer/device leads to a denial-of-service to additional users. Specifically, it is a malicious attempt to disrupt normal traffic of a targeted server, service or network by overwhelming the target with a flood of Internet traffic. A DoS attack comes from using a single computer to launch the attack. However, a distributed denial-of-service (DDoS) attack is a type of DoS attack that comes from many distributed sources. Thus, a DDoS is a large-scale DoS attack where the perpetrator uses more than one unique IP address, often from thousands of hosts infected with malware. A distributed denial of service attack typically involves more than around 3 to 5 nodes on different networks; fewer nodes may constitute a DoS attack but not a DDoS attack. Machines that are exploited can be computers and other networked resources such as IoT devices. [8]

Analytically, a DDoS attack requires an attacker to gain control of a network of online machines in order to carry out an attack. Each infected device is called a bot, each of these infected bots works together with other bots in order to create a disrupted network called a botnet. Botnets are created for several reasons, but they all have the same objective: taking web resources offline in order to deny your customer access. After the creation of botnet, the attacker can manage the machines by sending updated instructions to each bot via a method of remote control. When the botnet targets the IP address of a victim, each bot will give a response by sending requests to the target, causing an overflow to the capacity of the targeted server or network, leading to a denial-of-service to normal traffic. Since each bot is a legitimate Internet device, the separation of the attack traffic from normal traffic is quite difficult.[9]

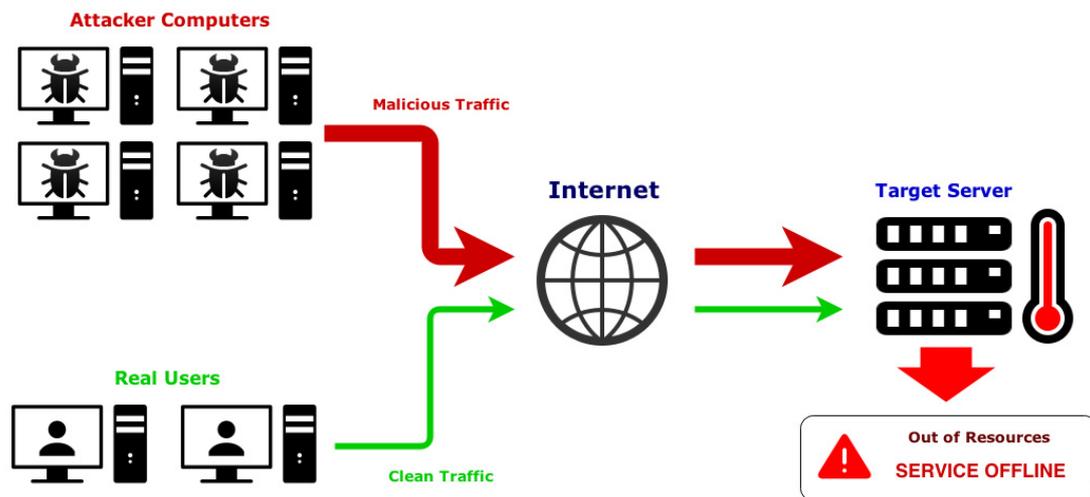


Figure 3.5: Operation of DDoS attack

Attackers of DDoS attacks often aim at sites or services stored on high-profile web servers such as banks or credit card payment gateways.

3.2.2 Types of DDoS

There are three basic categories of attacks: [10, 11, 12]

- volumetric attacks, which use high traffic to overwhelm the network bandwidth
- protocol attacks, which focus on exploiting server resources
- application attacks, which focus on web applications

Different types of attacks fall into different categories based on the traffic quantity and the vulnerabilities being targeted. It is essential to separate one type of attack from another so that we can then devise generalized mitigation strategies for each of them.

3.2.2.1 Volumetric attacks

A volumetric attack is a volume-based attack, which creates a high amount of traffic, or request packets, to a target in order to render the service unavailable. Packets are the units of data carried between networks and form a main part of the way data is communicated across the internet. The target can be any type of network components, like a web server or a flood of requests to the DNS. These attacks aim to slow or stop the network services and attempt to create congestion by consuming all available bandwidth between the target and the larger Internet. Numerous data are sent to a target by using a form of amplification or another means of creating massive traffic, such as requests from a botnet. It is important for the DNS and webserver to be public in order people request service from them, and they can be a direct target for the attacker. It is worth mentioning that in case of flooding, there is no need for the request to be properly formatted. That means that as long as the request packet arrives at the target the attack can potentially succeed. Examples of common volumetric attacks are UDP floods as well as ICMP floods. A UDP flood is a DDoS attack that floods a target with User Datagram Protocol (UDP) packets. The goal of the attack is to flood random ports on the receiving host. This causes the host to repeatedly check for applications with these datagrams and when no application is found, they reply with an ICMP âDestination Unreachableâ packet. As more and more UDP packets are received and answered, the system becomes overwhelmed and unresponsive to other clients. Thus it can completely lead to inaccessibility.

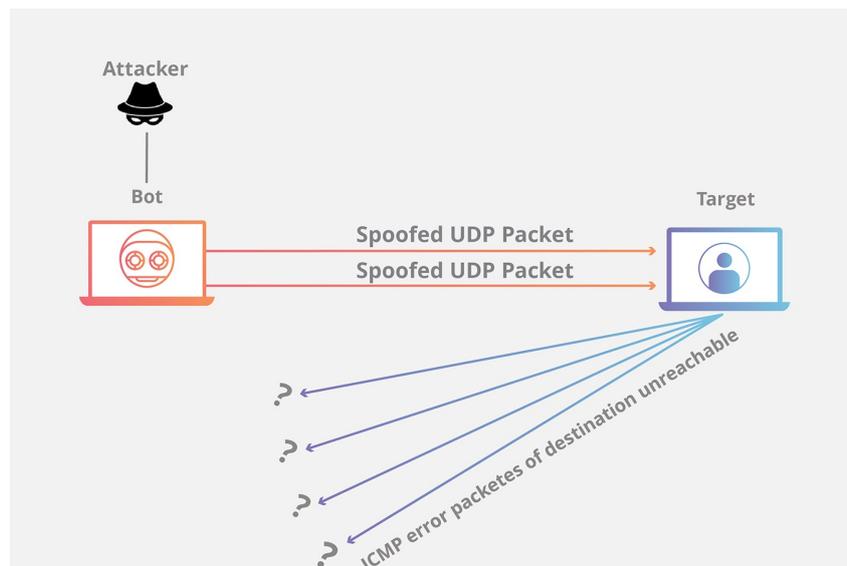


Figure 3.6: UDP flood

The ICMP (Internet Control Message Protocol) Flood, also known as Ping, is quite similar to the UDP flood attack. An ICMP flood overwhelms the target resource with ICMP Echo Requests, called pings, sending packets as fast as possible without waiting for replies. ICMP flood can consume both incoming and outgoing bandwidth, since the targeted servers will often attempt to respond with ICMP Echo Reply packets, resulting in a significant slowdown to the overall system.

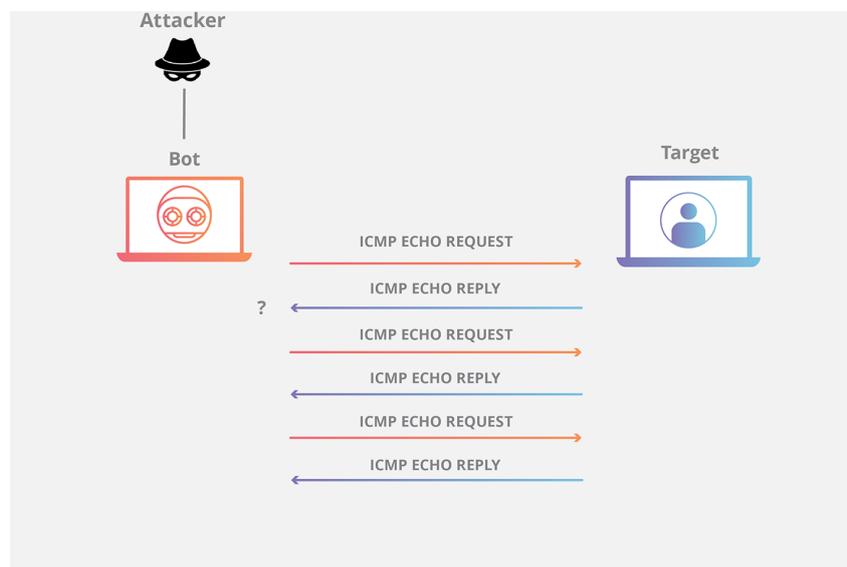


Figure 3.7: ICMP (Ping) Flood

3.2.2.2 Protocol attacks

A protocol attack takes advantage of the ways operating systems implement protocols. This type of attack uses actual server resources, or those of in-between communication equipment, for instance, firewalls and load balancers, and is measured in packets per second (Pps). It includes SYN floods, Ping of Death, Smurf DDoS and more.

In a SYN flood, the attacker sends a sequence of SYN requests to a targeted machine intending to consume enough server resources to make the system unavailable to legitimate traffic. The attacker send request packets to every port on the targeted server, often using a fake IP address, repeatedly and faster than the targeted machine can process. Thus the attacker can flood all available ports on the targeted server, causing the targeted device to respond to legitimate traffic slowly or not at all, creating a network saturation.

In the Ping of Death attack, ICMP (Internet Control Message Protocol) echo requests are sent to the targeted victim. The total data sizes of the requests are greater than the maximum IP standard size. As a result, the attack crashes or freezes the victim's machine and overflows the memory buffers allocated for the packet, causing a denial of service for legitimate packets.

A Smurf attack is a distributed denial-of-service (DDoS) attack in which an attacker attempts to overwhelm a targeted server with Internet Control Message Protocol (ICMP) packets. By making requests with the spoofed IP address of the targeted device to one or more computer networks, the computer networks then respond to the targeted server, reinforcing the initial attack traffic and potentially flooding the target, causing its inaccessibility.

3.2.2.3 Application attacks

Application attack is comprised of seemingly legitimate and innocent requests, exploiting vulnerabilities in applications. The goal of this type of attack is to not attack the entire server, but applications with known weaknesses to crash the webserver. Instead of attempting to overwhelm the entire server, an attacker will focus its attack on one or a few applications. Representative examples of application-specific targets are web-based email apps, WordPress, and forum software. They also include GET/POST floods, low-and-slow attacks, Windows vulnerabilities, attacks that target Apache, etc. This kind of attack may require more specific knowledge but not necessarily in-depth knowledge. For instance, understanding the basics of the HTTP protocol POST, you can easily launch a low-and-slow POST operation by posting one out of thousands of characters at a time to an HTTP server before the session times out.

3.2.2.4 Reflection Attacks

Reflection attacks are happening when the attack uses one or more intermediate hosts to generate DDoS attacks. Most of the time there are DDoS bot malware programs waiting for commands to be instructed to attack a particular host. In a typical case, numerous hosts are used against the intended target. The originating command and control (C&C) server sends the instructions for the bots to follow.

3.2.2.5 Amplified Attacks

Amplified DDoS attacks use noisy protocols, which respond with more than one packet when receiving a single packet (thus the amplification), against the intended targets. For instance, the DDoS attacker may send a single malformed request to a web server with the origination IP address being falsified as belonging to the victim. The intermediate webserver gets the malformed request and sends it back to the originating IP address (the target victim) with multiple responses or attempts at error correction. Another popular DDoS amplification attack abuses DNS servers by requesting larger amounts of legitimate DNS information to which the DNS server sends back multiple, if not dozens of, packets to the intended victim. The bigger the amplification, the happier the DDoS attacker. When amplification is coordinated with tens to hundreds of thousands of bots, huge DDoS attacks can be realized.

3.2.2.6 Multivector Attacks

Since the goal of the attacker is to make the service unavailable to other users, the attack can be a combination of the different types for a multivector attack. In several instances, we have seen the attack incident starts as a flood of traffic toward the network consisting

of classic floods, then changing into various other forms of attacks such as protocol-level attacks.

3.3 Economic Background & Risk Analysis

3.3.1 Figures

According to Accenture [20], the average cost of cyber crime has increased rapidly over the last years. From 2018 to 2019, large scale (5000+ employees) enterprises' average cost of cybercrime increased from 11.7 to 13 million US-\$, which is a plus of twelve percent in a year â which is as relevant as a growth of 72% over five years. The cost of Denial of Service attacks has nearly raised the same, by 10 percent to 1.7 Mio. US-\$ average per company. On the other hand, companies investments into cyber security are also rising. In the United States alone, 15 billion US-\$ were invested into cyber security defence according to US market research (1). The annual growth (from 2012-13 according to Market Research) was also significant with a value of six percent. Such immense sums demand for models to identify, assess and manage risks concerning cyber security. However, they are harder to develop than classic economic models, because the industry is rapidly evolving with new dangers popping up without much information neither data associated, managers not knowing all technical details about their systems and multiple layers correlating in potential damage.

3.3.2 Risk Identification

On the risk identification side, Microsoft developed a well-known risk identification framework called âSTRIDEâ, which is an acronym for the six threat categories Spoofing identity, Tampering with data, Repudiation, Information disclosure, Denial of service and Elevation of privilege. A Denial of Service attack is described as follows: âDenial of service (DoS) attacks deny service to valid usersâfor example, by making a Web server temporarily unavailable or unusable. You must protect against certain types of DoS threats simply to improve system availability and reliability.â [22]

3.3.3 Risk Assessment

3.3.3.1 Technical example: DREAD

There also exist several risk assessment frameworks that are needed to quantify risks. A technical framework to assess risks and to rate them is called DREAD. It considers the following five dimensions of risks according to OpenStack [19]

- Damage: If the vulnerability is exploited, how much damage will be caused?
- Reproducibility: How reliably can the vulnerability be exploited?
- Exploitability: How difficult is the vulnerability to exploit?
- Affected Users: How many users will be affected?
- Discoverability: How easy is it to discover the threat and to learn of the vulnerability

For every dimension, a score between 0 and 10 is given (0 meaning no impact, 10 meaning worst possible impact). The sum of these scores is then divided by five such that an overall risk score between 0 and 10 is again supplied based on which decisions may be taken. There exist conventions that say discoverability shall always be rated a ten.

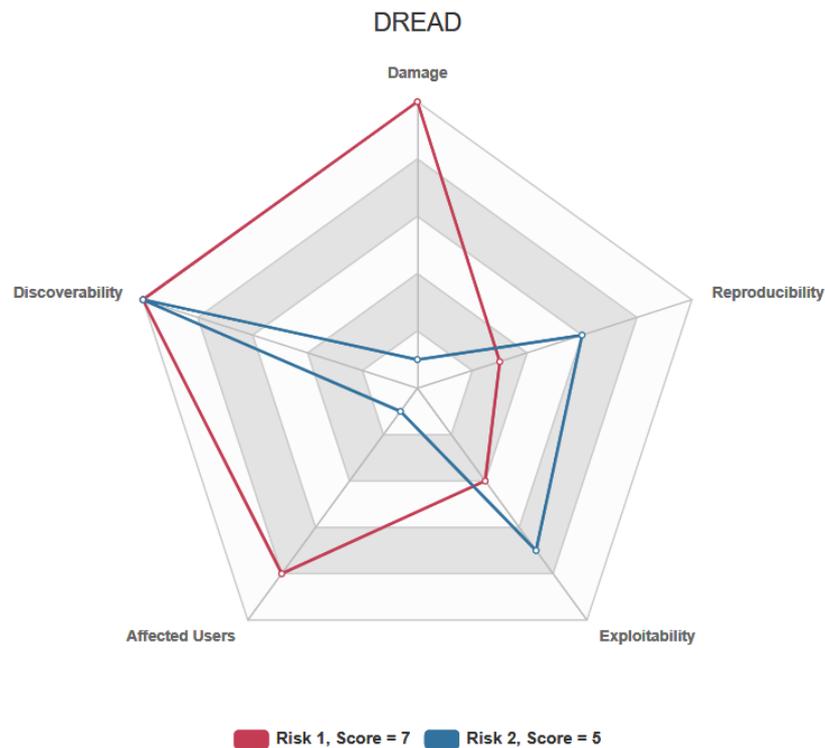


Figure 3.8: Visualization of applying DREAD on two risks making it easy to compare different risk dimensions of different risks

3.3.3.2 Economical example: ROSI

From a managerial point of view however it is not only important to quantify the risk itself, but also the return of an action that might be taken such that managerial actions can be taken in risk management. One framework that answers this question is called “Return On Security Investment (ROSI)”. It tries to give a figure that makes it easy to decide whether an action concerning security shall be taken, similar to the Return On Investment (ROI) often used in economics. In this model, Sonnenreich [16] takes the formula known for the ROI and redefines the Expected Returns as the Risk Exposure multiplied by the Risk Mitigated.

$$ROSI = \frac{(Risk\ Exposure \times Risk\ Mitigated) - Solution\ Cost}{Solution\ Cost}$$

Sonnenreich’s model also further defines risk exposure as the annual loss exposure, which is the product of the estimated annual rate of occurrence (ARO) multiplied with the single loss exposure (SRE). From a non-technical point of view the Expected Returns can be seen as “the probability of a security issue multiplied with the loss occurring from the security issue”. As the ARO and SRE may be hard to quantify, actuarial tables from insurance companies are often considered to have a rough estimation for these figures. However specific risks should also be taken into consideration, for example the sacking of a chief security officer may lead to a danger of him using insider information to launch a specific attack.

3.3.4 Risk management

Another source of well-defined risk assessment and even management frameworks is the national institute of standards and technology of the US. It provides several guidelines for risk assessment and even a full risk management framework. There also exist fully

fledged risk management frameworks and guides how to apply them.[18] One of them is NIST 800-37 released by the national institute of standards and technology (NIST) of the US. It is a guideline how to apply a risk management framework in practice. For the risk assessment part, it redirects to another standard (NIST 800-30) which is an extensive guideline on risk assessment for cyber security of US governmental organisations. The risk management framework consists of six steps and encourages to integrate these steps into system development.



Figure 3.9: Risk Management Framework NIST 800-300

In the categorization step (usually the first step to be taken), the goal is to identify the importance of the system, therefore the potential impact of non-functionality or leaks onto business and other systems. In a second step, appropriate control shall be selected. This selection is specified within another guideline and provides three different levels of security baselines as well as a catalog of security controls. In the next step those guidelines shall be applied or other selected controls implemented. NIST highly encourages to consider this phase during the system development process. The next step is the assessment, where the implemented controls are examined and tested, and a report is provided. In the authorization step afterwards, it is determined whether the resulting risk is acceptable or not. If yes, a continuous risk monitoring process is started, in which the system and its implemented controls are continuously checked for signs of attack and the effectiveness of the implemented controls is also continuously reassessed.

3.3.5 Application on DDoS

3.3.5.1 Identification

Applying the before mentioned frameworks to the topic of DDoS attacks is very easy for the risk identification part, as e. g. in STRIDE the letter D is for (D)DoS attacks. However, it gets way harder when trying to assess this risk or even managing it. Applying the DREAD formula, it is very challenging to estimate the reproducibility and exploitability because those do heavily depend on the given system and existing protection mechanisms that may already be implemented. The most interesting part, however is quantifying the revenues on an investment in DDoS mitigation, as this has become business for enterprises like Cloudflare offering DDoS protection as a Service at quite low rates. [21]

3.3.5.2 ROSI

To apply the ROSI formula on such an investment to decide whether it is worth, one would still have to quantify the risk mitigated by the solution as well as the risk exposure, which are heavily dependent on the business case. As DDoS attacks can happen on different layers, the risk exposure and risk mitigation of these solutions also need to be assessed on all layers. For the transport layer, however, Cloudflare advertises to have a bandwidth of 15 times the largest DDoS attack which would indicate a risk mitigation factor of 1.0 (equivalent to 100%).

3.3.5.3 Risk exposure

Calculating the risk exposure of DDoS however is not that straight forward. Depending on the business case, the exposure needs to be calculated in a very different fashion for example a sales company would have to take into consideration sale losses from a DDoS attack (based on at least the downtime multiplied by their yearly revenues), whereas public institutions like the university of Zurich would have to take into consideration the loss of bureaucratic efficiency, leading to unnecessary expenses for staff that can't do their work and then maybe leading to socioecological following costs like a loss of reputation. However one can say that with a growing institution size and therefore growing size of revenues (for businesses) or expenses (for public institutions or governments) the risk exposure grows too, leveraging the ROSI on (D)DoS mitigation.

3.4 SEconomy

In this section the SEconomy framework is introduced. This framework will be used in section 3.5 to analyze the impacts of the DDoS attack on Dyn in 2016. This introduction fully relies on the work of Rodrigues, Franco, Parangi and Stiller[17]. SEconomy is a strictly step-based framework to measure economic impact of cybersecurity activities in a distributed ecosystem with several actors. With the help of the SEconomy framework an organization should be able to decide which security investments should be applied, considering the ROSI, which was explained in Section 3.3.3.2. The framework is divided in 5 steps which are shown in Figure ???. Each step will be explained shortly in the following.

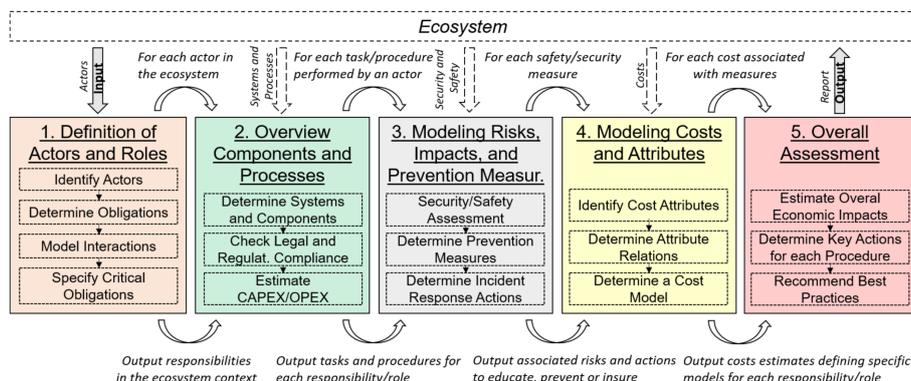


Figure 3.10: SEconomy Framework

3.4.1 Actors and Components

In the first step the actors involved in an ecosystem and their roles/obligations are identified. Critical obligations of the actors should be specified. Additionally, interactions

between actors in the ecosystem are modeled. Every ecosystem that requires an assurance of security and safety levels can be given as input here, for example the production chain of an aircraft system, which has some critical components considering security and safety. In the second step the systems/components and processes, which are performed by the actors identified in step 1, are determined. Safety and security regulations must be taken into account for the implementation of these systems. Obviously this further implies consequences for the capital and operational expenditures (CAPEX/OPEX).

3.4.2 Modeling Risks, Impacts, and Prevention Measures

For each system determined in step 2 its security and safety threats are analyzed in step 3. Risk models are generated and the possible impacts of the risks are estimated. To do so existing Cybersecurity Risk Management Frameworks(RMF) can support this step. For example the NIST framework, which was described in Section 3.3.4, can help in the risk analysis part. Other supporting RMF examples like STRIDE and DREAD can support in risk identification respectively risk assessments as stated in Section 3.3. In a rational approach to analyze the risks, three questions are examined.

- What are the potential vulnerabilities of the system and how high is the probability that they are successfully exploited?
- What does it cost if such a vulnerability is exploited?
- What does it cost to mitigate vulnerabilities?

Considering the probability of failure it is also important to analyze the correlation between different systems and risks. So there could be added two more questions.

- If system A fails, does the probability increase that also system B fails?
- If risk C occurs, does the probability increase that also risk D occurs?

Looking at these additional questions it is important to map dependencies between systems and between risks in this step. Finally, for each analyzed risk counter-measures to respond if it occurs are investigated. According to Rodrigues et al. the third step can be described as follows: "The risk analysis is the fundamental stage toward mapping costs associated with cybersecurity. It is responsible for determining, proactively or reactively, possible vulnerabilities/threats (i.e., probabilities) that may occur as a function of time as well as their associated counter-measures."[17]

3.4.3 Modeling Costs and Attributes

The risk analysis from the third step is considered in the fourth step to map costs in a fine-grained manner. In this step costs are associated to each threat defined in step 3 to estimate the economic impact of a threat. To calculate the associated cost of a threat also the dependent systems need to be considered, since it is possible that a failure in system A leads to a failure in system B as described in Section 3.4.2. With the help of this threat/-cost mapping an organization can better decide what measures it wants to apply to handle a specific threat. At this stage it is important to differ between proactive and reactive approaches in counter-measures. A proactive approach is a prevention measure, which tries to protect the system from a specific threat for example through education/training of personal or through redundancy of critical systems. Hence by applying a proactive approach the organizations aims that the threat never occurs. All the money spent for these activities is summarized as the Proactive Mitigation cost(PMC). The reactive approach implies measures that are applied when a threat already occurred. It comprises

monitoring of the threat that a possible failure is recognized at an early stage, as well as recovery measures to mitigate the impact of the occurred threat. All the money spent for these activities is summarized as the Reactive Mitigation cost(RMC). Through different key values like the ROSI, which was described in Section 3.3.3.2, it can be decided to which extent it is worth to invest in prevention measures. Also it is possible that an organization prefers to not invest in the prevention against a specific threat, because they decide it is not worth. Normally both proactive and reactive approaches are implemented in an organization, but the intensity can differ. Mainly proactive approaches are favored, since it is mostly cheaper to prevent a threat than handle its negative impact when it occurred. In such a case the return on the security investment is bigger than 1. Proactive approaches also increases customer trust.

3.4.4 Overall Economic Assessment

At the end of the framework step 5 concludes the analysis by estimating the overall economic impact of the cybersecurity measures defined in the previous steps, determining improvement actions and recommending best practices. The algorithm shown in Figure ?? summarizes the economic impact of possible security investment in any system, required by any role of any actor of the ecosystem. It is applied to estimate the cost for the entire ecosystem. Line 5 takes the dependency between the systems into consideration, which was explained in step 3 of the framework in Section 3.4.2. Afterwards with the result of line 5 the exposure cost, the Proactive Mitigation cost and the Reactive Mitigation cost are calculated. Finally, with applying the ROSI on the these newly calculated cost and the initial security cost an overall economic assessment can be concluded.

```

1 begin
2   for each Actor ∈ Ecosystem:
3     for each Role ∈ Actor:
4       for each System ∈ Role:
5         /* Correlation between linked systems in Equation 1 */
6          $p(x) \leftarrow dependence(System, \forall linkedSystems)$ 
7         /* Estimate exposure costs in Equation 2 */
8          $threat_{costs} \leftarrow T_{costs}(A, p(x))$ 
9         /* Estimate mitigation (Proactive and Reactive) costs
10          in Equation 3 */
11          $mitigation_{costs} \leftarrow PMC_{costs}(A)$ 
12          $mitigation_{costs} \leftarrow RMC_{costs}(A, p(x))$ 
13         /* Get Overall Economic Assessment (OEA) in Equation 4
14          */
15          $OEA \leftarrow ROSI(threat_{costs}, mitigation_{costs}, InitSecCost)$ 

```

Figure 3.11: Algorithm for Overall Economic Assessment (OEA)

At the end all the detected results should be written down.

3.5 Case study

3.5.1 Motivation

In October 2016 an enormous DDoS was launched on the DNS provider Dyn, Inc. Measuring a storm of 1.2Tbps of incoming traffic the attack was considered to be the largest attack at that time. 40 to 50 times of the normal traffic volume hit Dyn's servers[14]. Large parts of the domain name system infrastructure of the internet were controlled by Dyn and thus the impact spread to many industries[15]. It was widely covered by the media and it affected internet services across the globe. The company's own website, Amazon, Twitter, Paypal and other big players felt connectivity issues. Also, because

the attack was using the Mirai botnet, the incident added to further discussion about the future of internet security and the role of the Internet Of Things.

In order to be able to conduct an in depth study, detailed knowledge about Dyn's systems, structures and organization is required, knowledge which is not publicly available and kept secret. Therefore, due to an incomplete base of information regarding security investments and concrete prevention measures implemented by the concerned companies only a modified version of the SEconomy framework can be applied. In addition to that, as a comprehensive analysis would certainly exceed the scope of this paper, this case study focuses on the main aspects of the attack and tries to align them with the step-based approach of SEconomy. Moreover since vulnerabilities have already been exploited successfully, especially the third step of the framework must be treated differently in an analysis in comparison to an ex ante point of view. Risk can only be assessed regarding the future. Thus, as impacts are already known, it makes most sense to take them for an economic consideration. Furthermore incident response as well as general mitigation tactics applied shall be evaluated and prevention measures implemented after the attack shall be discussed. The last two steps of SEconomy are very important to apply the framework in practice and to concretely infer a guideline for security investments. However, since there are no detailed numbers available, step 4 on Modeling costs and attributes and step 5 with the algorithm for calculating overall costs will be handled much shorter. Instead a summary of this analysis will be provided, excluding a concrete numerical assessment of cost.

3.5.2 The Mirai Botnet

A significant amount of malicious traffic was analyzed to originate from devices that were infected by the Mirai malware. Thus a short overview about the key characteristics about the Mirai botnet shall be given.

Mirai is a software targeting devices that run on ARC-processors. Those processors run a stripped-down version of the Linux Operating System and are mainly used in smart home, mobile and Internet Of Things devices. The Malware scans the internet for vulnerable devices and after infecting them they are turned into a network of remotely controllable bots. A bot may be cleaned from the malware with a reboot, but since Mirai is constantly scanning for infectable devices, there is the chance that a device gets reinfected shortly after. The credentials for access to those devices must be changed before they get in touch with the malware again. Otherwise they immediately become bots again. As owners of such devices often do not have the required know-how to block an infection, a mirai-based botnet has the potential to become huge [?]. In Figure ??[24] the basic operation of a Mirai-Botnet is described involving a command and control server.

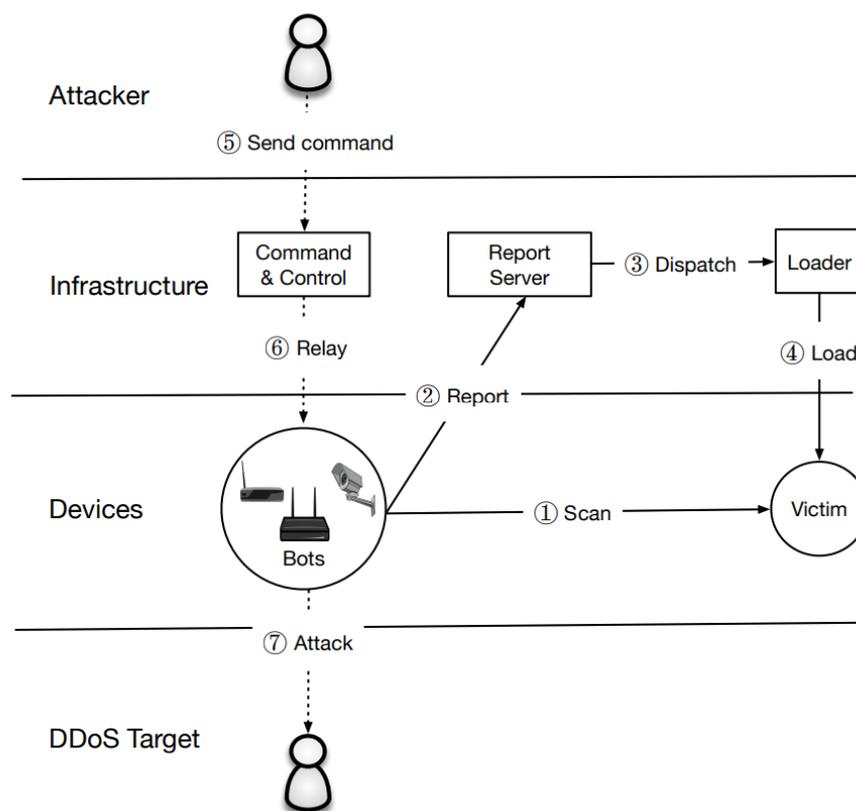


Figure 3.12: Operation of the Mirai Botnet

In a mirai botnet the bots scan the IPv4 address space for devices running SSH or telnet. They use a hardcoded dictionary of credentials and, in case of success the bot sends the IP address of the victim and the corresponding credentials to a report server. The report server asynchronously triggers a loader which in turn infects the device. After infection the bots may be steered by the command and control server[24].

3.5.3 Involved Actors

Back to the SEconomy framework, we will now try to apply the framework on this DDoS attack on Dyn. In the first step of the SEconomy framework the actors and their roles in the system need to be identified. Since there is no physical good created in the DynDNS service and no obvious production steps are present, there are different actors involved than in an organization, which produces a good. There are just a few actor involved in the DynDNS service. The main actor is the service itself, which has the critical obligation to resolve website names in IP addresses, which then would provide access to the website of companies relying on the "name solving". Beside the service a really important actor in the case of DDoS attacks in Dyn is the Network Operations Center(NOC) team[40]. This team is responsible for the monitoring and mitigation of DDoS attacks. Small DDoS attack can appear many times per day and most of them can be handled by the team. Summarized the service is the core actor, which delivers the value of the organization while the NOC team is a supporting actor, which defends the service against cyber attacks that it can work properly. However, at the 21. October the attack was so huge that the NOC team could not react in time and the service was not available for some customers for some time.

3.5.4 Overview, Components, Systems and Subsystems

For the second step of the framework it should be made clear that there are not a lot of components involved in the DynDNS service. However, some interesting concepts, which Dyn applies to secure that their service is running seamless can be mentioned at this part. It is not a big surprise that the DynDNS service is running on multiple servers, since it is a service with a huge amount of customers. With a proper load balancing, they secure that the traffic is distributed more or less equally on the servers and that one failing server is not fatal and does not impact any customers. Dyn applies a geographic based load balancing, which distributes traffic across data centers in different locations to increase the performance by shortening the distance between your customers and your data centers[41]. Because of this type of load balancing in the first DDoS attack on the DynDNS service only parts of the United States were suffering connection issues on different websites. This is because only the servers at this location were overloaded, while during the second attack all servers distributed over the world were affected. Hence these servers are a critical component of the DynDNS service, since the service does not work anymore when they are overloaded. They should be monitored and secured all the time. Obviously the effort performed by Dyn to secure them from attacks were not enough at this time and even though the problems were detected in a quite early stage, they could not be handled immediately[40].

3.5.5 Risks and Impacts

In the third step of the framework risks, their impacts and prevention measures for the organisation are investigated. Since we do research on a case, where the risk actually already happened we split this part in an investigation of the risk and its possible impacts and the countermeasures, which were applied by Dyn to avoid that this happens again. The countermeasures will be covered in the next section.

3.5.5.1 Overall Risk Assessment

Since the DynDNS service is a Web service to its customers, which resolves website names to IP addresses, the main risk of the organization is that this service does not work properly or does not work at all. The main threat that this could happen are cyberattacks like DDoS. The customers rely on the service, that it is working all the time and they pay a yearly fee that they can use it. Regarding this, if the service stops working at a specific time the provider of the service, in this case Dyn, does not loose money directly, since the fee for the service is already paid and they do not get more money for every minute the service is running properly. Hence for the provider organization itself, service downtime does not have a direct economic impact in the case of Dyn. However, loss of customer trust through service downtime tends to have a huge indirect economic impact. According to a DDoS impact survey conducted by the corero organization, which provides DDoS attack protection and mitigation services[38], 45% of the IT security people said that loss of customer trust and confidence were the most damaging consequences of DDoS attacks for their businesses, while 34 % said lost revenues were the worst effect[39]. As previously explained there is no direct lost revenue for Dyn. Hence it is important to consider the loss of trust aspect and therefore Dyn's customers. As already mentioned a few times, the DynDNS service resolves domain names like netflix.com to IP addresses that a customer of for example Netflix does not need to remember the random IP address and can just enter the domain name. Considering this, if the DNS service is not available anymore, the domain names of the websites, which are entered by internet users, can not be resolved anymore. Since a normal user does not know the real IP address, the websites are not accessible anymore. This can lead to direct economic impact or customer dissatisfaction

for Dyn's customers. Consequently this leads to loss of trust in the DynDNS service and some of the customers may change the service provider, which is an indirect economic impact. Considering the attack on the 21 of October 2016 roughly 8% of Dyn's customer base stopped using their services according to security ratings provider, BitSight. It is not clear whether the migrated customers left Dyn permanently or some of them used Dyn's service in a later stage again. Most of the customers, who stopped using Dyn's service after the attacks relied on a variety of different DNS providers already before the attack, and could easily re-route their traffic by using another DNS provider[40]. However, it seems obvious that Dyn experienced a loss of trust from their customers.

3.5.5.2 Impact to Different Customers

By looking at concrete examples of Dyn's customers we can try to estimate the economic impact on these. Some of the most famous services affected by the attack were[42]:

- Amazon
- PayPal
- Netflix
- Twitter
- Visa
- Electronic Arts
- Spotify

Regarding the missing information what percentage of customers and for how long they could not access different websites we make some limitations and assumptions here. We will focus on the second attack, because it was more globally distributed and hence for simplification we assume that no one could connect to the affected services, which is of course exaggerated, but it makes a calculation more clear. The second attack occurred at 15:50 UTC and was solved around 17:00 UTC. Hence its duration was around one hour[27]. Assuming that the use time of the services is equally distributed over the time we can estimate the impact of one hour downtime of a respective service. In the following some of the affected firms are investigated and the impact of one hour downtime on the sales is calculated. The following numbers regarding the sales of a company are taken all from the year 2018, so that the calculated numbers are most accurate to present the impact of a downtime of the corresponding service today. The numbers in the year 2016 were a bit smaller but it is still a good approximation to calculate the impact at this day. The amount of sales per hour is calculated with equation 3.1.

$$SalesPerHour = \frac{YearlySales}{365 \times 24} \quad (3.1)$$

For Amazon one hour service downtime means that nothing can be sold anymore during this time, which means that no money can be earned. With a yearly sales of 232,89 billion dollars this make 26,59 millions sales per hour [43]. It can be assumed that part of this loose can be written off, because a customer may decide to buy the wanted product on an other substitutional website or in a physical store, which is available for sure, while Amazon is not available. Hence the DDoS attack has a direct economic impact for Amazon. Similar to Amazon also PayPal is dependent on a running service that sales can be generated. With a yearly sales of 15.45 billion dollars this make 1.76 millions

sales per hour for PayPal [?]. In this case probably even a larger part of this amount can be written off, because a customer probably pay with credit card instead of waiting until PayPal is working again. Beside the direct monetary impact, both companies also have dissatisfied customers, which is possibly even worse. As a third organization Netflix is considered. Netflix is equivalent to Dyn an organisation, which provides its service against a fee, which is paid in advance. Hence there is no direct loose in sales. However, by looking at the number of hours, which could not be watched by its customer it seems obvious that there is a loose of customer trust. All customers together watch Netflix in average 164.8 million hours per day, which makes 6.87 million hours per hour [45]. This means that per hour Netflix is down 784.25 years of television can not be watched, which is a huge number. Therefore through the dissatisfaction of the customers, there is an indirect economic impact.

3.5.5.3 Impact to Financial Services Industry

At this point a very interesting point to look at is the impact to the financial service industry. U.S. banks are generally well prepared and defended against DDoS attacks. They invested heavily in DDoS protection, after they were the target of DDoS attacks in 2012 by an operation called Ababil run by the hacker group "Izz ad-Din al-Qassam Cyber Fighters"[34]. However, instead of attacking a bank directly, the attack in 2016 was directed to Dyn a service provider that was not a typical target for DDoS. Hence the attack resulted in major impact across the financial services industry. This impact in the bank sector was well monitored by Dynatrace, which is an organisation that measures performance and reliability of thousands of companies through monitoring. For the monitoring of U.S. banks, Dynatrace ran a test once per hour from twelve backbone agents across the United States of America against the home pages of dozens of banks. Figure ?? shows the results of these tests for 42 U.S. banks at the 21. October 2016, where each dot in the graph represents the home page test against one of these banks. With 12 tests per bank per hour this makes a sum of 504 samples per hour, which is more than 8 samples per minute, hence this figure gives us a nice view of the overall health of online banking in the U.S at this day[36].

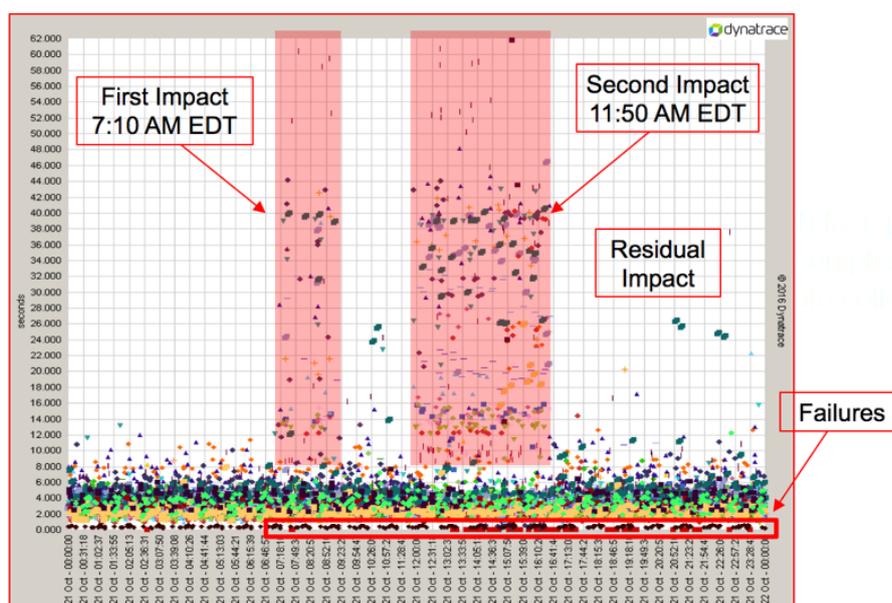


Figure 3.13: Impact on U.S. Banks

Regarding Bolstridge[36] five main observations can be made in the figure.

- At 7:10 EDT (11:10 UTC) the response time for some banking home pages increased drastically. It recovered approximately 2 hours later. This is the impact of the first DDoS attack.
- At 11:50 EDT (14:50 UTC) the response time for some banking home pages increased for a second time. This time the response time stayed at a high level until around 16:30 EDT (20:30 UTC). This is the impact of the second DDoS attack.
- Some outliers in response time stayed until late into the evening. This is because of the residual impact of the second attack.
- Some banks had complete failures of their home pages during the mentioned periods.
- Close examination shows that 28 of the 42 monitored banks had increased response time for their home pages due to this event.

Similar to the bank industry Dynatrace monitored the insurance industry with the same concept. Figure ?? shows the results of these tests for 25 U.S. insurance companies at the 21. October 2016. The observations, which can be made, are very similar to these in the banking sector. Hence also the insurance industry was impacted heavily.

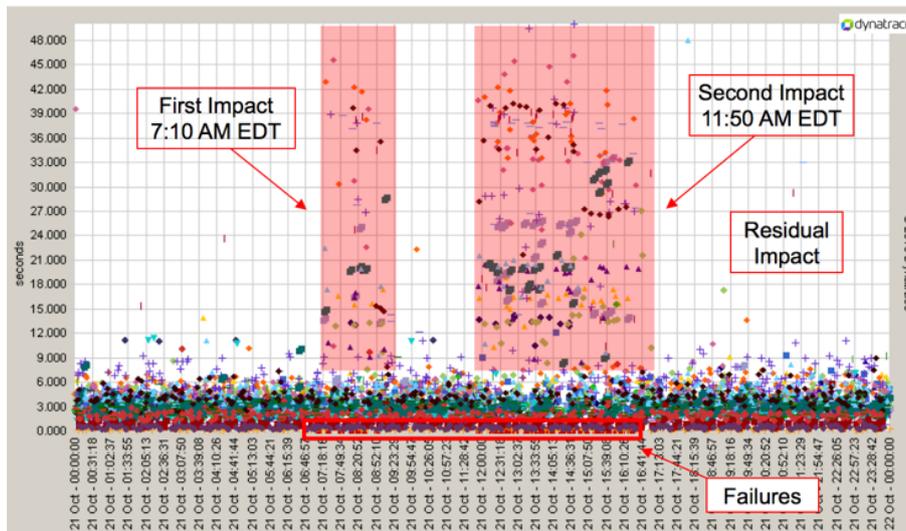


Figure 3.14: Impact on U.S. Insurance Companies

Additionally, similar observations could be made for online broker portal and credit card homepages in the United States [36]. Globally seen, for example in the United Kingdom, Dynatrace could not detect deviation in the response time during the first attack, since its focus was on the United States. During the second attack, which was more global, there were apparent deviation in the UK as well[37]. All the observations made in the financial industry covered in this section matches the time analysis of the attack from Dyn[26], which is described later in Section 3.5.6.1. However, there are an additional interesting point observed in the financial industry thanks to the Dynatrace monitoring. Even banks, which did not use DynDNS as service provider for resolving their domain name, had problems regarding their response time at the day of the attack[36]. This is because the bank websites (other websites as well) are loaded with tags, analytics, and content of many types from third parties. Each of these serve a purpose for the financial institution, but each also introduces risk[37] because of building a dependency. Because of that also non customers of Dyn had performance issues while loading their website.

3.5.5.4 General Impact

As discussed in this section, the DDoS attack on the DynDNS service had an impact on a huge amount of services. Even organizations, which were not customers of Dyn, suffered under the DDoS attack through connection problems. In some examples like Amazon and PayPal we can estimate a concrete number of economic impact of the DynDNS Server downtime. However, the most critical part is the indirect economic impact through loss of trust of the customers, which afterwards may change the service provider. Dyn suffered under this loss of customer trust and lost roughly 8% of their customers after the attack. The total revenue, which has been lost by all affected organizations as a result of the DDoS attack can only be speculated[40]. John van Sichen, the CEO of Dynatrace, estimates that the service disruption may have lost companies up to 110 million US-Dollars in revenue and sales[35].

3.5.6 Incident Response and Prevention Measures

As mentioned in the previous section, the main risk of the DynDNS service is a failure of resolving website names to IP addresses as desired by the clients. On October 21. 2016 exactly this scenario happened and the service was heavily overloaded due to a severe DDoS. The service was slowed down to such an extent that users, who wanted to access a website resolved by Dyn had to wait extremely long or could not even visit the page at all. What happened concretely directly during the attack will be discussed in the next chapter.

3.5.6.1 Direct Incident Response

The attack happened in multiple waves, as already mentioned in the previous chapter. The first one started at 11:10 UTC. At that time Dyn started to record elevating bandwidth against their Managed DNS platform in US-West regions, South America, the Asia Pacific, and Eastern Europe. Dyn launched their incident response protocols and shortly after, the attack suddenly shifted to their systems in US-East region. Huge floods of TCP and UDP traffic over port 53 and a large amount of different source IP addresses were registered. Dyn's Network Operations teams applied traffic-shaping tactics and internal filtering of packets. They manipulated anycast policies to rebalance the traffic and deployed scrubbing services [26]. It took Dyn approximately 2.5 hours to fully apply their mitigation tactics, and after a while the attack diminished. It was not only due to Dyn's own efforts, but also due to the help and mitigation efforts of upstream providers. A second wave of traffic flood occurred at 15:50 UTC and was more globally distributed, as can be seen in the figure below ??[25]. Therefore the mitigation mechanisms that were already deployed during the first wave had to be set to a global scale. Only by 17:00 Dyn recovered substantially, still having impact from other sources until 20:30 UTC. In the next few days further smaller incidents happened which Dyn effectively was able to mitigate using above mitigating mechanisms, and thus further customer impact could be prevented. [26]

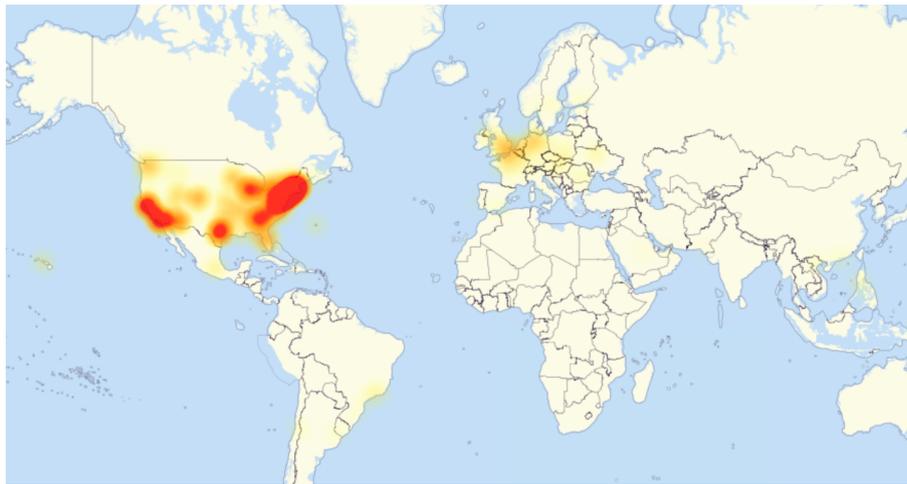


Figure 3.15: Globally Scattered impacts

A technical difficulty that Dyn faced during the mitigation of this attack was to distinguish between legitimate traffic and attack traffic. The DNS congestion created by the attack increased legitimate traffic volume as well. According to Dyn's analysis the normal traffic volume got increased by a factor 10-20 only because of legitimate retry activity of recursive servers that attempted to refresh their caches. That also made it difficult at the time of the attack to estimate a concrete number of malicious endpoints [26]. The code for the Mirai malware had been published open source in the internet a few weeks before the incident. This fact contributed further to the difficulty of investigation of adversaries [28].

3.5.6.2 Communication

To alleviate damage to reputation, as well as to customer and public relations, Dyn had to communicate effectively. As a matter of fact efficient PR-related work becomes very important for an incident with such dimensions. In various blogposts of Dyn the current state of investigation is described.[27][26] In addition to that those blogposts mainly try to clarify ambiguities and resolve rumours. Certainly, this is a form of preventing costs by the loss of trust of customers and partners.

In an early blogpost on the day after the attack, Chief Strategy Officer Kyle York writes that Dyn is practicing on a regular basis for such scenarios and mentions the tremendous effort that is being made by customers and the internet community. Moreover, the severity of the attack is stated. The attack is described as a "sophisticated attack across multiple attack vectors and internet locations". He also writes that they observed 10's of millions of different IP-addresses that were associated with the Mirai-Botnet. Another important point stated is that Dyn is working closely together with partners for mitigation. [27] In a further blogpost on Dyn's website five days after the attack they provide key findings from their analysis. They express again their appreciation for the great support of their partners, as well as their customers and all internet infrastructure stakeholders. Furthermore they state that Dyn's case has opened up a discussion about the internet infrastructure and its security in the future. Especially as in the case of the Mirai botnet, security in the internet of things and the general internet infrastructure must be considered. They express their will to contribute to developments in those areas jointly with the internet community. [26]

3.5.6.3 Collaboration

Unfortunately detailed technical information on collaboration and collaborative mitigation strategies are not made public and therefore also economic efforts cannot be assessed

reasonably. But the benefits of collaboration seem to be extraordinary considering the globally diverse properties and effects of the attack.

Collaborations that are well known are with the business risk intelligence companies Flashpoint and Akamai. Their efforts made it possible to confirm that the attack was in connection with the mirai botnet, and they partnered in further forensic investigation. [29]

Dyn's DNS lookup service is embedded in services that Dyn's clients in turn offer to their customers. As already described in previous chapters, relations to a range of stakeholders must be considered from an economic viewpoint. Therefore effective communication and collaboration is of uttermost importance. Even companies not directly affected by the attack can help in mitigation. For example the attack did not have direct impacts on the services provided by the CDN provider Akamai. Nevertheless they could help their clients that were utilising Dyn Managed DNS. Using a technique that reuses valid DNS resolutions from the past, they were able to continue identifying client locations effectively. This helped in resolving around 60'000 DNS requests per second. Without Akamai's help this load would have been left to Dyn's servers, and certainly a lot of damage was prevented. [30] However, as further information on clients from akamai is not available economic damage prevention can not be estimated numerically in a meaningful manner. As already pointed out by the previous section, Dyn is determined to contribute to developments for a secure internet in the future. In February 2018 one of the largest DDoS attack in the history of computing was launched on Github. Surprisingly it could be stopped about 20 minutes after. A storm of incoming traffic of 1.3Tbps was routed to Akamai Prolexic, a DDoS mitigation service provided by Akamai. [31] Akamai was definitely able to learn from Dyn's case and therefore economic benefit from security investment propagated to the future. Sharing cyber security incidents to promote research and increase understanding nation-wide or even globally is also a kind of collaboration which is less tight and further could be seen as a prevention measure.

3.5.6.4 Prevention Measures

Even though incident response mechanisms could be deployed successfully, Dyn must have failed at some point to implement appropriate prevention measures. A DNS system is a central part of today's web and should ideally not be subject to failure. It definitely must be able to robustly scale up to a huge amount of requests and must employ efficient strategies to block a DDoS. Obviously, especially in the internet of things this for sure is a huge challenge, as even with good prevention measures employed the system might just not be able to distinguish legitimate from malicious traffic fast enough and thus congestion occurs. The consequence is that questions about scalability, load balancing and security against DDoS become intertwined. In November 2016 Oracle's president, product development Thomas Kurian described Dyn's DNS as "immensely scalable", but obviously it was not enough. [33] Prevention measures taken by the company under assault go hand in hand with the actions performed as incident response. Dyn's traffic shaping and packet filtering endeavors were definitely not sufficient, as their services became temporarily completely inaccessible. Unfortunately, Dyn does not disclose their concrete prevention mechanisms at the time of the attack. But certainly there were some improvements and countermeasures put into place. An overview is given in the next section.

3.5.6.5 Improvements and Countermeasures

A month after the attack, Dyn announced that the company will be acquired by Oracle Corporation. According to Thomas Kurian was Dyn's global DNS a "natural extension" of Oracle's Cloud computing platform.[33] The integration with Oracle's cloud and the

experience that was taken from the incident in October 2016 led to the strategies of DDoS mitigation that are used today. Today Oracle Dyn provides a full protection stack against DDoS. According to Oracle's vision an integrated approach is needed that employs multiple layers to protect against DDoS attacks against applications, networks and DNS. They have a cloud-based, managed DNS that protects against DNS-infrastructure. Furthermore they offer a DDoS protection solution which is specifically designed for the layers 3 and 4. For the application layer, a WAF (Web application firewall) bot manager service is offered to protect against DDoS targeting web applications. [32] The protection stack is purely cloud-based and thus employs protection by redundancy.

3.5.7 Costs and Overall Assessment

Normally in step 4 of SEconomy, the risk analysis is taken to map costs appropriately. Costs should be evaluated for every task performed by an actor and associated with a risk. The estimation of overall cost is usually done in the last step. However, the costs of proactive approaches employed by Dyn (training of employees/education, traffic shaping, and redundancy of critical systems) as well costs of Dyn's reactive approaches (monitoring and recovery) cannot be quantified effectively. Inputs needed to calculate threat exposure costs, Pro- and reactive mitigation costs and return on security investments (ROSI) are not available. Also the mapping of all the mitigation efforts to specific actors is not straight forward as the attack was dealt with collaboratively. For instance quantifying the amount of impact prevented by Akamai which is described in the section on collaboration is simply not possible. We would need to calculate the economic impact on every customer of Akamai separately. Additionally we can not estimate the impact that would have been left to Dyn without Akamai's help.

From an ex post perspective of the attack, investments might seem to be just additional cost. But certainly, without Dyn's early security investment, the server overload could not have been dealt with so fast. Also considering the fact that Dyn took the right measures to collaborate with others might justify that they did not invest further in defense of DDoS beforehand. Moreover those early security investments provided the base for further investments and improvements of Dyn's defense capabilities against DDoS.

As already stated in previous chapters, the attack had diverse impacts across various industries and on a lot of Dyn's customers. As customers of Dyn's clients could not access their websites, direct impact was mostly on their side as they were not able to offer their usual services. Dyn was paid on an annual basis and hence the company did not face extensive direct costs, except from the costs of incident response and mitigation. However, from a pure financial viewpoint, it can be argued that the direct damage in form of lost income of Dyn's customers propagated into later incurring costs on Dyn's side. Dyn lost approximately 8 % of their customers as a result of the attack [40]. Further, the attack might have influenced the price to which the company was sold to the Oracle Corporation a month later. But it is not known whether the contract about the acquisition was negotiated before the attack and therefore this statement should be taken with a grain of salt. From the point of view of business optimization the economic impact on Dyn was definitely dominated by the loss of reputation and trust. Taking a macro-economic perspective, we can only speculate about the overall economic damage of this DDoS attack. What can be said with certainty is that this attack has raised awareness about the need to develop efficient concepts to deal with DDoS attack. The dialogue on internet security has definitely become much more intense and knowledge gained from the incident made it possible to develop today's mechanisms of DDoS mitigation.

3.6 Outlook

With the growth of the internet markets, with more and more segments coming into play, DDoS will also evolve. This makes important for every risk manager to take into consideration future developments of the internet to anticipate and prevent the security concerns of tomorrow.

3.6.1 5G

According to service providers' advertisements[50], 5G-enabled devices will have an available speed of 2Gbit/s up to 10 Gbit/s. This will drastically increase the amount of traffic a single internet-enabled device can receive from or send to a service provider - even exceeding the amount that was possible with the subscriptions based on carbon fiber. As much of an improvement this may be for industries like streaming platforms, this is also a great challenge for many service providers as now a single device is able to send traffic at a speed that was before only available to data centers. An existing botnet with 5G enabled would then gain massive power (e. g. from 100MBit/s to 1Gbit/s this would be factor 10). Therefore, service providers must take into consideration this *potential* immense increase of traffic in case of an attack.

3.6.2 IoT

3.6.2.1 General danger about IoT devices

With more and more devices connected to the internet, the number of potential attackers (intentional or unintentional due to malware making a device part of a botnet) will rise quickly. Every device with access to the internet needs proper security or it may be vulnerable. This will be challenging for different parties:

- Vendors: Need to include risk management into consideration for products that previously did not need *any* security risk management at all when enabling internet access.
- Users: Need to maintain more devices, performing updates if they are not done automatically and in case of mistrust need to renounce on more features
- Supporters: Need to be aware of a wider range of product vulnerabilities and be better educated for potential flaws in systems

3.6.2.2 Security issues in the past

However this did not really work out yet in the past, as the so-called "Mirai Botnet" could use standard-secured IoT devices to build up a botnet of 400'000 devices (see the chapter case study for further information) and one of it's successors called "Satori" could even compromise up to 700'000 devices. [46] The reason were always default users allowing to install software and users not changing this admin user's password because it was not obligatory. A problem that would be relatively easy to solve or at least make harder - either make random, strong, non-derivable default admin passwords like for wireless routers or force the user on the software side to change the password before he gains write access.

3.6.2.3 No patch for existing vulnerabilities

As always the implementation of such security measures would be bound to costs on the manufacturer side, which may be one of the reasons it was not done. Also according to Krebs, the manufacturer of many of those insecure devices was Xiongmai, a china-based manufacturer which did not exchange these insecure devices [47], which means this is still a problem as the end user most likely does not care about it as long as his device works properly for the designated use case (in this case the security camera delivering its data).

3.6.3 Open Source Malware

What played into cards of the "Satori" developer before was the public availability of the Mirai source code on github.[48] This allowed everybody to have a good basis to further develop it, react to new vulnerabilities in other devices and quickly create an even more powerful botnet. On the other hand the owners of the affected systems probably did not even know about being affected, neither how to react (e. g. changing the password). Also on the other hand more and more devices are entering the market - if they are not properly secured allowing for even more devices to be compromised by making minor adaptations to existing code.

This throws up the question whose responsibility it is to prevent malicious code from getting into use - github is on the safe side due to its terms of service stating that every user is legally responsible for the content he provides in his repository. However it can also be argued that making malware publicly available increases pressure on manufacturers and service providers to fix the vulnerabilities allowing them. Also it's of educational use to have malware code available to better get an understanding of how it works.

Releasing malware code, normally does not make sense however if the developer plans to use it commercially - releasing the code would then allow basically anybody with the necessary technical know-how to compete with the developer. In case of pure malicious intent however, it can make much sense as this will allow anybody to join as attacker.

3.6.4 X-as-a-Service

Seven out of ten companies already use the cloud for one or more of their services and the number is still growing, as in 2011 it was just five out of ten. Splitting it up into Software-, Platform- or Infrastructure as a Service, half of the cloud expenses are for Software as a Service.[49] In X-as-a-Service scenarios it is common to sign a service level agreement (SLA) between the two parties in which the service provider commits himself to a certain up-time of the service (normally 99%+). Therefore the service provider is in charge of risk management to fulfil this guarantee, as otherwise the service consumer could make him accountable for further outages and depending on the agreement even for data leaks or other security issues. This allows for service consumers to focus more on their core business, while the cyber security risk management can be centralized by service providers which may themselves even outsource this further to highly specialized companies. This kind of specialization makes sense from a macroeconomic point of view by potentially reducing overall costs for cyber security.

3.6.5 Software Defined Networking (SDN)

The idea behind Software Defined Networking is to centralize controlling logic to a controlling unit, which then decides what to do with the data (e. g. forwarding to a data processing node). This allows for centralizing the knowledge about packages to the controller, which may talk to other controllers and implement some mitigation against DDoS

and other attacks. However on the other hand it does again provide a single point of failure, also for security - which may as well be against the current measures taken with different layers of security, e. g. web application firewalls filtering traffic, switches distributing it to other switches or end nodes etc. On the other hand, it can reduce costs and improve maintainability by replacing physical devices with software.

3.7 Concluding Considerations

Risk management in information systems is a very complex challenge. Therefore, many of the standards and frameworks for risk assessment and management are very detailed and hard not only to understand, but also to apply. With Seconomy, the university of Zurich has developed a quite straight forward framework which is relatively easy to apply and covers the most important aspects of risk identification, assessment and management. However, this framework still relies on the availability of information and data, which is not always given in case of DDoS attacks. In fact, in today's data-driven world, the field of cybersecurity is adversely affected by insufficient data, largely because companies face a strong disincentive to report negative news. The more information is missing, the less concise analytics can be performed or even decisions be taken. Therefore, we think it would make sense to work closer together with an affected company for a further case study instead of applying it retrospectively. This would surely allow to gain more detailed information on how Seconomy could be integrated into existing companies' risk management concepts.

Bibliography

- [1] *From Disaster Scenario to Reality: Modeling the Dyn Cyber Attack, October 2016* <https://www.air-worldwide.com/Blog/From-Disaster-Scenario-to-Reality--Modeling-the-Dyn-Cyber-Attack/>
- [2] *2019 Data Breach Investigations Report Verizon Enterprise* <https://enterprise.verizon.com/resources/reports/dbir>
- [3] *DDoS attacks in Q1 2019 Securelist* <https://securelist.com/ddos-report-q1-2019/90792>
- [4] *Cloud in the crosshairs - Netscout* https://www.netscout.com/sites/default/files/2019-03/SECR_005_EN-1901%E2%80%93WISR.pdf
- [5] *Consumer Study on Aftermath of a Breach Ponemon* <https://www.ponemon.org/local/upload/file/Consumer%20Study%20on%20Aftermath%20of%20a%20Breach%20FINAL%20.pdf>
- [6] Zargar, Saman Taghavi, James Joshi, and David Tipper "A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks." *IEEE communications surveys & tutorials* 15.4 (2013): 2046-2069.
- [7] "State of the Internet: Security DDoS and Application Attacks Executive Summary" <https://www.akamai.com/us/en/multimedia/documents/state-of-the-internet/state-of-the-internet-security-ddos-and-application-attacks-executive-summary-2018.pdf>
- [8] What is a DDoS Attack? Online, URL: <https://www.cloudflare.com/learning/ddos/what-is-a-ddos-attack/>
- [9] Understanding DDOS Attack Online, URL: <https://medium.com/@kapil.sharma91812/understanding-ddos-attack-15dd2cbce2a>
- [10] *Distributed Denial of Service (DDoS) by Rich Groves, Eric Chou, Publisher: O'Reilly Media, Inc., Release Date: April 2018* <https://learning.oreilly.com/library/view/distributed-denial-of/9781492026181/>
- [11] Distributed denial of service attack (DDoS) definition <https://www.imperva.com/learn/application-security/ddos-attacks/>
- [12] DDoS attacks and defense mechanisms: classification and state-of-the-art by Christos Douligeris *, Aikaterini Mitrokotsa October 2003 <http://www.cse.chalmers.se/~aikmitr/papers/COMNET.pdf>
- [13] Martin Waldburger, Patrick Poullie, Burkhard Stiller: *Guideline for Seminar Reports*, Communication Systems Group, Department of Informatics, University of Zurich, January 2013. <http://www.csg.uzh.ch/teaching/guideline-seminar-report-v05.pdf>.

- [14] Aishwary Sreekanth, Prashant Sri, Teemu Vartiainen: *Dyn DDOS Cyberattack - a case study*, ..., Department of Informatics, Aalto University, Summer 2016.
- [15] <https://www.thesslstore.com/blog/largest-ddos-attack-in-history/>.
- [16] Sonnenreich, Wes and Albanese, Jason and Stout, Bruce *Return On Security Investment (ROSI): A Practical Quantitative Model.*, Journal of Research and Practice in Information Technology, p. 239-252, January 2005
- [17] Bruno Rodrigues, Muriel Franco, Geetha Parangi, Burkhard Stiller: *SEconomy: a Framework for the Economic Assessment of Cybersecurity*, Communication Systems Group, Department of Informatics, University of Zurich, October 2019.
- [18] NIST: *Guide for Applying the Risk Management Framework to Federal Information Systems: A Security Life Cycle Approach*. Tech. rep., National Institute of Standards and Technology (NIST), February 2010
- [19] *Security/OSSA-Metrics: DREAD*, OpenStack, November, 2019 <https://wiki.openstack.org/wiki/Security/OSSA-Metrics/#DREAD>
- [20] *The cost of Cybercrime*, Accenture, 2019 https://www.accenture.com/_acnmedia/pdf-96/accenture-2019-cost-of-cybercrime-study-final.pdf
- [21] *Advanced DDoS Protection*, Cloudflare, November, 2019 <https://www.cloudflare.com/ddos/>
- [22] *The STRIDE Threat Model*, Microsoft, November 2009 [https://docs.microsoft.com/en-us/previous-versions/commerce-server/ee823878\(v=cs.20\)?redirectedfrom=MSDN](https://docs.microsoft.com/en-us/previous-versions/commerce-server/ee823878(v=cs.20)?redirectedfrom=MSDN)
- [23] *What is the Mirai Botnet?*, Cloudflare, November 2019. <https://www.cloudflare.com/learning/ddos/glossary/mirai-botnet/>
- [24] Manos Antonakakis et al: *Understanding the Mirai Botnet*, usenix, August 2017. <https://www.usenix.org/system/files/conference/usenixsecurity17/sec17-antonakakis.pdf>
- [25] *CenturyLink outage map*, Downdetector, October 2016. <https://downdetector.com/status/centurylink/map/>
- [26] Scott Hilton: *Dyn Analysis Summary Of Friday October 21 Attack*, Oracle Dyn, October 2016. <https://dyn.com/blog/dyn-analysis-summary-of-friday-october-21-attack/>.
- [27] Kyle York: *Read Dyn's Statement on the 10/21/2016 DNS DDoS Attack | Dyn Blog*, Oracle Dyn, October 2016. <https://dyn.com/blog/dyn-statement-on-10212016-ddos-attack/>.
- [28] Nick Statt: *How an army of vulnerable gadgets took down the web today*, The Verge, October 2016. <https://www.theverge.com/2016/10/21/13362354/dyn-dns-ddos-attack-cause-outage-status-explained>.
- [29] *An After-Action Analysis of the Mirai Botnet Attacks on Dyn*, Flashpoint, October 2016. <https://www.flashpoint-intel.com/blog/cybercrime/action-analysis-mirai-botnet-attacks-dyn/>.

- [30] MÃ©lodie Reynaud: *INTERVIEW: AKAMAI AT THE HEART OF THE RECENT DDOS ATTACKS*, FIC, July 2017 <https://observatoire-fic.com/en/akamai-at-the-heart-of-the-recent-ddos-attacks/>.
- [31] Ai Lei Tao: *How traffic scrubbing can guard against DDoS attacks*, ComputerWeekly, January 2019. <https://www.computerweekly.com/news/252456702/How-traffic-scrubbing-can-guard-against-DDoS-attacks>.
- [32] *DDoS Mitigation*, Oracle Dyn, 2019 <https://dyn.com/ddos/>.
- [33] Ingrid Lunden: *Oracle acquires DNS provider Dyn, subject of a massive DDoS attack in October*, techcrunch, November 2016. <https://techcrunch.com/2016/11/21/oracle-acquires-dns-provider-dyn-subject-of-a-massive-ddos-attack-in-october/>.
- [34] DDoSPedia: *Operation Ababil*, radware, 2019. <https://security.radware.com/ddos-knowledge-center/ddospedia/operation-ababil/>.
- [35] Samuel Burke: *Massive cyberattack turned ordinary devices into weapons*, CNNTech, October 2016. <https://money.cnn.com/2016/10/22/technology/cyberattack-dyn-ddos/index.html>.
- [36] Rich Bolstridge: *DYN DDOS ATTACK: WIDE-SPREAD IMPACT ACROSS THE FINANCIAL SERVICES INDUSTRY (PART 1)*, Akamai, Octobre 2016. <https://blogs.akamai.com/2016/10/dyn-ddos-attack-wide-spread-impact-across-the-financial-services-industry-part-1.html>.
- [37] Rich Bolstridge: *DYN DDOS ATTACK: LESSONS LEARNED FOR THE FINANCIAL SERVICES INDUSTRY (PART 2 OF 2)*, Akamai, Decembre 2016. <https://blogs.akamai.com/2016/12/dyn-ddos-attack-lessons-learned-for-the-financial-services-industry-part-2-of-2.html>.
- [38] corero: *DDoS Protection for Service Providers*, corero, 2019. <https://www.corero.com>.
- [39] *DDoS impact survey*, Professional Security Magazine Online, March 2016. <https://www.professionalsecurity.co.uk/products/computer-systems-and-it-security-news/ddos-impact-survey/>.
- [40] Stephanie Weagle: *Financial Impact of Mirai DDoS Attack on Dyn Revealed in New Data*, corero, February 2017. <https://www.corero.com/blog/797-financial-impact-of-mirai-ddos-attack-on-dyn-revealed-in-new-data.html>.
- [41] Chris Gonyea: *What Is DNS Load Balancing Why Is It Important?*, Oracle Dyn, July 2013. <https://dyn.com/blog/what-is-dns-load-balancing-why-is-it-important/>.
- [42] *Famous DDoS Attacks | The Largest DDoS Attacks Of All Time*, Cloudflare, 2019. <https://www.cloudflare.com/learning/ddos/famous-ddos-attacks/>.
- [43] L. Rabe: *Umsatz von Amazon in Deutschland und weltweit in den Jahren 2010 bis 2018*, statista, August 2019. <https://de.statista.com/statistik/daten/studie/374731/umfrage/nettoumsatz-von-amazon-in-deutschland-und-weltweit/>.

- [44] L. Rabe: *Statistiken und Daten zu PayPal*, statista, May 2019. <https://de.statista.com/themen/2499/paypal/>.
- [45] Jackson Pacheco: *Netflix Users Watch 165 Million Hours Per Day, Study Says*, Paste, April 2019. <https://www.pastemagazine.com/articles/2019/04/people-watch-165-million-hours-of-netflix-per-day.html>.
- [46] *âSatoriâ IoT Botnet Operator Pleads Guilty*, Krebs on Security, September, 2019 <https://krebsonsecurity.com/2019/09/satori-iot-botnet-operator-pleads-guilty/>
- [47] Krebs on Security: *Naming & Shaming Web Polluters: Xiongmâi*, Krebs on Security, October, 2018 <https://krebsonsecurity.com/2018/10/naming-shaming-web-polluters-xiongmâi/>
- [48] *Mirai Source Code*, Github/jgamblin, November, 2019 <https://github.com/jgamblin/Mirai-Source-Code>
- [49] *State Of Enterprise Cloud Computing, 2018*, Forbes, August, 2018 <https://www.forbes.com/sites/louiscolombus/2018/08/30/state-of-enterprise-cloud-computing-2018/#7173d2f7265e>
- [50] *The best network. Now with 5G*, Swisscom, October, 2019 <https://www.swisscom.ch/en/about/company/portrait/network/5g.html>,
- [51] Martin Waldburger, Patrick Poullie, Burkhard Stiller: *Guideline for Seminar Reports*, Communication Systems Group, Department of Informatics, University of Zurich, January 2013. <http://www.csg.uzh.ch/teaching/guideline-seminar-report-v05.pdf>.

Chapter 4

An Overview of Cyber Insurance Models

Christian Birchler, Michael Nadig, Sandro Padovan and Louis Preisig

When there are risks in a business, there often are corresponding insurance products for it. Usually, companies have a fire insurance and they have to accomplish requirements defined by the insurance company to minimize the risk and also to reduce the premiums. Companies can get damaged not only through fire, but also through cyber crimes. The number of cyber crimes is increasing dramatically e.g., DDoS attacks or phishing E-mails. Companies can be a target for such crimes and are vulnerable. The world is in a movement of digitization and it is likely that the revenue of a company is sensitive for such cyber attacks. It might be logical to also have insurance for damages based on cyber attacks. In this seminar report, we summarize the challenges to establish a functioning market for cyber insurance products. Further, we present existing cyber insurance products and show their similarities and differences. The goal is to achieve a better understanding of the need for cyber insurance products in the context of cyber threats and its economic influence.

Contents

4.1	Introduction and Problem Statement	101
4.1.1	Framework and Stakeholders	101
4.1.2	Problems	101
4.2	Related Work	101
4.2.1	Correlated Risks	101
4.2.2	Inadequate Reinsurance Capacity	102
4.2.3	Information Asymmetry	103
4.2.4	Lack of breach disclosures	104
4.2.5	Inexperience	105
4.2.6	Free-Rider problem	105
4.2.7	Moral Hazard	106
4.3	Empirical data and current state of research	106
4.3.1	Cyber resilience	106
4.3.2	Risk Analysis	108
4.4	Data Collection	110
4.4.1	NetDiligence Main Findings	110
4.5	The cyber insurance market	112
4.5.1	Business Insurance	113
4.5.2	Personal Insurance	115
4.6	Solutions	117
4.7	Evaluation and Discussions	117
4.7.1	Overpricing	117
4.7.2	Low policy limits	118
4.7.3	Further research	118
4.8	Summary and Conclusion	118

4.1 Introduction and Problem Statement

With the growth of the internet, security breaches have become an everyday occurrence in today's world. Every single internet user is exposed to a variety of different risks associated with computers. Security systems become better on a daily basis and the topic of security breaches seem to be omnipresent especially when big companies or celebrities are involved. If you have an online business and you get hit by such attacks it can have devastating effects. A possible solution to such breaches could be cyber insurances. Yet despite the gigantic presence of regular insurances, cyber insurances are lagging behind other insurance models and are fairly unpopular or even unheard of by many.

In this paper we will analyse as to why we are seeing such a lack of cyber insurances and what the current problems might be. Those problems can be because of economic, scientific or other circumstances. We will see why we might still see a massive growth despite the current and past problems. We will also briefly analyse how different threats can affect a company and why it can be beneficial for an individual or an entity to use such insurances and what problems they could solve. We will also look at the currently existing work concerning risk analysis when it comes to cyber insurances to see, in what kind of scenarios companies could insure themselves against those risks coming from computers.

4.1.1 Framework and Stakeholders

Cyber insurances exist both for individuals and companies to limit the risk and therefore we will cover insurances for those two. We also limit ourselves to cases and companies mostly in Europe or north America as there is little information or research from anywhere else and we do not consider those to be relevant enough for this paper.

4.1.2 Problems

There are various problems with cyber insurances which still have to be tackled and will be an issue in the future. One clearly being that the current IT-systems are heavily correlated which can lead to chain reactions and massive sums of money being lost just because of one failure or breach of security. Insurance companies also rely on reinsurance companies to spread the risk however there is a big lack of at the moment. Those just mentioned and even more problems which we will dive deeper into in a later section dedicated to the problems of the market have been an issue in the past and will be a burden in the future as well. We also will see how those problems might be tackled in the future.

4.2 Related Work

The CSI/FBI survey in 2004 which asked several companies for their cyber security measures clarified that only 25 percent of the respondents purchased cyber insurance policies from underwriters [12]. In year 2008 the CSI/FBI survey results are based on 522 responses of various U.S. companies and institutions which shows that 34 percentage of the respondents have external cyber insurance policies [22]. This increase seems not to be significant over the past years which can be explained by several challenge aspects that the cyber insurance market has to face. The sections 4.2.1 till 4.2.7 describe the most known challenges in the cyber insurance market according to the current state of research.

4.2.1 Correlated Risks

The main idea behind insurance companies is that they must have a big policy holder base. In the general case of insurances, the claims are uncorrelated which means that the fact

that one policyholder files a claim does not necessarily mean that also another policyholder will file a claim. Due to this uncorrelation the insurance companies can handle the individual claims without going into the risk that they cannot payout the customers. However, in the case of cyber insurance the market behaves differently. Since many IT-systems are interconnected there is a higher probability that the insurance companies receive multiple claims which are due to the same cyber attack at once. This, of course, can bring the underwriters into trouble by paying out the customers, because the insurance companies might not be liquid enough [5].

A good example is the case of a ransomware like WannaCry. WannaCry affected over 300'000 systems in over 150 countries. This infection could happen because there was a security bug in Microsoft Windows. Since the most systems run on Windows the range of the affected system was broad. Fields like the financial industry, the health care or the telecommunication branch were affected [1]. Broad attacks like WannaCry will lead to a significant amount of claims which the cyber insurance companies might deal with.

Another example of a correlated risk is the the case of the ILOVEYOU virus. This virus was distributed via email and its subject title was „ILOVEYOU“. This email had an attachment which was named as a love letter. Also here, the virus exploited a Microsoft Windows bug which makes it easier to infect a larger amount of computers. Approximately 45 millions desktop PCs were infected by this virus [28]. Further, the estimation of the damage is going to be roughly 10 billions USD [25].

The interdependence between IT-systems is one of the main reasons why the claims in cyber insurance are often correlated. Nowadays, many software depend on other software namely libraries and frameworks. Such dependencies are high risk factors in IT-security. The example of the npm package „left-pad“ has shown that already a small piece of code which many other systems depend on it can produce a big damage if it will be unpublished. Basically in this case a naming conflict occurred of a npm package so that whose developer removed the said package with 272 other packages. One of the packages was the „left-pad“ package which was crucial for dependent systems. This lead to unavailability of many systems and produced therefore correlated damages [27].

4.2.2 Inadequate Reinsurance Capacity

The cyber insurance market faces also the problem that there is an inadequate reinsurance capacity. Reinsurers are basically insurance companies which insures other underwriters. The idea behind this construction is that in case of a natural disaster, which leads to many claims, the insurance companies are still able to pay out the policy holders. This is usually done by spreading the risk globally to reinsurance companies so that it is very unlikely that due to an event (e.g. natural disaster) the reinsurance companies are unable to cover the damage [5].

Nowadays, the reinsurance companies use investment tools like bonds to spread the risks even further. With bonds the reinsurance companies can acquire more capital from other markets. The existing bonds are usually called „cat bonds“ since those bonds are mostly for catastrophic events. If one of the well defined catastrophic events occur then the reinsurer can keep back the interest of the investors for paying out the claims. Sometimes the bonds allow the reinsurance companies to keep the invested capital to ensure the coverage of the claims. As mentioned before those bonds are mainly so called cat bonds which are mainly an instrument for getting a larger amount of capital for covering only claims due to natural catastrophic events. In case of cyber insurance there exist no such bonds [5, 19].

In the case of cyber insurance the market seems to be different. As mentioned in section 4.2.1 above, damages caused by cyber attacks are likely correlated. In this situation a problem with reinsurers occur. Spreading the risk of cyber attacks globally is harder or

even not possible since the damages which occur due to cyber attacks might be globally like the ILOVEYOU virus or the WannaCry ransomware. Such attack can produce high losses which can not be full covered by the insurers and reinsurers [5].

The conclusion of having an inadequate reinsurance capacity is that it will affect the price for cyber insurance policies. Due to the fact that it is hard to spread the risk globally for cyber attacks caused damages, the insurer and reinsurer have problems in minimizing their own financial risks. According to this issue the underwriters usually offer only contracts with low policy limits. Such low policy limits are very unattractive for large IT-intensive companies which would like purchase a cyber insurance policy.

4.2.3 Information Asymmetry

A common term in the field of insurances is „Information asymmetry“. The information asymmetry refers to a common problem which insurances go into. Information Asymmetry is defined as follows:

„Situation that favors the more knowledgeable party in a transaction. In most markets (especially where the goods being traded are of uncertain quality, such as used equipment), a seller’s is usually in a more advantageous position because his or her store of information is based on numerous sales conducted over the years. A buyer’s information, however, is based usually on an experience of only a few purchases. A similar situation exists between a commercial lender and a borrower“ [9].

To clarify the information asymmetry there are many simple examples. One of these examples is the car market. Let say there exist new and old cars on the market and those cars are either good or bad. Further, the new cars are more expensive than the old cars which is quite obvious. To decide if a car is good or bad, the owner need a certain amount of experience. So a potential buyer of a new car does not know if it is a good or bad car but the buyer has a probability estimate \hat{p}_0 to buy a good car and $\hat{p}_0 - 1$ to buy a bad one. An owner of a new car will get an accurate estimate \hat{p}_1 after experiencing the car for a while. With the new estimate an information asymmetry can be developed since the seller has usually more experience with the car than the buyer who will possibly pay more for the car than its actual value [2]. According to the mentioned definition of information asymmetry above, the car seller takes here clearly an advantageous position.

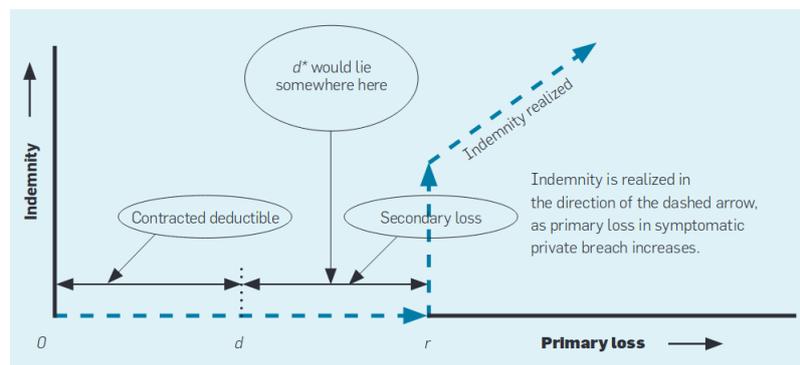


Figure 4.1: Relationship between de facto deductible r and realized indemnity l .

The market of cyber insurance faces also the challenge of information asymmetry. This is mainly due to the fact that the so called secondary loss cannot be perfectly quantified. Secondary losses are for example a diminished reputation, goodwill, consumer confidence etc. As illustrated in [6](cf. Fig. 4.1), the realized indemnity for a policyholder differs to the paid indemnity of the insurer’s point of view. The insurer has the assumption that the policyholder realize the indemnity already at point d . The difference between d and r is the secondary loss. So the secondary loss takes an important role in the analysis of

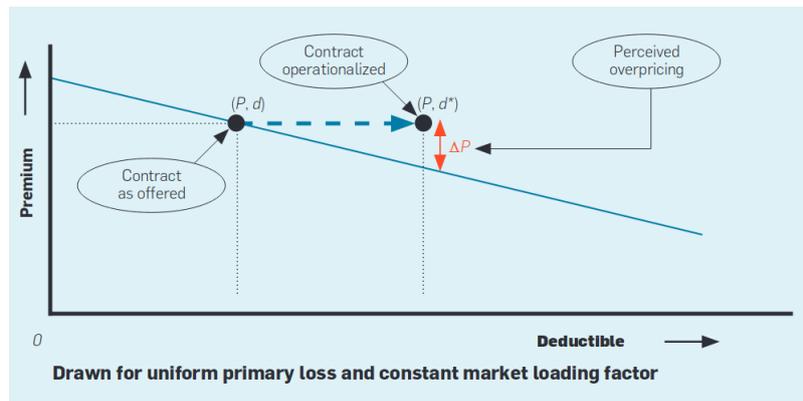


Figure 4.2: Perceived overpricing of a cyber-insurance contract.

cyber insurance contracts. Before claiming a breach, IT managers will balance between the indemnity payout and the sum of the primary loss and the secondary loss. Further, the choice of the deductible is essential for calculating the premiums. These two measures define at the end the insurance contract. In the second figure [6](cf. Fig 4.2) it is possible to observe that due to the secondary loss, which introduce an artificial higher deductible, will lead to an overpriced contract. Therefore, an estimation of the secondary loss must be as much accurate as possible otherwise the pricing of the cyber insurance contract will be imperfect [6].

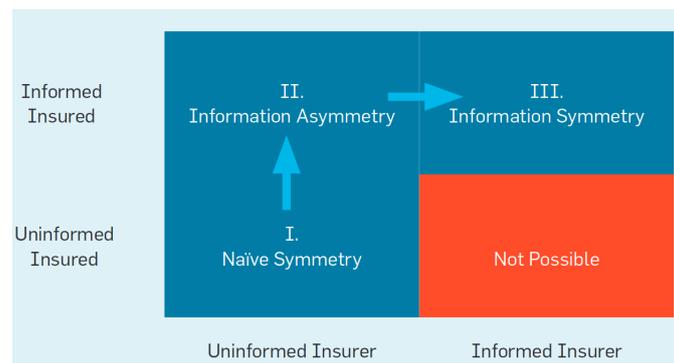


Figure 4.3: Information asymmetry and market transition.

Next to the term of information asymmetry there exist a naïve symmetry and an information symmetry. In the illustration [6](cf. Fig. 4.3) there are different states in which policyholders and insurers can be in. Both parties are in the naïve state if none of them is aware of the secondary loss of a breach. If the policyholder is aware of the existence of a secondary loss but the insurer is not then the parties are in the state of information asymmetry. In the case that the insurer is also aware of the secondary loss then the parties are in the state of information symmetry. Moreover, the state that only the insurer is informed about the secondary loss is considered as not possible [6].

To reduce the information asymmetry the insurance companies typically apply some mechanism (e.g. surveys). However, it is nearly impossible to eliminate completely the information asymmetry, since risk depends on too many factors. Due to this fact a premium differentiation is difficult and thus the price is not optimal for the policyholders [5].

4.2.4 Lack of breach disclosures

A problem for insurance companies is the lack of information about cyber threats. Some measures like the Security Breach Information Act in California, which went into effect in 2003, provide next to the actual goal to have more security for personal data, also the

possibility to get more information about the cyber threats [5]. This law in California applies only for security breaches where personal data of a Californian resident were leaked. Moreover, the fast moving environment for cyber risks makes collecting historical data about security breaches almost unnecessary [11]. The idea to enforce disclosures by law for all kind of security breaches seems to have only a little political support [5].

Many companies with cyber insurance contracts do not disclose all of their cyber security breaches to its insurers. The reason for such behaviour is the secondary loss which can be higher than the realized indemnity (cf. section 4.2.3). Especially for larger companies which receive their capital from stock exchanges, security breaches can take a negative influence on the stock price in the short term [6]. The CSI/FBI Computer Crime and Security Survey 2005 shows that only 63 percent of the respondents did share their security breaches. Further, the percentage of the companies which disclose their breaches shrank slightly in the duration between 1999 and 2005 [12]. This development shows that the problem of having less information about security breaches is getting more serious.

Possible cases of secondary loss which leads the company not to disclose the breach is the loss of customer's confidence [6]. This particular loss is high especially for IT-intensive firms whose costumers expect that their cyber security measures do not allow such security breaches. A good example how a IT-intensive company can have high secondary loss is the case of Swisscom with the MyCloud product. MyCloud is a cloud storage platform which had some programming issues. MyCloud deleted irreversible some of the costumers data. Swisscom did targeted advertising by saying the data are safe because they are only on Swiss servers. The loss of data reduced the confidence in storing data on MyCloud and further the image of „Swissness“ was damaged. This kind of secondary loss is high but not quantifiable. The breach get published a year after it has occurred [14]. It is clear that Swisscom did not have the intention to disclose this case because the secondary loss would be high.

The future direction should be to have a better information base [5]. In Switzerland there is a platform, „Melde- und Analysestelle Informationssicherung“ (MELANI), where private persons and companies can disclose cyber security breaches. Further, MELANI provides more detailed reports and gives news about the recent security issues. Nevertheless, reporting security breaches on MELANI is not mandatory and it provides only a compact pool of information about computer security issues [24].

4.2.5 Inexperience

Since the market of cyber insurances is quite young in comparison to other insurance markets (e.g. fire insurance market), there is also the problem of having inexperience in this new market field. Usually underwriters have experience in their tangible asset classes like vehicles or properties. Data and information which are being damaged through cyber security breaches are not considered as a tangible property. There were many legal disputes about the tangibility of electronic data which is a reason why insurers write exclusion terms for electronic data in the policy contracts [5]. Those exclusion terms led to an slower growth of experience in evaluating such intangible property. This inexperience is also connected to the issue of having less public breach disclosures (cf. 4.2.4). In general the inexperience leads to conservatism pricing of the cyber insurance contracts [6].

4.2.6 Free-Rider problem

A problem which occur more in general on the insurance market is the free-rider problem. The free-rider problem is the case if one party benefits from a service which is paid by the community. This phenomena can occur whenever a collective good can be shared and it's not possible to exclude a single individual [29]. A good example is the case of healthcare.

A person with an emergency will be treated in a hospital despite of not having a health insurance. The cost will be covered from the community namely the health insurance policy holders. To notice is also that the private profit in having a health insurance is small (will be treated anyway in case of an emergency) in comparison to the social benefit which is high if many people pay their premiums.

In the cyber insurance market this problem also occur. The reason for such behaviour is due to the fact that the most computer systems are interconnected. We see in comparison to the health care case above that also here the private benefit of investing in cyber security is lower than the social. To be more precise, the investment of one individual into a single computer system may probably not be beneficial since connected third party systems can remain insecure which can have an negative security influence on the own computer system. Further the social benefit is high if multiple individuals of a network invest in the security of their IT-infrastructure. Due to this issue, there is only a low incentive to invest in cyber security measures [5].

4.2.7 Moral Hazard

Another general problem of the insurance market is the Moral Hazard problem. The Moral Hazard problem simply says that due to the fact of having an insurance it will effect the risk behaviour of the policy holder. For example, the case of having a full coverage car insurance will lead the driver to be more riskily because the driver knows that in case of an accident (except negligence) the driver will not loose the asset [5].

Like in other insurance markets this problem occurs also in the case of cyber insurance. There are less incentives to invest in cyber security if a cyber insurance product is purchased. A possible measure is to demand a higher deductible or to set some compulsory security measures. But it is not possible to observe all actions and therefore an occurrence of the Moral Hazard phenomena can not be excluded [5].

4.3 Empirical data and current state of research

The following section presents the problem of limited data about cyber risk and cyber insurance and tries to give an insight in existing work in this area. The aim is firstly, to point out the importance of the work of data collection in this field and to present some relevant results. Secondly, to show what type of work that is necessary to establish a more elaborated system of cyber insurance and policy writing.

4.3.1 Cyber resilience

Before presenting the work about risk analysis, it is necessary to show how well prepared the firms at the moment are. Thus, we want to show data about the resilience of the cyber market toward any kind of risk, as it has been pointed out several times that this topic has, for some reason, been neglected in the last years, even though there seems to be a lot of potential.

On one hand, this is important for the cyber insurance companies, as they want to know how secure and aware the market at the moment ist. On the other hand, it is important for the cyber market in general, as there might arise significant financial deficits due to insecurity of investing and loss of money due to attacks and leaks.

In the report of the World Economic Forum 2014, the authors investigate the readiness of firms in terms of cyber risks/cyber resilience and the possible financial consequences [16]. They looked at the cyber resilience of firms of the private and the public sector across different industries by questioning more than 250 industry leaders. They found

that even large companies lack efficient cyber security systems and often can not make efficient decisions about cyber resilience. Most of the questioned firms were rated either "nascent" or "Developing" (cf. Fig. 4.4), where the scale goes from "Nascent" (1) over "Developing"(2) and "Mature"(3) to "Robust"(4). In short: according to the World Economic Forum 2014, the firms score quite badly in terms of cyber resilience and have poorly developed mechanisms to protect against cyber risks.

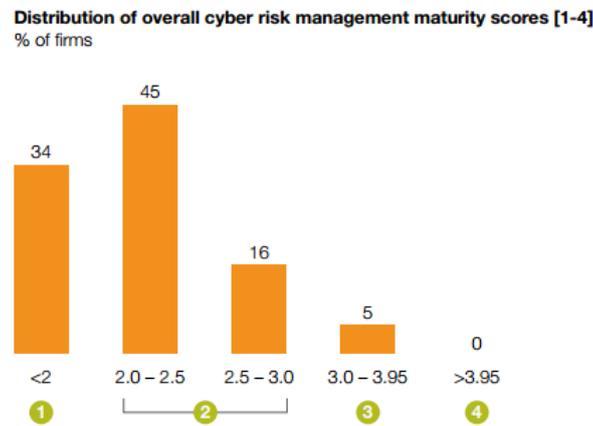


Figure 4.4: Cyber resilience rated according to WEF.

There are different reasons for this, for example the fact that the phenomenon is quite new and the firm leaders have not yet recognise its importance, or the dynamic and rapidly changing nature of the digital market, which makes it extremely difficult to take appropriate actions.

The relevance of the lack of resilience is twofold. Firstly, it is important for the cyber insurance market, as the security of the cyber systems is a key aspect to consider when writing a policy. If a company has only nascent precautions, the cost of the same insurance policy might be higher than for a company that has a mature resilience. Further, it might be an indicator that the awareness of the risk is generally low among the leaders even though the threat is there. Thus, education about the risk, data collection and formalised processes are fundamental in developing a solid cyber resilience.

Second, it is important for the cyber market in general, as there are predicted significant financial losses due to the insecure environment of the cyber market. This could lead to delay of adopting technology, legal restrictions, hesitation of investing and so on. The report from the World Economic Forum 2014 tries to predict the generated value of different technologies in the future based on different present baselines. The results are shown in figure 4.5. The table is directly taken from the WEF 2014 report with the lower part omitted, as there were no different outcomes for these sections predicted. In the first two columns are the different sectors shown and the value they could generate by 2020 when leveraging their full potential. It is divided in private and public sector, where the purple coloured bar represents the public and the orange coloured bar the private sector.

	Est. value created by 2020		Impact of alternative future scenarios		
			Muddling	Backlash	Cyber resilience
Business & technology innovation total	9,630-21,630		(410)-(1,020)	(1,230)-(3,060)	Full value captured
Cloud technology	1,020	2,700 ²	(130)-(470) ⁴	(390)-(1,410) ⁴	-
Internet of things	1,600	2,150 ²	(90)-(210)	(270)-(630)	-
Mobile internet	1,330	1,550 ²	(70)-(150)	(210)-(450)	-
Rapid entry into new markets	170	50 ¹	(10)	(20)-(40)	-
Automation of knowledge work	2,500	720 ²	(80)-(100)	(240)-(310)	-
Social technologies	750	350 ³	(20)-(30)	(70)-(100)	-
E-commerce	270	240 ¹	(10)	(20)-(40)	-
Autonomous & near-autonomous vehicles	1,020 ²	120	(20)	(10)-(70)	-
Next-generation genomics	420	540 ²	(10)	(20)-(40)	-

Figure 4.5: Future financial impacts of cyber resilience.

In the last three columns are the scenario shown that depend on different developments of cyber resilience and cyber threats. The over all value that could be created is estimated between 9.630 and 21.630 Billion USD, if the firms could implement the new technologies on time and without major disruptions. However, if the cyber threat is developing faster than the resilience, due to hesitation and "muddling", a lot of potential value might get lost, because of delayed adoption of technologies and fear of data breaches. In this scenario, 130 to 470 Billion s could not be realised by 2020. In an even darker future scenario, the lack of resilience and growth of threats would lead to sever security gaps and to stricter regulations of the market, which could cause a "backlash" and leave up to 1410 Billion s not realised.

Due to the little awareness and the low resilience scores of the firms, the "muddling" scenario seems the most likely which is why it was considered the baseline. This shows how important the awareness for this topic is. Even though the mentioned numbers are only predictions, it becomes visible how much impact a functioning cyber resilience environment could have. These findings are extremely important for the cyber insurance market, as cyber resilience and cyber insurance are directly connected topics and influence each other greatly. The more secure the environment, the less insurance might be needed, however, insurances are a strong carrier for security and contribute directly to the resilience itself. Thus, when performing risk analysis and providing insurance policy, it is essential to consider the over all market/environment situation first.

4.3.2 Risk Analysis

In this section, we summarize the work concerning risk analysis in the field of cyber insurance. The purpose is to show the essential tools necessary in this field and to give an overview over existing data. These findings are considered crucial for cyber insurance companies, because they constitute the base upon which insurance policies can operate. The first step of analysing risk - or in any kind of scientific work - is to classify the different phenomenon into groups with similar attributes. Thus, before analysing the risk, it is necessary to put it into categories. This is a difficult task, as the cyber market is extremely diversified and rapidly developing and there exist numerous different threats and incidents which are hard to classify. Also, there are many different approaches in classifying cyber risks and many authors come up with different classification system, such

as the NIST or ISO framework. The uniqueness of the incidents in the cyber market and its fast developing character are often considered a significant problem for the establishment of cyber insurances.

Despite the difficulties, there are certain patterns that can be observed to be present in most of the classification systems. For example Cebula and Young classified the cyber risk types as follows [10]:

Category 1: Actions of People Includes the subcategories inadvertent, deliberate and inaction. Under inadvertent fall all the unintentional actions without harmful intent, deliberate is the exact opposite and encompasses all the actions with harmful intentions. Inaction addresses the cases in which the lack of action lead to an incident.

Category2: Systems and technology failure Consists of hardware, software and systems issues. The hardware section addresses the issues of physical structure failures, while "software" includes all the software problems. In the category "systems" are the cases included in which the implemented system did no perform as expected and lead to losses.

Category3: Failed internal processes Includes process design and execution, process controls and supporting processes. These category address the cases in which badly performed process or badly controlled processes lead to losses or in which supporting processes failed to provide the necessary support.

Category4: external events Encompasses catastrophes, legal issues, business issues and service dependencies.

Of course it is not always clear to what category an incident should belong. But this category-system can be seen as a general guideline to classify cyber risk incidents. Based on this system it is possible to analyse cyber risk properly. Biener et all. for example used this classification for their study [8] on cyber risk analysis. They took data from the SAS OpRisk Global Data, which is a database that consists of publicly reported operational losses with 22,075 incidents from 1971 to 2009.

They only included data considered as insurable based on Berliner's criteria. Berliner et all. distinguished between insurable risk and non-insurable risk [7]. Thus, Biener et all. based their process of distinction on Berliner's approach, considering 9 criteria: randomness of loss occurrence, maximum possible loss, average loss per event, loss exposure, information asymmetry, insurance premium, cover limits, public policy, legal restrictions[8].

Based on these criteria, they excluded all the cases that were not insurable and only considered those that were. A descriptive table for these cases is provided in figure 4.6.

Christian Biener *et al.*
Insurability of Cyber Risk

9

Table 4 Losses per risk type (in million US\$)

Category	N	Mean	Std. dev.	Min.	Quantiles			VaR	TVaR	Max.
					25%	50%	75%	(95%)	(95%)	
<i>Panel A: Cyber vs non-cyber risk</i>										
Cyber risk	994	40.53	443.88	0.10	0.56	1.87	7.72	89.56	676.88	13,313
Non-cyber risk	21,081	99.65	1,160.17	0.10	1.88	6.20	25.37	248.97	1,595.27	89,143
<i>Panel B: Cyber risk subcategories</i>										
Actions of people	903	40.69	463.25	0.10	0.55	1.83	6.87	84.36	679.04	13,313
Systems and technical failure	37	29.07	77.33	0.10	1.10	5.03	11.65	168.95	329.04	370
Failed internal processes	41	47.72	205.92	0.14	0.42	2.04	9.05	158.65	743.40	1,311
External events	13	39.40	115.73	0.28	0.56	1.03	13.77	192.88	422.71	422

Figure 4.6: Descriptive analysis of cyber risk incidents from Biener et all.

The table shows in the first row a comparison between cyber and non-cyber incidents and the related losses. It is clear, that the overall losses of cyber-incidents are by far smaller than the non-cyber incidents, as the mean and the maximum loss are significantly higher for non-cyber risks. This might be due to several reasons. One might be that the available data for cyber incidents is less complete than the one for non-cyber incidents and therefore, it might be underestimated. Another reason might be that it is more difficult to insure cyber-incidents and thus, when selecting based on Berliners criteria, a lot of incidents were filtered out because they were classified as "non insurable". In other words, it might be easier to classify non-cyber incidents as insurable, which would lead to a bias in the category of insurable incidents, because the relative amount of non-cyber incidents would be higher and thus, the total amount of loss as well.

In the second Panel of figure 4.6, The authors displayed a differentiation between the above mentioned subcategories. The most expensive category is by far the "actions of people" category. Although the mean of loss is quite similar across all the categories, the max loss is clearly the highest for actions of people. This includes unintended actions, like mistakes or carelessness as well as intended, such as hacker attacks and data thefts. Further, Biener et al. analysed additional aspects of cyber risk, such as the distribution of cyber risk compared to the distribution of non cyber risk and found, that the distribution of non cyber risk has a much heavier tailed shape than the one of cyber risk, which might partially explain the results of Panel B in figure 4.6. Also, they differentiated between different continents and cyber risk categories, which will not be discussed as it is too specific and does not serve the overview character of this work. In conclusion, this section shows that cyber incidents are, on average, as well as in the maximum amount less costly than non cyber incidents, which might be due to different reasons. For example, it might be its nature or, it might be some sort of measurement error, such as lack of data for cyber incidents. Also, the highest losses for cyber incidents were clearly related to actions of people, even though on average no cyber incident category was significantly higher or smaller than others. These type of findings are highly important for insurance companies, as they allow the policy writers to adjust their conditions and reduce information asymmetry to a certain degree. For example, seeing that actions of people is by far the most expensive category it seems necessary to adjust the maximum loss premium for these type of insurance.

4.4 Data Collection

Other than examining existing operational loss databases such as the SAS OpRisk Global Data Base, it is also necessary to directly gather data about cyber incidents. It was mentioned that, because of the newness of the topic and the little societal attention, the available data concerning cyber loss and insurance claims concerning cyber incidents is rather scarce. Thus, it is essential to gather and investigate additional data. The Net Diligence Cyber Claim Study is, among others, a project that investigates and collects data in this field annually. In the following section, we are going to present their approaches and main findings in order to give some insight in this type of work [18].

4.4.1 NetDiligence Main Findings

The study gathered data of incidence that occurred between 2013 and 2017 and for which all a claim has been paid. 500 alone in 2018. They categorised the data based on different attributes, such as type of data, sector, revenue size and so on. Thus, the difference to the SAS OptRisk data is firstly, that it was gathered by the authors themselves rather than relying on already existing data and second, that for each incident, there has actually been

paid a claim. This means, these cases do not have to be checked if they can be insured or not, as they actually have been insured. These property makes the data valuable as it is "real life" data taken from the cyber insurance market itself and did not undergo any processes before being analysed.

In their analysis, the authors categorise the data based on different properties. One is similar to the categorisation above, which they call "cause of loss". It consist of hacker attacks, ransomware, malware, stolen device, phishing, legal action, staff mistake, rogue employee, business email compromise, third party, paper records, programming error and "all others". Their key finding was that hacker attacks, ransomware, malware and lost/stolen laptops were the biggest causes of loss. This is congruent with what Biener et al. in their analysis found.

Further, the NetDiligence study views the data from other angels, such as the over all costs or the affected sectors. Their main findings were that the total breach cost was on average 603'900 USD, the average crisis service cost 307'000 USD and large company breach on average 8.8 million USD. Concerning affected sectors, their main findings were that, on average, the retail sector had the highest losses with 1.2 millions USD, second was the financial sector with 854 000, followed by the healthcare sector with 555'000 USD on average. The smallest average losses were found in the professional service sector. An overview over the claims by sector can be found in figure 4.7.

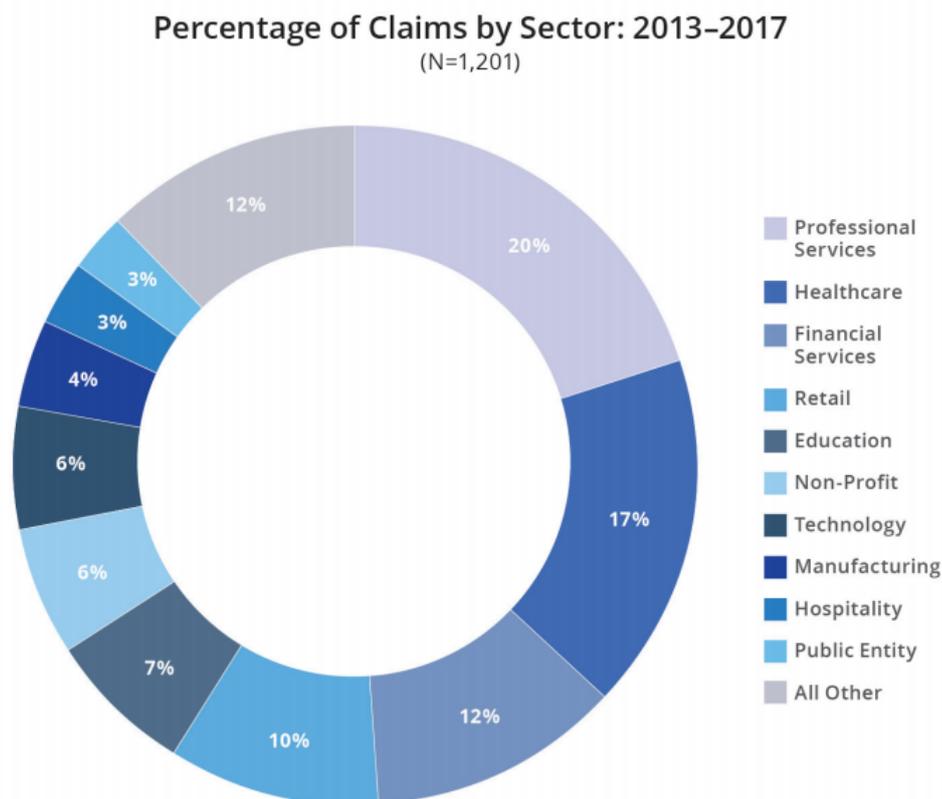


Figure 4.7: Figure taken from the NetDiligence Cyber Claim study 2018

The figure shows the percentage of claims per business sector from 2013 to 2017. For example, the highest percentage of claims has the professional service sector with 20 percent. This type of classification is an example of how to classify risk and helps to estimate the probabilities and costs for insurance companies.

In conclusion, in this chapter it has been as shown that the cyber market is mostly not aware of the risks and the possible financial consequences in the future of a low cyber resilience. For cyber insurance companies and the general market, it is important to

collect data concerning cyber risks and to analyse it from different points of view in order to initiate appropriate actions.

4.5 The cyber insurance market

This section gives an overview over the existing market and what such insurance can cover in different scenarios. With real world examples we want to outline what the current market looks like and then take a look at future expectations of the cyber insurance market.

According to KPMG's 2018 Global CEO Outlook many CEO's believe believe that a cyber attack is only a matter of time and are inevitable with 68% saying it is just a matter of time and only 51% believe they are well prepared for a cyber attack [15]. This pessimistic view of many CEO's can help funnel more money into cyber insurance markets, as they do not expect their company on their own to handle the problem of cyber threats.

There is a general optimistic view of the cyber insurance market as a whole as the Allianz reports that they expect cyber insurance premiums to grow globally from just 2 billion USD a year to over 20 billion USD which would mark an annual growth rate of above 20% [3].

Also for personal cyber insurance, the market is expected to have a high level of growth, however the personal sector is lagging behind the business insurance sector. Being estimated to be under 500 million USD of written premiums at the moment, the personal cyber insurance sector is expected to reach between 1.6 billion USD and 3.1 USD billion by 2025 [26].

Personal cyber insurance market size estimation

Gross written premium in million USD

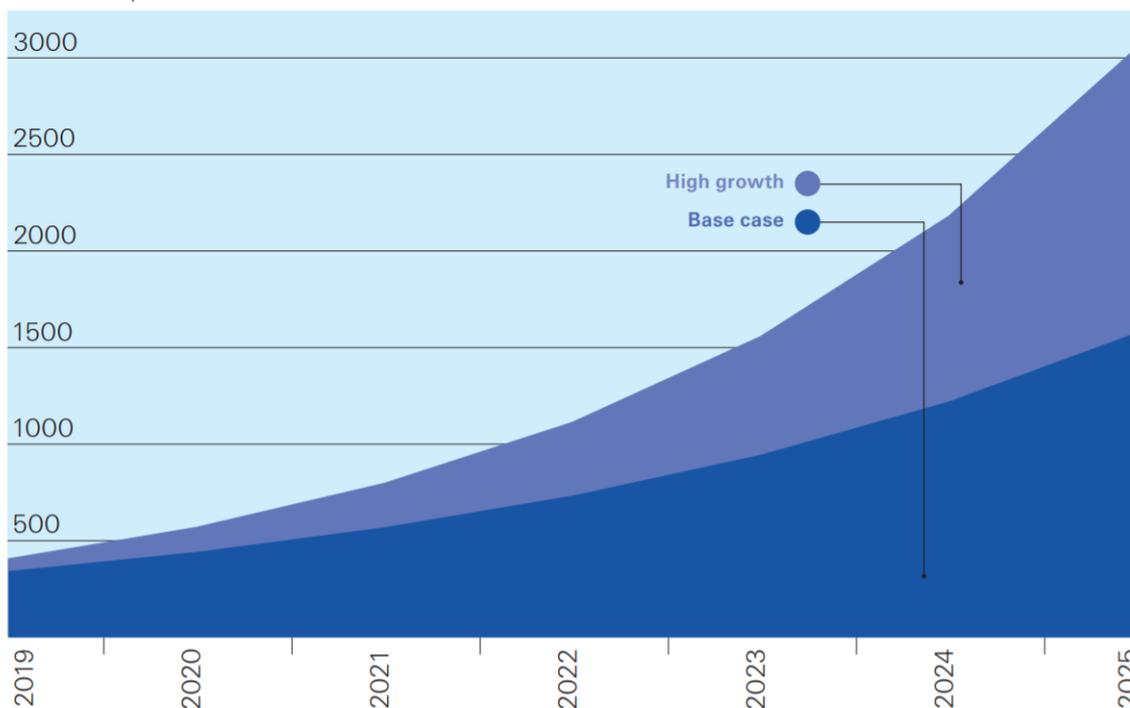


Figure 4.8: Expected growth of individual cyber market [26]

4.5.1 Business Insurance

The fact that cyber insurance for businesses often requires custom solutions for a company has to be considered when looking at example contracts of insurance companies. Often a standard solution does not exist.

There are two types of cyber insurance contracts offered by insurance companies. On one hand, there is first-party insurance, which means the coverage of direct losses such as theft of money or business interruption. On the other hand there is third-party insurance. This on the contrary means the coverage of indirect losses such as litigation or settlements [17].

4.5.1.1 Typical coverage

In this subsection, typical coverage will be explained, split in first party and third party coverage. Typical coverage in this context means the coverage which usually is offered, taken from a paper which analysed the content of available contracts [23].

- First party coverage
 - Data compromise response (coverage for expenses from personal data compromise)
 - Identity recovery (coverage for recovery of identity after identity theft)
 - Computer attack (expenses arising from attack on the computer system)
 - Cyber extortion (coverage for cost of investigation and any amount paid by the insured)
- Third party coverage
 - Data compromise liability (coverage for defense and settlement costs in legal affairs due to data compromise)
 - Network security liability (coverage for defense and settlement costs due to third party business information breach, unintended forwarding of malware etc.)
 - Electronic media liability (coverage for defense and settlement costs due to electronic communications which resulted in defamation etc.)

4.5.1.2 Buyers perspective

In this subsection we take a look at who the buyers of cyber insurance actually are, and what reasons they have for purchasing insurance.

According to a survey with insurance brokers and underwriters conducted by PartnerRe and Advisen, most companies buying cyber insurance contracts are small or medium-sized businesses [21]. One reason for this could be that bigger companies already have existing insurance contracts and smaller and medium-sized businesses are just starting to become aware of the exposure and therefore are buying insurance more.

As one can see in figure 4.9, most new buyers come from the professional services segment, closely followed by manufacturing / industrials [20]. In 2018, healthcare was listed as the top segment of new buyers [21], which is interesting since healthcare was often affected by cyber crime [13].

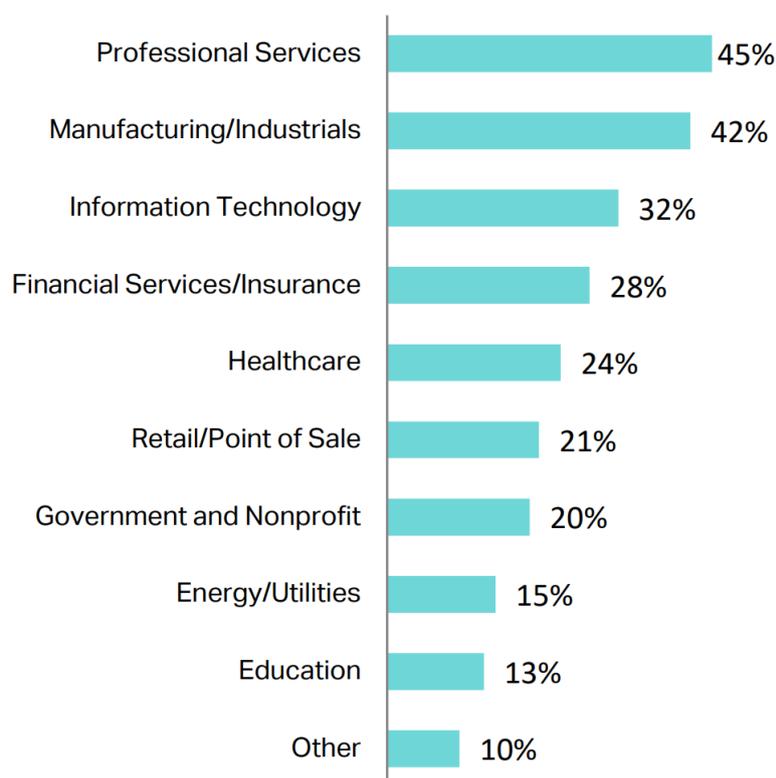


Figure 4.9: Market segments of insurance buyers [20]

The top driver of motivation to buy cyber insurance for a company is news of cyber-related losses of other companies. Similarly, one of the main obstacles for a broker to sell insurance is that their customers do not understand the exposure to cyber-related risk [20]. Through the news of other companies having losses through cyber related risks, companies are made aware of those risks and opt for cyber insurance.

Another possibly big driver for insurance products are regulations and laws. In 2003, a law was introduced in California, which requires a state agency, person or business to disclose any data breach. As a result of this regulation, the cyber insurance market grew significantly [17].

4.5.1.3 Example contract

In this subsection we want to give an example of an existing contract by Zurich. The key coverages of Zurich's cyber insurance are as follows [30]:

- Security and Privacy Liability
- Regulatory proceedings defense costs
- Civil fines and penalties associated with Payment Card Industry (PCI) and General Data Protection Regulation (GDPR)
- Media liability coverage
- Privacy breach costs (including: Forensic investigations, legal and public relations, credit and identity monitoring, identity restoration, call center costs)
- Business income loss
- Digital asset replacement

- Cyber extortion
- System failure
- Reputational damage
- Social engineering funds transfer
- Claims avoidance coverage

Coverage limits are available up to 25 million USD. A company will be initially assessed by a risk engineer to identify existing risks. There is also an option to get fee-based ongoing assistance for cyber security based on three pillars: people, process and technology. This includes education of management, user awareness training, security strategy, incident response and disaster recovery and recommendations for technology solutions. Also, there is an option to subscribe to a 24/7 network monitoring service.

As one can see, the typical coverages (cf. section 4.5.1.1) are included in this example. However, there are many more services included or available that go beyond just insurance, e.g. security consulting.

4.5.2 Personal Insurance

Personal insurance is quite different compared to cyber insurance for businesses. The needs are obviously different. For instance, business interruption being the most requested coverage for businesses [20], this is by far not as important for individuals.

While there exists a wide variety of offers by insurances, personal cyber insurance is not a mass market product yet. Most products offer first party coverage and only few offer some form of third party coverage [26].

According to a global survey by Swiss Re conducted with almost 900 participants, 56% of people asked were willing to buy personal cyber insurance. Interestingly, women are much more willing to buy insurance than men. Almost two thirds of women stated they would go for insurance, however, almost half of the men said they would not. The main reason to opt for an insurance contract is the fear of being affected by some cyber risk. Therefore, as the risk awareness rises, it is expected that the demand for individual cyber insurance products will rise significantly [26].

4.5.2.1 Typical coverage

In this subsection, we show what typically is included in most personal cyber insurance contracts [26].

- Financial fraud:

Financial loss when your online payment services are misused is covered. Also costs for reacting to an incident like blocking the credit card is usually covered. A possible scenario would be when you enter your bank credentials in a phishing attack and you lose money or it costs you a certain amount of money to fix it again.

- Online shopping

This includes protection for goods which are bought online but are never delivered. For example, a lot of online shoppers are buying items from China through a third party web shop and those items might be fake, don't even exist or never get delivered in the first place. Those incidents are covered.

- Identity theft
Costs to rectify your records with banks and authorities and also costs of an unpaid leave to deal with this are usually covered. Your credit card details being used by someone else in your name would be a possible scenario.
- Data restoration
The costs of restoring your compromised data is usually covered, for example if an infected memory stick causes loss of data.
- Cyberbullying
Sometimes, legal advice, psychological consultation, relocation cost and parent's liability in relation to cyberbullying is included in such a contract.
- Cyber extortion
Ransom payment is expected to be covered. An example is if some ransomware infects your computer and encrypts the data. To resolve this, a payment is demanded. The policyholder might decide in such a scenario that least damage is dealt if the payment is done or it is solved in another way. Those kind of costs are to be covered.
- Cyber liability
Liability for different claims are usually covered. An example is that your network is used as part of an attack.

As described in the list above, most of the times only first party losses are covered by an insurance contract.

4.5.2.2 Example contract

The following list describes what the coverage of an existing insurance offer by Allianz looks like [4]:

- Removal from reputation-damaging or defamatory content by an IT specialist
- Psychological support
- Data recovery, removal of malware
- Online legal protection (for computer and internet offences, personality rights and copyright infringement)
- Online purchase protection (in case of non-delivery, incorrect or damaged delivery of online purchases)
- Online account protection (financial losses due to identity theft)

Comparing this example contract with the typical coverage in section 4.5.2.1, one can see that the example aligns with the typical coverage.

4.6 Solutions

The first problem mentioned is correlated risk. The technology sector heavily favors interconnected systems and it is not in the power or the desire of the insurer to change that. In order to tackle this problem, the contract for an insurance policy would have to be written incredibly carefully and also has to be heavily limiting in order to prevent liability in case of some security breaches of a connected, yet independent system for which it is not providing coverage.

The lack of reinsurance companies which limit the risk of the insurance companies could be partially solved by the solution mentioned above. It might encourage reinsurance companies slightly however there still is the problem of world wide viruses which is almost unheard of when it comes to more traditional insurances. A possibility would be that policies are implemented in which at a certain impact the companies affected would have to be liable themselves however that defies the purpose of cyber insurance companies as a whole and would negatively impact the demand of such insurances.

Artificial intelligence has become popular and might pose as a solution for the information asymmetry existing between the insurer and the policyholder. Current security measures could be automatically detected and with the help of AI assess the risk more accurately. The question still remains however if an insurance is appropriate when an AI is present that can detect a lack of security measures, because those problems might as well be fixed when the AI was able to detect them.

Human errors are a common source of vulnerabilities in systems and therefore influence cyber insurances. Decreasing the amount of human errors is not the goal of the insurance company, however a general skill level regarding the employees average knowledge in cyber security threats might be assigned and play a role in the price of the premium a company has to pay. Another solution is introducing some form of certificate for having absolved a course about cyber threats and if most employees are a holder of such a certificate the price of a premium would decrease.

A possible solution to solving the free-rider problem would also be stricter and better defined policies. Through precisely defined policies and stakeholders eliminating some free-riders is possible. However this problem can also never be completely eliminated and will have to be dealt with indefinitely.

The problem of moral hazard exists in cyber insurances however also in most other insurances and we can see how other insurances have tackled such a problem. Having a low information asymmetrie also helps to tackle this problem, because the premium could be higher if such behavior is discovered by an AI for example as mentioned above.

4.7 Evaluation and Discussions

The market for cyber insurances did not grow as fast as the digitization process. Since the threats in cyber space are more serious than in the past the use of cyber insurance products might be interesting. However, the overall limitation is that the field of cyber insurance is quite young and therefore not much research and data available. Nevertheless, there are some observable and evaluations which can be done.

4.7.1 Overpricing

Of course due to the imperfect cyber insurance market the price will be high. The imperfection of the market can be explained by all the problems mentioned above. Particular the information asymmetry can lead to overpricing in the case if the policy holder does not know about possible discounts on his contract. The weak growth of the market imply

also a weak growth of experience in the field of cyber insurance products. The insurance companies work with prediction models for their risk management. Since there are several limitations in this market field the predicting models are not very accurate. The insurers will apply a conservative price to avoid their own financial risks. The high prices are no incentives to invest in cyber insurances for companies. In short the problems are interdependent and the market is in a vicious circle.

4.7.2 Low policy limits

We also see that the insurer sets low policy limits for their customers. The interdependent and correlated risks which the cyber insurance market has to face do not allow to offer high policy limits in the contracts. Such low policy limits makes the cyber insurance products very unattractive for larger companies which can face high losses due to cyber security breaches.

4.7.3 Further research

We conclude that there is still a lack of research in order to make cyber insurances more mainstream and the following ways points may help the further development of cyber insurances:

- Explore and research multiple different ways of how to decrease the information asymmetry whether manually or automatically.
- Research what kind of laws are appropriate in the country in question to guarantee easier risk management for example how can pinpoint the stakeholders and the errors more accurately.
- Find out how exactly more reinsurance companies can be attracted and whether the help of an administration is needed or not.

4.8 Summary and Conclusion

Despite the clear problems currently existing in the cyber insurance market there does exist an almost universal positive outlook between insurance companies that this market segment is expected to grow massively in the coming years as we have seen in the past years in companies.

This market segment is unfortunately somewhat nontransparent because for many companies it is uncertain whether they do purchase cyber insurances and in what form making it hard to state clear numbers but according to a survey conducted by Statista the majority of companies supposedly already do purchase some form of cyber insurance. When cyber insurances become more mainstream and the general public has a more positive view of cyber insurances companies might also disclose more openly if they have some form of cyber insurance which will make it easier for further and more precise research in the future.

When it comes to individuals, purchasing cyber insurance is still fairly unheard of. There are clear barriers in the general public which are to be crossed like the lack of knowledge about those kind of risks when it comes to computer usage in general. However in the far future maybe we will see mandatory cyber insurances which every individual has to purchase like the current requirement of having some form of health insurance in Switzerland?

Cyber insurance is and will be an interesting topic for many insurance companies as the market is still young and therefore it still has a lot of room for growth and we are certain that this topic will become more and more relevant in the future.

Bibliography

- [1] M. Akbanov, V. G. Vassilakis, M. D. Logothetis: *WannaCry Ransomware: Analysis of Infection, Persistence, Recovery Prevention and Propagation Mechanisms.*; Journal of Telecommunications and Information Technology, 1(1), 2019, pp. 113-124.
- [2] G. A. Akerlof: *The Market for "Lemons": Quality Uncertainty and the Market Mechanism.*; The Quarterly Journal of Economics, 84, 3, 1970, pp. 488-500, www.jstor.org/stable/1879431.
- [3] Allianz: *Businesses must prepare for new generation of cyber risks*; <https://www.agcs.allianz.com/news-and-insights/news/cyber-risk-guide.html>, November, 2019.
- [4] Allianz: *Customer Information and General Terms and Conditions of Insurance Secure Cyber*; https://www.allianz-assistance.ch/v_1546961964660/media/downloads/TC_2018/TC_Secure_Cyber.pdf, November, 2019
- [5] W. S. Baer, A. Parkinson: *Cyberinsurance in IT Security Management.*; IEEE Security & Privacy 5.3, 2007, pp. 50-56.
- [6] T. Bandyopadhyay, V. S. Mookerjee, R. C. Rao: *Why IT Managers Don't go for Cyber-insurance Products.*; Communications of the ACM-Scratch Programming for All 52.11, 2009, pp. 68-73.
- [7] B. Berliner: *Limits of Insurability of Risks*; Englewood Cliffs, NJ: Prentice-Hall,1982.
- [8] C. Biener, M. Eling, J. H. Wirfs: *Insurability of cyber risk: An empirical analysis.*; The Geneva Papers on Risk and Insurance-Issues and Practice 40.1, 2015, pp. 131-158.
- [9] BusinessDictionary: *information asymmetry*; <http://www.businessdictionary.com/definition/information-asymmetry.html>, November, 2019.
- [10] J.J. Cebula, L.R. Young: *A Taxonomy of Operational Cyber Security Risk*; Technical Note CMU/SEI-2010-TN-028, Software Engineering Institute, Carnegie Mellon University.
- [11] CRO Forum: *Cyber Resilience – The cyber risk challenge and the role of insurance*; CRO Forum, 2014, <https://www.thecroforum.org/2014/12/19/cyber-resilience-cyber-risk-challenge-role-insurance/>.
- [12] L. Gordon, M. Loeb, W. Lucyshyn, R. Richardson: *2005 CSI/FBI Computer Crime and Security Survey.*; Computer Security Institute, 2005, <http://www.detomaso.it/materiale/FBI2005.pdf>.
- [13] M. Greisiger: *Cyber Liability & Data Breach Insurance Claims*; 2013, <https://netdiligence.com/wp-content/uploads/2016/05/CyberClaimsStudy-2013.pdf>, November, 2019.

- [14] C. Kolbe, S. Zaugg: *Nach Swisscom-Panne – Cloud gelöscht, wie weiter?*; <https://www.blick.ch/news/wirtschaft/nach-swisscom-panne-cloud-geloescht-wie-weiter-beteuerungen-der-cloud-anbieter-nicht-blind-vertrauen-id15418409.html>, November, 2019.
- [15] Kpmg: *Optimistisch, vorgewarnt*; <https://home.kpmg/de/de/home/themen/2018/05/ceo-outlook-2018.html>, November, 2019.
- [16] A. Marcus, D. O’Halloran, E. Kvochko, R. Vora: *Risk and Responsibility in a Hyperconnected World*; World Economic Forum, 2014.
- [17] A. Marotta, F. Martinelli, S. Nanni, A. Orlando, A. Yautsiukhin: *Cyber-insurance survey.*; Computer Science Review, 24, 2017, pp. 35-61.
- [18] NetDiligence: *Net Diligence Cyber Claim Study*, v. 1.0, 2018.
- [19] T. Nguyen: *Vorteilhaftigkeitsanalyse von Katastrophenbonds versus traditionelle Rückversicherung.*; Fakultät für Mathematik und Wirtschaftswissenschaften, Ulm Universität, 2019.
- [20] PartnerRe & Advisen: *Cyber Insurance - The Market’s View.*; 2019, https://partnerre.com/wp-content/uploads/2019/10/Cyber_Insurance_The_Markets_View_2019-1.pdf, November, 2019.
- [21] PartnerRe & Advisen: *2018 Survey of Cyber Insurance Market Trends*; 2018, <https://partnerre.com/wp-content/uploads/2018/10/2018-Survey-of-Cyber-Insurance-Market-Trends.pdf>, November, 2019.
- [22] R. Richardson: *2008 CSI Computer Crime & Security Survey.*; CSI, 2008, <http://www.kwell.net/doc/FBI2008.pdf>.
- [23] S. Romanosky, L. Ablon, A. Kuehn, T. Jones: *Content Analysis of Cyber Insurance Policies: How Do Carriers Write Policies and Price Cyber Risk?*; Research Paper, Journal of Cybersecurity, 2019, pp. 1-19, <https://academic.oup.com/cybersecurity/article/5/1/tyz002/5366419>
- [24] Schweizerische Eidgenossenschaft: *Melde- und Analysestelle Informationssicherung MELANI*; <https://www.melani.admin.ch/melani/de/home.html>, November, 2019.
- [25] Spiegel: *Onel de Guzman ist frei!*; <https://www.spiegel.de/netzwelt/web/i-love-you-prozess-onel-de-guzman-ist-frei-a-89973.html>, November, 2019.
- [26] Swiss Re: *Personal cyber insurance: Protecting our digital lives*; White Paper, 2019, https://www.swissre.com/dam/jcr:68e4d8fb-509c-4182-a219-c803f7d23af1/ZRH-18-00632-P1_Personal_cyber_insurance_Publication_WEB.pdf.
- [27] the npm blog: *kik, left-pad, and npm*; <https://blog.npmjs.org/post/141577284765/kik-left-pad-and-npm>, November, 2019.
- [28] M. Ward: *A decade on from the ILOVEYOU bug*; <https://www.bbc.com/news/10095957>, November, 2019.
- [29] Wirtschaftslexikon24: *Trittbrettfahrer*; <http://www.wirtschaftslexikon24.com/d/trittbrettfahrer/trittbrettfahrer.htm>, November, 2019.
- [30] Zurich: *Cyber security and privacy*; <https://www.zurichna.com/en/industries/technology/secpriv>, November, 2019.

Chapter 5

An Economic Analysis of Cloud Storage Providers and Private Cloud/NAS Systems

Clara-Maria Barth

In this report, public and private cloud storage solutions are presented and compared according to three examples each. Both options are evaluated concerning their price and their security. The report concludes, stating that public cloud storage is cheaper but lacks security, while private cloud storage is more expensive yet more secure. Two alternative options, which partially mitigate the price while keeping sensitive data secure, namely hybrid cloud storage and community cloud storage, are presented. However, they introduce different drawbacks, such as the increased complexity of the system as well as the rapid scaling and moving of the data. The report concludes that the choice of cloud storage type depends on the sensibility of data that is being stored.

Contents

5.1	Introduction	125
5.2	Cloud Storage	126
5.2.1	Architecture of Cloud Storage	126
5.2.2	Service Models of Cloud Computing	127
5.2.3	Cloud Storage Types	128
5.3	General Security Techniques	129
5.4	Public Cloud Storage Providers	131
5.4.1	Pricing	131
5.4.2	Security	131
5.4.3	Encryption	132
5.5	Private Cloud Storage Solutions	135
5.5.1	Network Attached Storage (NAS)	135
5.5.2	Web hosting providers	136
5.5.3	Nextcloud	136
5.5.4	Seafile	137
5.5.5	Pydio Cells	138
5.6	Stress Test Performance	139
5.7	Discussion	140
5.7.1	Pricing Differences	140
5.7.2	Privacy Differences	140
5.7.3	Alternative Solutions	141
5.8	Conclusion	143
5.9	Limitations and Future Work	144

5.1 Introduction

Cloud computing is a well-studied concept that has many applications to support the processing of volumetric data [1]. This movement to cloud services was adopted by cloud storage providers, striving to monetize storage space [2]. Cloud storage allows the user to back up their data and rapid scaling, ubiquitous access and paying only for the resources needed which makes it attractive to businesses there are still some drawbacks concerning security and privacy [3]. To address this drawback, private cloud solutions were developed, which provide more control over the user's data and its security but requires the acquisition of more hardware and specialized personnel. Acknowledging these benefits and drawbacks, it is crucial for a company to choose the option from which they benefit the most.

The goal of this seminar report is to survey the cloud storage options, public and private cloud storage, and alternative solutions. The advantages and drawbacks are presented and discussed along with three concrete examples for each option. The focus of the report is on the pricing of the hardware and/or software, and security concerning the location and encryption of the data.

The remainder of this report is structured as follows. Section 5.2 presents an overview of cloud storage and cloud computing models. Then, Section 5.3 discusses general security techniques regarding cloud computing. Further, Section 5.4 presents public cloud storage providers. Section 5.5 describes private solutions, being complemented with a performance evaluation in Section 5.6. Section 5.7 compares and discusses these solutions in terms of pricing, and privacy. Section 10.9 summarizes the report and presents its results. Finally, Section 5.9 Introduces limitations of the report and presents future work.

5.2 Cloud Storage

This section introduces cloud storage, its architecture, service models, and two cloud storage types. Further, solutions available on the market today will be presented and analysed according to their pricing and security.

Cloud data storage can be defined as a technology that enables the user to access their stored data from any device using the Internet [4, 5]. Therefore, the data from the user is not stored locally but remotely on a server provided by their company or a cloud storage provider [4, 5].

5.2.1 Architecture of Cloud Storage

Cloud Storage generally consists of a front-end that accesses the back-end storage via an Application Programming Interface (API) [6]. For example via the method Representational State Transfer (REST) [6]. An important feature of REST is that it does not require a session and since all the information for retrieval is contained within the REST call there is no issue concerning latency [6]. Using a REST API has the advantage of being language-neutral it, therefore, describes how a storage cloud is accessed in a general way [6]. The layer behind the front-end is the storage logic which manages the data-placement as well as other features, such as data reduction. The back-end storage manages the physical data storage using an internal protocol or a traditional back-end [5]. The storage Architecture mentioned is depicted in Figure 5.1.

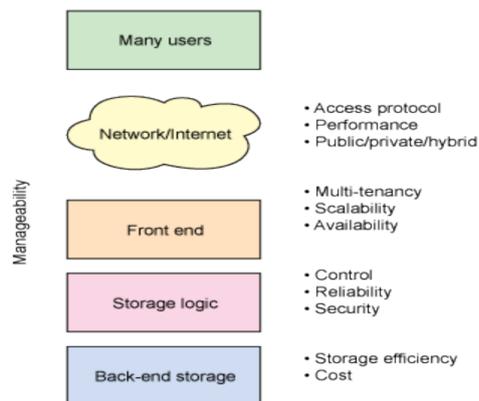


Figure 5.1: Storage Architecture illustration[7]

The characteristics depicted in Figure 5.1, are not assigned to the layers but are there to help understand specific features that are addressed within that layer [7]. These characteristics are presented in Table 5.1.

Table 5.1: Cloud Storage Characteristics [7]

Characteristic	Description
Manageability	The ability to manage a system with minimal resource
Access method	Protocol through which cloud storage is exposed
Multi-tenancy	support for multiple users (or tenants)
Scalability	Ability to scale to meet higher demands or load gracefully
Data availability	Measure of a system's uptime
Control	Ability to control a system, in particular, to configure for cost, performance, or other characteristics.
Storage efficiency	Measure of how efficiently the raw storage is used
Cost	Measure of the cost of the storage (commonly in dollars per gigabyte)

5.2.2 Service Models of Cloud Computing

Cloud computing service models can be divided into three main categories, *(i)* Infrastructure-as-a-Service (IaaS), *(ii)* Platform-as-a-Service (PaaS), and *(iii)* Software-as-a-Service (SaaS). These models are described in the next items.

- i **Infrastructure-as-a-Service (IaaS)**: IaaS provides computing resources to a consumer which can include servers, storage, and security [4]. Therefore, the user is able to combine the options of *e.g.*, the hardware they prefer, without the hassle of managing the cloud infrastructure [4].
- ii **Platform-as-a-Service (PaaS)**: PaaS enables the consumers to write applications, it further has a set infrastructure which means less control over the hardware as in IaaS, as well as the environment needed to deploy an application [4].
- iii **Software-as-a-Service (SaaS)**: SaaS refers to consumers using the applications of the SaaS Providers in the cloud, which means less control but also less managing of the application itself [4]. The software and scale of the resources used from the SaaS provider are flexibly adjustable to the current need of the consumer and therefore the cost of using the SaaS is variable [4].

5.2.3 Cloud Storage Types

This section introduces two cloud storage types; public cloud storage and private cloud storage. The general differences are highlighted, and some examples will follow in the next sections. In Figure 5.2 a graphical representation of the private and public cloud storage is illustrated.

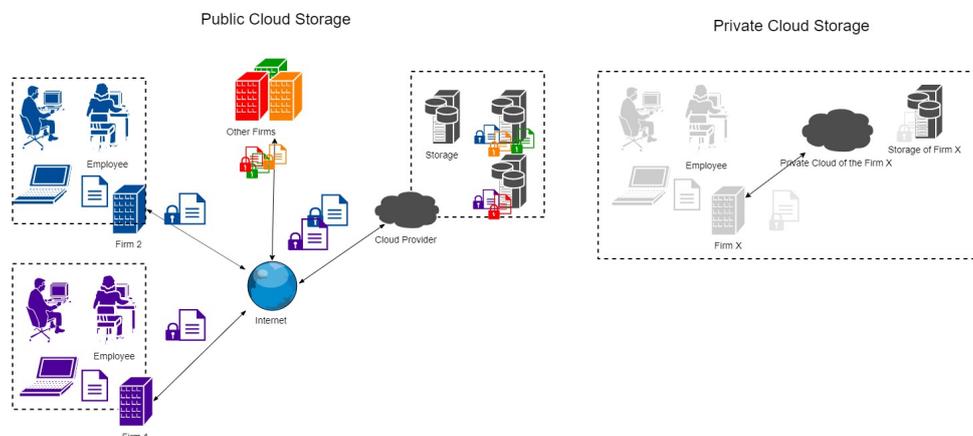


Figure 5.2: Illustration of public and private cloud storage [8]

5.2.3.1 Public Cloud Storage

In public cloud storage the user can access the resources provided using the internet, the storage of their data is therefore handled by the provider of the public cloud storage [4]. The stated advantages of public cloud storages are:

- Data availability and continuous uptime,
- 24/7 technical expertise,
- on-demand scalability,
- easy and inexpensive set-up, and
- no wasted resources.

The technical expertise and on-demand scalability, as well as the not wasted resources, can be especially helpful for small companies that would not have access to such an extensive set of resources otherwise [4]. However, drawbacks exist in such a cloud storage type. In [4] the following are listed:

- Data security concerning, *e.g.* the storage location of the data as well as, who would be able to access it.
- Privacy and reliability.

Security is a major issue when looking at public cloud storage [4]. Of course, there are also opportunities to make use of that cloud with *e.g.*, security services that can be purchased and used by the user of the cloud storage or service [4]. The drawback concerning this solution is that the economies of scale, that are a benefit of the public cloud storage, could be affected when looking to store sensitive data which might lead to a higher price [4].

5.2.3.2 Private Cloud Storage

Private Cloud Storage means that the Storage Infrastructure is operated, maintained and accessed by one company [4]. This can be the consumers within that company (on-premise) or it can be provided by a third party to the consumer company (off-premise) [4]. Therefore, private cloud storage can be seen as a subset of public cloud storage since it is stored within a cloud but only accessible by one company [4].

The general advantages and drawbacks of a private cloud solution are the following. In [4] the drawbacks stated are the cost of acquisition and maintenance of the resources needed. Further, the following advantages in comparison to public cloud storage are mentioned in [4]:

- More security,
- cost-efficient, and
- control over infrastructure and resources.

5.3 General Security Techniques

In this section, general security techniques that can be used to protect the data that is stored in the cloud, are presented. To secure information means to protect it from unauthorized activities, which does not only include access but also modifications deletions [5]. According to [5], the following securities should be provided:

- **Confidentiality:** Confidentiality is about preventing unauthorized access to the data stored in the cloud. Two solutions are cryptography and isolation.
- **Data Integrity:** Data Integrity is the prevention of deletion, modification of data in the cloud storage. This can also happen within an organisation and therefore that kind of data rights are well managed.
- **Availability:** Availability ensures the access of the owner at any time. This can be supported by data redundancy and hardening.

To provide such security aspects, [5] proposes the following solutions:

- **Encryption:** To encrypt something means, turning the data into a format that is not understood by unauthorized people [9]. Traditionally, providers employ, such as asymmetric cryptography which uses two keys, a private and public one, which are mathematically related. Another traditional approach is to use the symmetric key algorithm which uses the same key for encryption and description. Finally, there is the hash function, which produces static length data from the original data. Advanced Cryptography techniques are searchable, homomorphic, identity-based and attribute-based encryption.
- **Identity access management:** Determines, who has access to which information and is authorized to perform which actions. This can be managed by the cloud service provider, or an identity provider, that can be managed in the same cloud, in a different cloud, on-premise. Concerning access management, the solutions are of the form; attribute-based access control or role-based access control.
- **Data Protection:** Ensures that the data is undamaged without having a copy of the data. There are two approaches Provable Data Possession (PDP) and Proof of Readability (PoR). For more information on these two concepts, one can refer to [10].

- Cloud Storage Availability:** Cloud Storage Availability significantly affects the price and storage available in a private cloud set-up discussed in later sections. The data that is stored in the cloud uses a redundant array of independent disks called RAID, which can have different architectures also called levels [5]. To increase performance RAID 0 shown in Figure 5.3, can be used which splits the data over two disks, therefore the search/retrieval time is n times faster considering n disk [5]. RAID 1 shown in Figure 5.4, mirrors the data from one disk to the other, therefore, requiring at least 2 disks, this is useful to prevent data loss if one hard drive fails [5]. A combination of performance and redundancy is RAID 10 visible in Figure: 5.5 which requires at least 4 disks and splits the data over two disks increasing the retrieval performance, while the other two discs are used for redundancy [5].

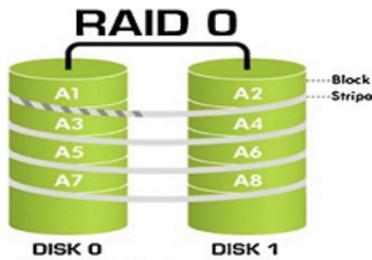


Figure 5.3: RAID 0

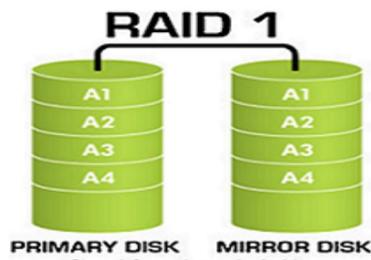


Figure 5.4: RAID 1

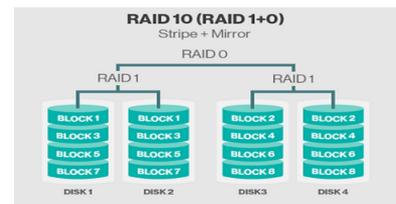


Figure 5.5: RAID 10

5.4 Public Cloud Storage Providers

In this section, relevant public cloud storage providers will be presented, as well as their set-up and differences. Therefore, Dropbox Business, Egnyte, and Box will be compared according to price and security. It is important to note that the plans these providers were evaluated according to their business or enterprise plans and not personal plans.

Dropbox claims to have created the world's first smart workspace, trying to connect all the content and tools that the user needs to use for work [11]. They further state that their system brings people that need to work together, together while using machine intelligence suggesting contents for *e.g.*, a meeting [11]. Egnyte was founded in 2007 and is an enterprise file system built for businesses and claims to deliver secure content [12]. Finally, Box introduced in 2005 is a cloud content management system, that focuses on collaboration any-time and anywhere containing today 95'000 companies [13]. According to [14] Egnyte, Box and Dropbox Business belong to the best enterprise file sync and share of 2019.

5.4.1 Pricing

All three providers offer deals per user per month with a minimum of 3 users. The pricing policy between the three is based on different factors. While Egnyte provides no information concerning the price of unlimited users and unlimited storage, Dropbox provides this for 15 Euros and Box even at 13.5 Euros per user per month, as can be seen in Figure 5.6. This means that if storage space and the number of users is the biggest concern, Box would be the cheapest option available on the market.

Drop Box Business		
	Standard	Advanced
Price: per user/month	10€	15€
Storage	3TB	unlimited
Maximum of users	unlimited	unlimited

Egnyte			
	Office	Business	Enterprise
Price: per user/month	8\$	20\$??
Storage	5TB	10TB	unlimited
Maximum of users	25 employees	100 employees	unlimited

Box				
	Starter	Business	Business Plus	Enterprise
Price: per user/month	4.5€	13.5€	22.5€	??
Storage	100GB	unlimited	unlimited	unlimited
Maximum of users	10	unlimited	unlimited	unlimited

Figure 5.6: Pricing of public cloud solutions [15, 16, 17]

However, the advanced features that are the reason for the higher price tag in the different public cloud provider subscription plans, are important concerning security and functionality. Therefore, it is dependent on the business, which option with which features is best suited for them. Important features concerning security are presented in the following sections.

5.4.2 Security

Concerning the security of public cloud two aspects were evaluated. First the location of where the user's data will be stored and secondly the encryption of the data stored in the public cloud. In Figure 5.7, an overview of the security of the three public cloud service providers that were analysed.

Box		
	Default Option	Add ins
Geolocation	??	Box Zones Free choice of data location
Encryption	AES 256-bit Encryption	Box KeySave Own encryption keys stored by AWS.

Egnyte		
	Default	Description based additions
Geolocation	All European data is stored in their European datacentres / ??	??
Encryption	AES 256-bit Encryption	Allows for customer key management, rotation and storage

Dropbox Business		
	Default	Description based additions
Geolocation	United States	15+ Users: Germany, Australia and Japan
Encryption	AES 256-bit Encryption	Device Approvals, Enterprise Mobility Management (EMM)

Figure 5.7: Overview over the differences concerning security in the three public cloud providers [18, 19, 16, 20, 21, 15]

5.4.2.1 Location

According to [18], Box does not have its own data storage center but uses commercial data center providers mostly in the United States to store their data. Concerning the location where the user's data is stored Box provides the add-in Box Zones which is available to purchase but the pricing is not available on their website. Box Zones allow the user to decide where to store his or her data. The zones that are available in Box Zones are Europe, the United Kingdom, Asia, Japan, Canada, Australia, and the United States. The information where the data is being stored, if Box Zone add-in is not purchased, was not available on the website.

One interesting fact that was encountered, when trying to get information about Egnyte connect was that to receive any information at all, the user has to provide its personal information, such as name, email address, its encryption, the company it is working for as well as the companies size.

Furthermore, they do not disclose where the user's data will be stored on their website. The only hint is that they promise that if the user's company is located in the EU then its data will be stored in a data center located in the EU.

On their website [19] they state that the data of Dropbox business customers is stored in the United States. Business customers have the option, with the condition of having 15+ seats, to store their data in Germany, Australian, and Japan. If this option would include additional costs is not disclosed on their website.

5.4.3 Encryption

Box uses the AES 256-bit encryption for all its data at rest and in transit. As can be seen in Figure 5.8, the additional options available by purchasing a more expensive solution are a quite a few. Some of them can be relevant to businesses for example the 2-fact authentication for external users, *e.g.*, if a company has a lot of external users.

Security			
Full encryption in transit and at rest	✓	✓	✓
Granular access and collaboration controls	✓	✓	✓
User-enabled 2-factor authentication	✓	✓	✓
Shared link password security and link expiration controls	✓	✓	✓
Usage logs	✓	✓	✓
File statistics	✓	✓	✓
Bulk managed user provisioning	✓	✓	✓
Box Verified Enterprise	✓	✓	✓
Box Accelerator	✓	✓	✓
Box Network Connect (with AT&T and NTT)	✓	✓	✓
Content manager		✓	✓
Admin activity tracking	✓	✓	✓
Admin role delegation		✓	✓
Test environments (up to 5)		✓	✓
Security tab		✓	✓
Password policy enforcement		✓	✓
2-factor authentication for external users		✓	✓
Device trust		✓	✓
Document watermarking		✓	✓
Box Zones*** In-region data storage	💰	💰	💰
Options include (pick any 6):			
US (Box Service Providers)			
UK/Germany (IBM London-IBM Frankfurt)			
UK (AWS London-Azure Cardiff)			
Germany/Ireland (AWS Frankfurt-Dublin)			
Japan/Singapore (AWS Tokyo-Singapore)			
Japan (AWS Tokyo-Azure Osaka)			
Canada (AWS Montreal-IBM Toronto)			
Australia (AWS Sydney-Azure Melbourne)			
Box KeySafe	💰	💰	💰
Independent key control,			
Unchangeable audit log			
Content kill switch			
Options include:			
AWS KMS, AWS KMS Custom Key Store, AWS KMS GovCloud			

Figure 5.8: Additional features of Box, upon purchase of a more expensive solution [22]

A more interesting feature is the paid add-in Box KeySave. Box Keysave offers the user the possibility of using its encryption key. The process can be seen in Figure 5.9. The keys are stored (not in plain text) by AWS which has a 99.99999999% durability.



Figure 5.9: Optional Box KeySave [23]

Based on their website [16] and as can be seen in Figure 5.10, Egnyte bases the level of security on the subscription bought. The general set up is equipped with AES 256-bit encryption as well as Tier II SSAE-16 Compliant Facility Storage. The additional options provided by the more expensive subscriptions are shown in Figure 5.10.

DATA SECURITY			
Tier II SSAE-16 Compliant Facility Storage	✓	✓	✓
256-bit AES encryption ⓘ	✓	✓	✓
Advanced Authentication ⓘ		✓	✓
Multi-factor Authentication		✓	✓
Centralized Device Management		✓	✓
Egnyte Key Management SM ⓘ			✓

Figure 5.10: Subscription based security options of Egnyte [16]

In their paper [20] Egnyte Key Management is introduced, that allows on-primers storage as well as Amazon AWS ad Microsoft Azure Cloud storage for the keys used by the customers. This approach is especially useful as it provides the customers with a granular control of the security needed for their different datasets.

According to their whitepaper [21], Dropbox business uses also 256-bit AES encryption for data at rest. The files are fragmented and the fragments are encrypted and stored in multiple data center, while papers are stored over multiple reliability zones. Secure Sockets Layer (SSL)/Transport Layer Security (TLS) is used to protect the user’s data in transit using AES encryption.

		Standard €10/user/month	Advanced €15/user/month	Enterprise Contact us for pricing
		Try for free or purchase now	Try for free or purchase now	Contact us
256-bit AES and SSL/TLS encryption ?		✓	✓	✓
Advanced data protection				
Version history and file recovery ?		120 days	120 days	120 days
Advanced sharing permissions, including disabling downloads ?		✓	✓	✓
Password-protected and expiring shared links ?		✓	✓	✓
Remote device wipe ?		✓	✓	✓
Require two-factor authentication (2FA) ?		✓	✓	✓
Granular permissions ?		✓	✓	✓
Account transfer tool ?		✓	✓	✓
Enables HIPAA compliance ?		✓	✓	✓
Device approvals ?			✓	✓
Enterprise Mobility Management (EMM) ?				✓

Figure 5.11: Subscription based security options of Dropbox Business [21]

Further options are available on their website [15], as shown in Figure 5.11, provided if the customer decides to use advanced or enterprise packages are device approvals where the number of devices connected can be controlled as well as the security aspect of disconnected devices. Enterprise Mobility Management refers to the ability to manage security by a third party.

5.5 Private Cloud Storage Solutions

In the previous section Egnyte, Box and Dropbox business as an example for public cloud storage providers were covered. In this section, relevant private storage solutions and their differences will be covered. The evaluated private cloud storage solutions are all open source. Moreover, two ways to store the user’s data in personal cloud storage are introduced. The first is a Network Attached Storage (NAS) and the second is using a web hosting provider *e.g.*, Hostinger.

5.5.1 Network Attached Storage (NAS)

Network Attached Storage (NAS), is a storage system that can be accessed via the network and provides central management, backup options, high availability, data sharing, rapid deployment, as well as file system functionality [24]. An important aspect of using a NAS system compared to an external hard drive is, that using the NAS one can share and see the same type of data on different machines accessing the network of the NAS between client and file system [24]. Furthermore, they represent different storage disks as if it was one [24]. The cost of buying a NAS system is shown in Figure 5.12, where one can see the options for 4 TB, 10 TB and 140 TB available on digitec.ch [25]. The number of TBs was chosen in such a manner to make it comparable to the pricing of the public cloud. Since there was no “unlimited” option it was decided to display the price of the biggest storage available on digitec.ch which is 720 TB. This are all fully equipped solution but it would be possible to buy a disk-less NAS as well and choose the hard-drives used independently.

NAS options on Digitech (accessed: 27.10.2019)				
Price	160 CHF	1165 CHF	499 CHF	3434 CHF
Model	Buffalo LINKSTATION 210	Buffalo TERAStation 5200 WSS 2016	Synology DS118 Softbundle	ioSafe 1019 + Diskless
Storage	4TB		10TB	140TB

Figure 5.12: Pricing of NAS available on digitec.ch: [25]

5.5.2 Web hosting providers

Web hosting providers offer packages for hosting, *e.g.*, storage space, in a private cloud [26]. Examples for web hosting providers are Amazon EC2, Elastra, VMware [26]. To create private cloud storage, while taking advantage of the web hosting providers packages and not having to set up everything, can be achieved using a virtual private cloud provided by one of those web hosting providers. The difference to a public cloud is that a virtual private cloud is isolated from the internet. This can be achieved by creating, for example, custom IP addresses, and routing tables [27].

5.5.3 Nextcloud

Nextcloud is a platform for file sharing and collaboration using an on-premise environment [28]. Nextcloud is a server that runs on Apache or Nginx. The databases for storage, that are supported by Nextcloud are MariaDB, PostgreSQL and MySQL [28]. Furthermore, any storage protocol can be used *e.g.*, Network File System (NFS) depicted in the image as well as object stores compatible with SWIFT and S3. Moreover, it is possible to integrate WebDAV and external cloud storage services, such as Google Drive and others [28]. The data is accessible through a web browser as well as android and IOS apps. Nextcloud storage can be mounted on their server or through its interface supporting object stores or compliant systems, such as NFS, FTPs [28]. In their whitepaper [28] they suggest using object storage instead of NAS, as it should make the system of the user most suitable for handling large files. But further options like local storage, cloud storage, FTP and anything that can be locally mounted on the server are supported.

Each user can have his/her storage path using LDAP or Active Directory and therefore, different users can have different allowances, on which storage solutions to use *e.g.*, if they want to integrate public cloud storage, such as Dropbox. The pricing options for Nextcloud is depicted in Figure 5.13.

Nextcloud Enterprise									
	Basic			Standard			Premium		
Maximum of users	<50	<100	100<	< 50	<100	100<	<50	<100	100<
Price per year	1900€	3400€	??	3400€	6100€	??	4900€	8900€	??
Some of the additional features	<ul style="list-style-type: none"> access to security and stability fixes direct access to the most advanced security expertise 			<ul style="list-style-type: none"> Better customer support ONLYOFFICE for real-time 			<ul style="list-style-type: none"> SLA on fixes Clustering support Global Scale setup 		

Figure 5.13: Nextcloud solution options [29]

5.5.3.1 Security

Concerning the security of their application they provide an extensive description of how they try to develop securely, concerning their developers, as well as using regular security scans and a security bug bounty program [28]. They apply passive security hardening capabilities, such as content security policy, same-site cookies and brute force protection as well as active security measures, such as two-factor authentication with device specific passwords etc.

The server-side encryption allows for making external storage acts as a blind storage server. Therefore, the user does not need to trust third-party storage providers [28]. Furthermore, keys can be managed externally or internally and are not stored unencrypted on permanent storage [28].

File Access Control blocks access requests based on rules, such as IP address, user/group, file type and more positive security model compatible with *mod_security* compliant Web Application Firewall (WAF) solutions.

5.5.4 Seafile

Seafile makes use of libraries that can be synced and encrypted by a password separately while having an owner that sets its read and write permissions [30]. The architecture of Seafile, as can be seen in Figure 5.14 consists of a Seahub which is the web front-end and runs on Gunicorn or WSGI mode, a Seafile server which handles upload and download and Ccnet server that allows for internal communication between components [30].

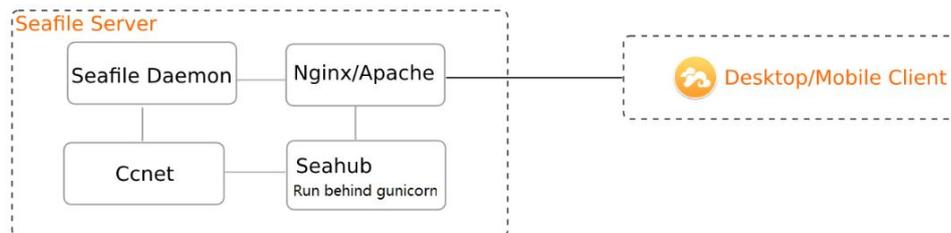


Figure 5.14: Seafile client access file, configured behind Nginx/Apache [30]

Furthermore, Seafile supports different Storage backends including Amazon S3, Ceph, OpenStack Swift. Some of these options can only be used when purchasing the professional edition depicted in Figure 5.15. The pricing options of Seafile. The community edition is open source and therefore free, while the professional edition which comes with features concerning management, support and security, are illustrated in Figure 5.15.

Seafile								
	Community Edition	Professional Edition						
Maximum of users	unlimited	<3	<9	<249	<499	<749	<999	1000+
Price per year	free	Free	100\$	44€/User	40€/User	35€/User	30€/User	??
Some of the additional features		<ul style="list-style-type: none"> File locking Fine grained folder permission Syncing LDAP/AD Users and Groups Single Sign On with ADFS 			<ul style="list-style-type: none"> Role based Account Management Remote Wipe Audit Log Antivirus Integration Scalability/HA AWS S3/Ceph RADOS 			

Figure 5.15: Seafile solution options [31]

5.5.5 Pydio Cells

Pydio Cells is a highly available and distributed system that uses micro-services structuring their architecture as loosely coupled services, that communicate via HTTP connections [32]. In Pydio Cells data storages can be created, such as mounting file systems on a server. The supported sources are local and remote file-systems, as well as object storages. They further allow for each data source to maintain its indexing in a dedicated database, therefore, allowing for fast indexation. To control access within Pydio Cells, they define roles which consist of an access control list, actions, and parameters. Those can, therefore, be attached to the user [32]. Pydio offers an IP ban, which bans a user with a specific IP address if it connects more than 10-times under 5 seconds. Furthermore, they can white list and blacklist the different IP addresses [32]. Pydio uses AES-GCM 256 bits encryption key to protect the data to allow for a third party storage provider [32]. They further argue that AES-GCM is more secure. This option does not provide a stream mode, therefore they divide the stream into different fixed-size blocks and encrypt each block with unique parameters [32]. Furthermore, Pydio Cells implements security policies that can be base on different criteria like source IP access time, resource type endpoint type.

Pydio Cells				
	Open source home	Enterprise		
Maximum of users	Unlimited	< 20	<100	100<
Price per year	Free	1620€	4980€	??
Some of the additional features	-	<ul style="list-style-type: none"> AD/LDAP connector Security policies on REST request Role-based Access Control Lists and security policies Auditable logs 		

Figure 5.16: Pydio cells solution options [33]

The paid enterprise solution as can be seen in Figure 5.16 offers the integration of authenticators like google authenticator app for multi-factor authentication [32]. They further offer an LDAP plug-in for external authentication.

5.6 Stress Test Performance

In January 2018 [34], a paper about stress test on Nextcloud, Seafile and Pydio Cells was published. The authors performed these tests for disaster scenarios, such as the earthquake that happened on April 16, 2016, in Manabi. This earthquake shut down a whole educational sector [34]. The servers used for the test were: Intel (R) Xeon (R) CPU E3-1220 v3 @ 3.10GHz with 4 cores; 4 GB RAM; Ubuntu Server 14.04 with a kernel 3.13.0-85-64-GNU/Linux Ubuntu; two hard drives of one TB. They further used JMeter version 2.13.20 to test performance and functional behaviour; Cacti that stores and graphically display the data from the database; Mrtg a daemon that makes data available for other users [34]. A request was the download of a 500 KB file.

In Figure 5.17, is depicted how many concurrent user requests could be performed before the system crashed with Seafile only supporting less than 2300 and Nextcloud 10603 concurrent user requests [34]. In Figure 5.18, the CPU consumption after 1, 5 and 15 minutes of execution of the stress test is illustrated. Further in Figure 5.20, the Memory consumption when performing 6000 requests is presented. Finally in Figure 5.19, the meantime to execute the concurrent user requests with a set of 1000, 6000 and 10572 requests are displayed [34]. According to this stress tests, Nextcloud seems to be the best performing private cloud storage under stress conditions.

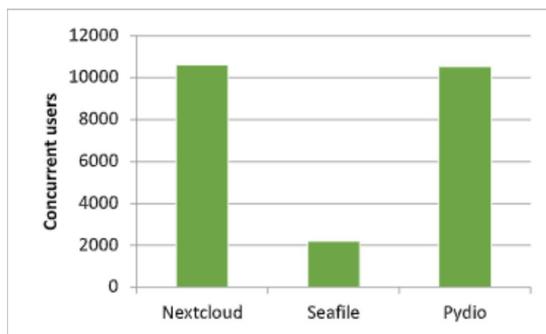


Figure 5.17: Concurrent user requests [34]

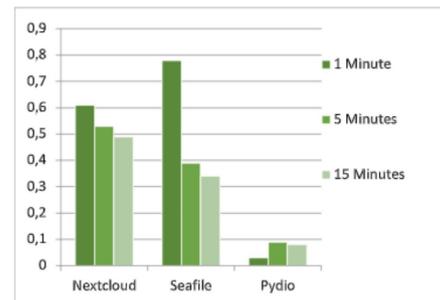


Figure 5.18: CPU consumption [34]

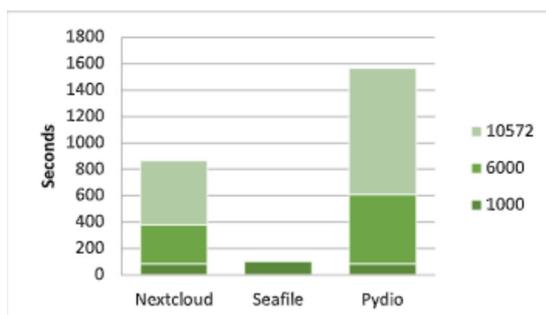


Figure 5.19: Time for user requests [34]

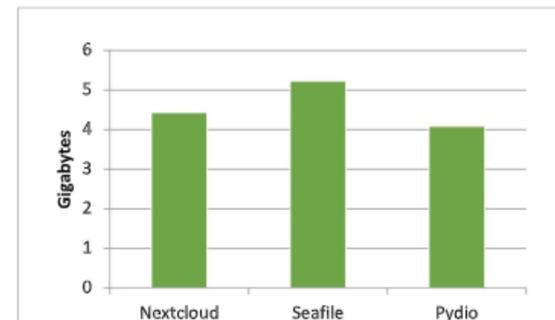


Figure 5.20: Memory consumption [34]

This was interesting for this report as in a private cloud solution the user is responsible for the backup and redundancy of its data and therefore the eventual loss of data that could occur if an earthquake happens.

5.7 Discussion

In this report, three Private Cloud Service Providers Nextcloud, Pydio Cells, and Seafile, as well as three public cloud providers Dropbox Business, Egnyte, and Box, were presented and analysed in respect to pricing and security. The following section discusses the differences between the two and also present two alternative solutions.

5.7.1 Pricing Differences

Private cloud storage is by far more expensive concerning their set up especially considering that the company using a private cloud is responsible for backups as well as redundancy to avoid data loss, therefore, they are only able to use about half of their storage. The prices to buy a NAS that is fully equipped, is expensive, and the cost that should be included using a private cloud is not only the set-up but also the maintenance and technical expertise that are needed to run it. In comparison the public cloud storage is much less expensive and consists of a subscription plan and is incremental; thus, there are no wasted resources. An advanced version that is more expensive might be needed when looking at the public cloud since it provides very useful additional functionality.

To give an example comparison, it was presented as a reference the price of 4 TB, which choosing the cheapest set up in digitec.ch is 160 CHF, but there is the problem that the user is responsible for redundancy and backup. Choosing RAID to avoid data loss the private cloud user would only be able to use half of the storage space available; therefore, to compare the standard for Dropbox Business with 3 TB and office option for Egnyte an user would have to purchase storage of at least 6/8 TB respectively. Therefore 8-10 Euro per user per month can be compared with the set-up cost of 499 CHF excluding the personnel needed for the set up, as well as the maintenance of the data storage costs. This difference might not be as big as one might have guessed but looking at more storage consumption, such as unlimited in comparison to 720 TB if the usage of RAID 360 that is available on digitec.ch for 47'448 CHF, was decided.

5.7.2 Privacy Differences

When looking at privacy the difference between public and private cloud is noticeable. The information where the user's data will be stored by default is not available in some of the solutions and the United States in Dropbox business. Box and Dropbox allow the data to be stored in a country of the user's choice, if they have a data center in that location. However, to make use of this option the user has to upgrade its subscription and/or pay more. To comply with the General Data Protection Regulation (GDPR) is up to the user of the public cloud. The location of the data in a private cloud is up to the user. The advantage is, that the user knows who has physical access to the data center and that they are in compliance with GDPR. However, the drawback of this solution is that the user is responsible for backing up the data and creating data redundancy *e.g.*, using RAID to avoid data loss. In a public cloud, the data is secured using AES 256-bit encryption, in most cases. Moreover, all three options that were showed allow for an upgrade of this security, such as Enterprise Mobility Management, Custom key management, and Box Keysave. When looking at security in private cloud solutions, the security is up to the owner of the private cloud. Therefore, the user can decide how secure it wants to store its data and are in no way dependent on the upgrades or usage of new security solutions of their cloud storage provider.

5.7.3 Alternative Solutions

In this section, two alternative solutions to the categories public and private cloud storage, are shown. Hybrid cloud storage and community cloud storage are introduced and their advantages as well as drawbacks are discussed.

5.7.3.1 Hybrid Cloud

Hybrid cloud storage is a cloud storage type, where at least one part of the data is stored in a public cloud and at least one part of the data is stored in a private cloud [4], which are connected such that data can be transferred from one to the other. [4] lists the following advantages:

- *Security*: Sensitive data can be stored in the private cloud, therefore, increasing the overall security.
- *Cost*: Not all resources need to be bought and maintained by the consumer company since some can be used by the public cloud storage provider.
- *Scalability*: Rapid scaling is possible using the public cloud.

In [4], the following drawbacks are listed:

- *Attack Surface*: The attack surface of the company is greater since a part of the hybrid infrastructure is managed by the service provider.
- *Complexity and Security*: Since a complex cloud needs to be managed, the company has to decide whether they want to use a management tool and what kind of security risks this will introduce.
- *Data movement*: One of the advantages is the ability of data movement between the different cloud storage. This will tough introduce security concerns as the two have very different privacy controls.
- *Encryption keys*: As there are two different clouds that handle encryption keys differently, different security considerations are needed.

5.7.3.2 Community Cloud

A community cloud concerning storage is formed by a small set of companies that share their cloud storage within them and therefore it is neither just for one consumer and neither for everyone available [4]. Therefore, the companies must share privacy and security regulations. According to [4] community cloud storage advantages are:

- *Cost*: Since the cloud storage is shared the cost of purchase is shared and therefore smaller.
- *Management*: The management can be handled by a third party that is unbiased concerning the different companies and therefore avoid unfair distribution of resources.
- *Tools*: The information stored within the community cloud can be used to the advantage of the customers *e.g.*, if the companies are connected through a supply chain.

The drawbacks that are identified by [4], are:

- *Cost*: Even though the cost is lower than that of a private cloud storage, it is still higher than the cost of a public cloud storage solutions.
- *Amount of Storage*: The amount of storage that is available to each consumer of the public cloud storage is limited.

5.8 Conclusion

In this report, public and private cloud storage solution according to price and security were compared. The price of private cloud storage is much higher than the price of public cloud storage concerning the initial set up and eventually the maintenance of the storage and expertise needed for this storage. Looking at the security aspect it was found that the most used encryption in the public cloud examples is AES-256 encryption while the information about where the data of the user is stored is missing or poor. In a private cloud, the user is responsible for how secure its data is concerning *e.g.*, firewall and anti-virus software but it is initially protected since it is not sending its data over the internet and concerning the storage, it knows where its data is stored and who has physical access to it. Therefore, private cloud storage is more secure than public cloud storage.

Further, it was introduced two alternatives, hybrid and community cloud storage. The options of a hybrid cloud might also be interesting to some businesses that have a limited amount of sensitive data requiring higher security standards and are willing to accept the drawback such as limited data movement between public and private cloud storage. Finally, the community cloud storage, which can be an option for companies that share some common interest and user information, was introduced. In summary, it depends on the specific company and the data they want to store, which solution should be chosen.

5.9 Limitations and Future Work

It is important to note that this report covered the main examples in public and private storage in the market. However, this market is growing and novel solutions appear. Thus, it is not a trivial task to provide a recommendation of which is the best alternative or solution, as the choice is directly dependent on the use-case, and on security, price, and data requirement. One major drawback is that there was no communication with companies that use public, private, hybrid or community cloud storage to understand their experience and how they deal with security and what the price *e.g.* of maintaining the private cloud is. Moreover, the information for the different examples in this report is all available on the website of the providers. Therefore, information is missing which could only be obtained when contacting the cloud storage providers. With this report, a non-exhaustive list of providers was compared. Thus, it would be interesting to further cover how much public cloud providers gain when a firm subscribes to their services and what kind of subscription/amount of users per subscription would be needed for them to make a profit.

In this report, the pricing of public and private cloud storage was analysed, but it would be interesting to analyse *e.g.*, maintenance costs for private cloud storage. It would be further interesting to look at the users and their experience with the different cloud storage types. For example, talking with companies that use public or private cloud storage and ask them about their experience and their reasons for choosing one or the other. Besides the Hybrid and Community cloud storage solution, there are also other alternative solutions and/or approaches that could be analysed.

Bibliography

- [1] B. P. Rimal, E. Choi, and I. Lumb, "A taxonomy and survey of cloud computing systems," in *2009 Fifth International Joint Conference on INC, IMS and IDC*, pp. 44–51, Ieee, 2009.
- [2] Cloudwards.net, "Best Cloud Storage Reviews 2019." [Online] <https://www.cloudwards.net/cloud-storage-reviews/>, last visit September 25, 2019.
- [3] D. Slamanig and C. Hanser, "On cloud storage and the cloud of clouds approach," in *2012 International Conference for Internet Technology and Secured Transactions*, pp. 649–655, IEEE, 2012.
- [4] S. Goyal, "Public vs private vs hybrid vs community - cloud computing: A critical review," *International Journal of Computer Network and Information Security*, vol. 6, pp. 20–29, 02 2014.
- [5] O. Arki and A. Zitouni, "Cloud Storage and Security Overview," in *International Conference on Advanced Aspects of Software Engineering (ICAASE18)*, (Constantine, Algeria), pp. 26–33, December 2018.
- [6] A. P. Rajan and S. Shanmugapriyaa, "Evolution of cloud storage as cloud computing infrastructure service," *CoRR*, vol. abs/1308.1303, 2013.
- [7] V. Spoorthy, M. Mamatha, and B. S. Kumar, "A survey on data storage and security in cloud computing," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 6, pp. 306–313, 2014.
- [8] C. Schwarzenegger, *Nutzung von Cloud-Diensten durch AnwÄltinnen und AnwÄlte = Utilisation des services de cloud par les avocates et avocats*, vol. volume 4 of *Schriften aus dem ITSL*. ZÄrich: Schulthess, 2019.
- [9] V. Geetha, N. Laavanya, S. Priyadharshiny, and C. Sofeyikalaimathy, "Survey on Security Mechanisms for Public Cloud Data," in *International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS 2016)*, (Pudukkottai, India), pp. 1–8, Feb 2016.
- [10] C. V. Desai and G. B. Jethava, "Survey on data integrity checking techniques in cloud data storage," *International Journal*, vol. 4, no. 12, 2014.
- [11] Dropbox, Inc., "Welcome to the smart workspace.." [Online] https://www.dropbox.com/en_GB/smart-workspace, last visit November 10, 2019.
- [12] Egnyte, Inc., "About Us." [Online] <https://www.egnyte.com/about-us.html>, last visit November 10, 2019.
- [13] Box, Inc., "Meet Box." [Online] <https://www.box.com/en-gb/about-us>, last visit November 16, 2019.

- [14] Cloudwards.net, “Enterprise File Sync and Share REviews of 2019.” [Online] <https://www.cloudwards.net/enterprise-file-sync-reviews/>, last visit November 16, 2019.
- [15] Dropbox, Inc., “Choose the right solution for you,” 2019. [Online] <https://www.dropbox.com/business/plans-comparison>, last visit September 15, 2019.
- [16] Egnyte, Inc., “Egnyte Connect Plans,” 2019. [Online] https://www.egnyte.com/corp/plans_pricing.html, last visit November 26, 2019.
- [17] Box, Inc., “Choose a plan that’s right for your business.” [Online] https://www.box.com/en-gb/pricing?utm_source=Google&utm_medium=SEM&utm_campaign=DM-Google-SEM-Global-ENG-Sitelinks_Branded_Keywords&utm_keyword=box%20pricing&id=7010e00000axJc&utm_content=267298059333|aud-302644801190:kwd-301876084226|c&gclid=Cj0KCQiAn8nuBRCzARIsAJcdIf0e0EYyo2Or_6JxzME1MxNV6y7Q9qsTrKxdEB1GwvwpVnBK0Qsm1fwaApRFEALw_wcB, last visit November 4, 2019.
- [18] Y. Sverdlik, “Five Most Interesting Things for Data Center Pros in Box’s Pre-IPO Filing,” 2014. [Online] <https://www.datacenterknowledge.com/archives/2014/07/14/five-most-interesting-things-about-box-data-centers-in-sec-filing>, last visit November 1, 2019.
- [19] Dropbox, Inc., “Where is my data stored?.” [Online] <https://help.dropbox.com/accounts-billing/security/physical-location-data-storage>, last visit November 2, 2019.
- [20] Egnyte, Inc., “Egnyte Key Managment Gives Enterprises Control Over Privacy and Securiry,” 2016. [Online] https://egnyte-www-static.egnyte.com/assets/pdfs/press-releases/2016-03-01-egnyte-key-management-gives-enterprises-control-over-privacy-and-security.pdf?_ga=2.86382232.1034260575.1571928885-257809673.1571567588, last visit October 25, 2019.
- [21] Dropbox, Inc., “Dropbox Buisness Security,” 2019. [Online] https://aem.dropbox.com/cms/content/dam/dropbox/www/en-us/business/solutions/solutions/white_paper/dfb_security_whitepaper.pdf, last visit September 28, 2019.
- [22] Box, Inc., “Box - Feature Matrix - Box Business Plans (External) [Aug 06, 2019].pdf,” August 6 2019. [Online] <https://cloud.app.box.com/v/BoxBusinessEditions>, last visit September 15, 2019.
- [23] Box, Inc., “Box KeySave Manage your own encryption keys.” [Online] <https://www.box.com/en-gb/security/keysafe>, last visit October 16, 2019.
- [24] G. A. Gibson and R. Van Meter, “Network attached storage architecture,” *Commun. ACM*, vol. 43, pp. 37–45, Nov. 2000.
- [25] Digitec Galaxus AG, “Search history,” 2019. [Online] <https://www.digitec.ch/de/s1/producttype/netzwerkspeicher-nas-68?tagIds=7-1171>, last visit September 20, 2019.
- [26] R. Prodan and S. Ostermann, “A survey and taxonomy of infrastructure as a service and web hosting cloud providers,” in *2009 10th IEEE/ACM International Conference on Grid Computing*, pp. 17–25, Oct 2009.

- [27] D. Yang, H. Wei, Y. Zhu, P. Li, and J. Tan, "Virtual private cloud based power-dispatching automation system-architecture and application," *IEEE Transactions on Industrial Informatics*, vol. 15, pp. 1756–1766, March 2019.
- [28] G. Nextcloud, "Nextcloud solution architecture," whitepaper, Nextcloud GmbH, 2018.
- [29] Nextcloud, GmbH., "Plans and Pricing for Nextcloud Enterprise." [Online] <https://nextcloud.com/pricing/>, last visit October 16, 2019.
- [30] Seafile, Ltd., "Manual for Sefile server." [Online] <https://download.seafile.com/published/seafiler-manual/home.md>, last visit November 8, 2019.
- [31] Seafile, Lnt., "Deploy Seafiler on your Own Server," 2019. [Online] https://www.seafile.com/en/product/private_server/, last visit October 19, 2019.
- [32] G. Hecquet, "Achieve high availability with cells enterprise," white paper, pydio cells, 2019.
- [33] Pydio, "Pydio Cells Pricing." [Online] <https://pydio.com/en/pricing>, last visit October 19, 2019.
- [34] A. Delgado-Domínguez, W. Fuertes, and S. Sanchez-Gordon, "Enterprise file synchronization and sharing services for educational environments in case of disaster," *Revista Facultad de Ingeniería UPTC*, vol. 27, 01 2018.

Chapter 6

Data Collection and Wireless Communication in Internet of Things (IoT) Using Economic Analysis and Pricing Models

Matej Jakovljevic, Jeremy Kubrak, Dylan Puser, Marc Zwimpfer

Business processes have largely changed during the age of digitalization. The internet of things (IoT) will probably have a great impact on several business sectors by allowing further interconnection between different processes. Hence, these businesses are forced to adapt their traditional pricing-models to the newly emerged IoT-systems.

This paper aims towards a deeper understanding of the economical impact of IoT in different business sectors. To do so, selected protocols of IoT-systems and pricing-models are described. Furthermore, the pricing models will be applied to several business sectors in which IoT enables new possibilities. Moreover, a use case of a AWS IoT Core illustrates a possible pricing model for IoT services. Finally, a short overview of the challenges and limitations the IoT-systems face, are provided and discussed. As IoT-systems rely heavily on available data, the critical aspects of data processing are reviewed.

Contents

6.1	Introduction	151
6.2	IoT-Systems	152
6.2.1	Wireless Sensor Networks	152
6.2.2	Machine-To-Machine Communication (M2M)	153
6.2.3	IoT-Architecture	153
6.3	A Case Study on Different IoT-Platforms and their Respective Pricing	155
6.3.1	IoT-Platforms	156
6.3.2	Difficulties in Comparing Different IoT Platforms	158
6.3.3	Compare IoT Platforms Qualitatively	159
6.3.4	Compare IoT Platforms by Use Cases	161
6.4	Pricing-Models	163
6.4.1	Economic Concepts Based Pricing	163
6.4.2	auction-based Pricing	165
6.4.3	Game Theory	166
6.4.4	Utility Maximization and Knapsack Problem	167
6.4.5	Combination of Pricing-Models	168
6.5	Application of Economic and Pricing-Models for IoT-Systems	168
6.5.1	Pricing-Models for Data Exchange and Topology Formation	168
6.5.2	Application for Resource and Task Allocation	172
6.5.3	Pricing-Models for Sensing Coverage and Target Tracking	173
6.5.4	Pricing-Models for Improving Privacy Security	175
6.5.5	Pricing-Models in M2M Communications	176
6.5.6	Applicability on Business Sectors	176
6.6	Challenges and Limitations	177
6.6.1	Technical	177
6.6.2	Business	178
6.6.3	Social	178
6.7	Discussion	178

6.1 Introduction

The internet of things (IoT) is the interconnection of physical objects to the internet containing electronic parts in order to communicate and sense interactions between those physical objects or their external environment. The goal of IoT is to gather data from the environment and gain insights from the data or even allow devices to automatically respond accordingly to the data gathered. There is a huge variety of products IoT has brought forth, such as bike helmet crash sensors, smart tennis rackets, soil monitoring, retail analytics and elderly care monitoring.

From an economical perspective, the IoT applications can be interpreted as services or products. To sell those goods, the pricing model for the offered goods is key to remain profitably in the market and satisfy the customers' demands. Although most IoT products may not be compared with traditional goods, the need to take over existing pricing models and adapt them to the business is essential. Pricing models are part of our everyday life, as we encounter them by paying our mobile subscription or any other subscription. Depending on the product and customer, a flat rate for a mobile subscription might make sense, whereas the same customer might ask for a usage-based billing for his TV subscription. The customers as well as the nature of the good to sell allow to create different pricing models. This of course applies also to IoT-systems and their services.

In order to apply the traditional pricing models to IoT, the factors to calculate a price have to be determined first. Through the introduction of IoT, some services may be automated with IoT which would then lead to the classical pricing models being affected by IoT.

With IoT emerging, business sectors will want to adapt their products and test whether IoT may be beneficial to change an existing product into something more viable. Business sectors that are potentially affected by IoT are insurance, agriculture, healthcare and logistics.

This paper aims to foster the understanding of IoT and applicable pricing models for products that emerge from IoT. Moreover, some prices of existing IoT-platforms are calculated with an example use case in order to compare their pricing model, their functionality and the costs for a exemplary IoT-system.

6.2 IoT-Systems

6.2.1 Wireless Sensor Networks

WSNs can be described as wireless networks that consist of autonomous devices that are spatially distributed. These devices carry sensors that enable it to monitor physical and environmental conditions, such as temperature, pollution, sound and so forth [7]. The nodes in the WSN may exchange their sensed data from node to node, but also the more common server-client architecture is possible to transmit the data to a central instance, such as the server or gateway. The different interaction paradigms are depicted in figure 6.1. The figure immediately shows the main difference between the mentioned paradigms.

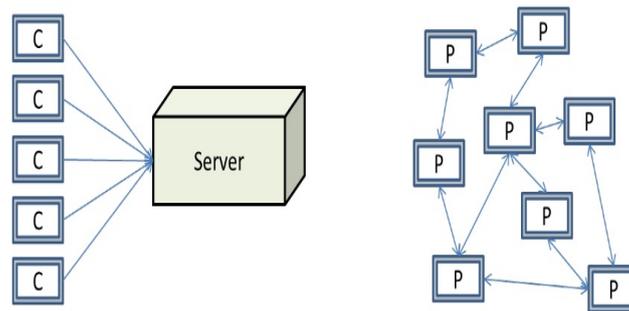


Figure 6.1: Client-Server interaction paradigm and the peer-to-peer interaction paradigm. [7]

In a server-client interaction paradigm, the data is transmitted to a central unit, which then distributes the data to the individual nodes again, whereas in a peer-to-peer network each node is able to exchange data with other nodes within the network. Apart from the interaction paradigm, other properties of networks, such as stateful or stateless servers are to consider when deciding for an IoT-system. Further information about the architecture and protocols will be given in section 6.2.3.

To foster the understanding of a WSN, the upcoming figure shows a schematic structure of a WSN. Furthermore, due to choice of protocol sensor nodes often do not have a global ID, which are usually included in TCP/IP, in order to reduce the overhead to establish and maintain the connection from device to device. The unavailability of global ID's for sensor nodes also has the effect that a cyber attack may not target the sensor nodes directly, but rather the gateway is attacked to gain control over the whole sensor network.

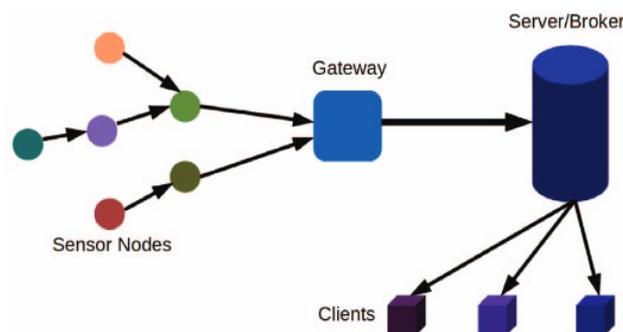


Figure 6.2: Illustration example of a WSN with sensor nodes and a gateway. [35]

Figure 6.2 shows a typical network scheme of a WSN. It is clearly visible that the sensor nodes (small circles) propagate their sensor data to the next available node and finally send it to a gateway node. The gateway node furthermore transmits the collected data to a server, in which typically the aggregated data will be processed to gain insights out of

the sensor data. To transmit the data, often a publish-subscribe architecture is chosen in order to serve several applications with the same data. The figure may only be seen as an example, as the data transmission and aggregation is heavily dependent on the structure of the network. This figure illustrates an example network that transmits the data up to a server which then analyses data and sends the information to client applications, typically called cloud computing. In section 6.2.3.4 we will cover the possibilities of data processing more in-depth.

6.2.2 Machine-To-Machine Communication (M2M)

As IoT depends on the interconnection and exchange of data between various devices, M2M communication is a concept that is applied in the field of IoT. M2M can be defined as the data exchange between devices without the need of human interaction to transmit the data. According to a paper [7] a big amount of connected devices will be deployed by 2020. The key features of M2M offers are [19]:

- Time Controlled, meaning that data can only be send or received at certain pre-defined methods
- Time tolerance implies that the data transfer may be delayed
- Low Power Consumption
- Small Data Transactions in order to enable real-time transmission of data on low-bandwidth networks
- Monitoring

Especially the monitoring aspect is interesting, as most protocols that are used in IoT do not offer the possibility to monitor devices. M2M communication could serve different applications in various areas such as security, tracking, payment, facility management and remote maintenance [19].

6.2.3 IoT-Architecture

This subsection presents information about the different topologies and the actors within an IoT network as well as the gateways to connect two different networks with different protocols. Furthermore, some of the most prominent protocols for messaging and connection establishment are described.

6.2.3.1 Requirements

This section describes a small set of naturally formulated requirements for the implementation of an IoT-System.

As the requirements differ from perspective to perspective, the requirements in terms of business as well as technical requirements are mentioned. For business requirements, a successful IoT-System should include real-time needs, availability of applications, data protection, user privacy, efficient power consumption and a dynamic resource demand. Ideally, the system should be accessible over an interoperable cloud-system [23]. A more technical view suggests that IoT-systems need the hardware, such as sensors, actuators and gateways, middleware, such as storage and computing tools for data analysis and a presentation layer that facilitates the generating of insights with the available data.

6.2.3.2 Interaction Paradigms

As mentioned before, the interaction paradigms are a crucial part of IoT-systems. Furthermore, the interaction paradigm chosen for a network may affect the price for an IoT-platform during the usage. Depending on the interaction paradigm, a different number of messages is sent within the network to achieve the same data transmission. From the examples above, imagine one sensor requesting data from two other sensors. In a peer-to-peer network, the sensor node requests data from the two sensors and receives two responses resulting in four messages sent. In a client-server interaction, the sensor node sends a request to the server to get the sensor data of the two other sensors. The server then requests the data from the two sensors and gets a response each. Only then, the server is able to pass the data to the requesting sensor node, resulting in 6 messages sent within the network. Depending on the IoT-platform chosen, which will be covered in a later section, the number of messages sent within a network has an influence on the costs of the usage of an IoT-platform. Therefore, we can conclude that the interaction paradigm chosen for an IoT-system may influence the costs directly.

6.2.3.3 IoT Session/Messaging Protocols

As proposed in the section above, the choice of protocols may also directly influence the costs of an IoT-platform. Below some of the basic protocols used today are described.

MQTT stands for 'Message Queuing Telemetry Transport' and is often used for the intercommunication of devices in a WSN. It is known to be extremely lightweight and simple and is a good fit for the use in constrained environment, where often the network bandwidth is relatively low. Moreover, the publish-subscribe architecture is designed to easily transmit messages within the network [7]. Through the small message size, MQTT minimizes the needed network bandwidth for real-time transmission and allows IoT devices, which are usually cheap and have limited memory and processing capabilities, to transmit data more reliably. A disadvantage of MQTT is that data transmitted is unstructured, meaning that the data needs to be preprocessed before it is possible to gain insights from it.

HTTP is the probably the most common protocol used by today. HTTP nevertheless can not be described as suitable for IoT-systems. IoT-devices usually require only little interaction, have only limited power and often face connectivity issues. Due to the heavy overhead in the HTTP protocol, it is generally too heavy to be used in IoT. Additionally, HTTP is a session-based protocol that establishes a session between two nodes and therefore sends more messages than a stateless protocol [7]. A HTTP request needs at least 9 TCP packets to submit an answer (without packet loss), so the amount of messages sent and the message size are too big to be using it in IoT. As stated before, the message size and the number of messages may influence the costs of an IoT-platform directly.

CoAP is another transfer protocol for constrained nodes and networks. CoAP uses the REST architectural style and is based on UDP [6]. In comparison to MQTT, CoAP offers security mechanisms built-in on the Datagram Transport Layer Security protocol. Just like HTTP, CoAP, unlike MQTT, is able to carry different types of payloads such as XML and JSON. Moreover, as CoAP is REST-based, getting and setting values from the nodes inside a network looks quite similar to an ordinary HTTP request. Nevertheless, through the choice of payload the message size is bigger compared to MQTT. Additionally, CoAP is primarily used in a server-client interaction paradigm (one-to-one protocol),

whereas MQTT enables a many-to-many communication through its publish-subscribe architecture [6].

6.2.3.4 Data Processing

With the ubiquity of sensors in the environment, much more data in raw form is available that needs to be processed in order to make predictions or send commands from it.

Cloud Computing makes computing resources (hardware, applications) available as services over the internet. Costly hardware and analysing tools have no longer to be bought separately but are often accessible over a subscription [12]. Cloud computing has revolutionized the existing data analysis. However, IoT also pushes cloud computing to its limits, as processing in a central entity requires all nodes in the network to send the data to the processing entity first. This both increases latency and saturates the network bandwidth [11]. Thus, several latency-sensitive applications, such as emergency-response systems, need a different approach to get a quick response.

Fog Computing is a distributed paradigm that provides the cloud services to the network edge. Fog computing involves the components of data-processing in distributed cloud and edge devices. Fog computing deals with data locally by utilizing edge devices near users to carry out the calculations [11]. Through the proximity to the sensors the latency is reduced, throughput is maximized and the results are computed faster than in a centralized cloud computer and are therefore able to provide the clients with real-time analytics. Nevertheless, fog computing is claimed to be more power-consuming than cloud computing [11].

Edge Computing takes the approach of fog computing to the edge of the network, more specifically where the data is produced [33]. In order to enable edge computing in a network, the edge nodes of the network need to have enough processing power to process the data. Although this decentralized approach guarantees the fastest response from data processing, it encounters several problems. One of the problems is that often these edge devices are not reliable enough. A loss in connectivity or an empty battery is often enough a challenge for edge nodes and by putting processing power in them, the reliability rather declines. The edge devices would also cost more than ordinary ones without processing power. Moreover, if an edge node is battery-powered, the battery will run out of energy much more quickly when the edge node is processing data [33].

The costs for a IoT-system are heavily dependent on the interaction paradigm chosen and the protocols used to transmit messages between the actors inside the network. Finally, we would like to mention that now only the costs of a IoT-platform depend on the aforementioned properties, but also the choice of hardware and the network on which the transmission takes place influence the total costs.

6.3 A Case Study on Different IoT-Platforms and their Respective Pricing

Before starting with the theoretical analysis of pricing models in IoT in the later part of this thesis, we establish two sample use cases in order to give a first idea of what the problems in IoT pricing consist. For this reason, we first introduce the concept of IoT platforms and their usage throughout IoT. Secondly, we compare three popular IoT platforms of large technology companies in means of their functionalities as well as their pricing models in regard to our established use cases.

6.3.1 IoT-Platforms

Designing and implementing a functioning IoT system brings certain difficulties with it. A large-scale IoT system should enable gathering big amounts of data from independent sensors, allow big numbers of devices to be connected and react to changes autonomously to some extent [15]. Furthermore, as in any communication system, the elements security, reliability and performance play a big role in the success of such a system as a whole. However, how can such a system be built in manner that all these requirements are met? An IoT platform is a kind of middleware software which acts as a hub connecting the big variety of devices in an IoT environment [30]. With the increasing number of heterogeneous devices and the difference in the protocols they use (e.g. standard HTTP or MQTT), an additional layer between the hardware (devices) and the software services is needed. An IoT platform provides a variety of functionality, however the main functionalities include the simple communication between the devices and the abstraction of the whole IoT ecosystem to the software services [29].

6.3.1.1 IoT Middleware

Since acting as an IoT middleware in an IoT system forms a big part of an IoT platform, we first establish the definition of IoT middlewares. The general term middleware stands for a middle level between different applications [29]. The big advantage of using middleware is that it provides a set of services to applications from outside which then are processed or delegated to other parts of a system. Due to this, the remainder of the system must not consider the differences in the outside application. In a broader sense, a middleware represents a mediator between a service provider and a service consumer, without revealing them to each other directly.

In the case of IoT, middleware is of utter importance as a big heterogeneity exists between the "things". Nowadays, in a big IoT ecosystem there are devices of many different manufactures. However, these devices should interoperate without any problems and other software parts of this IoT infrastructure should not need to handle the different kinds of devices. Furthermore, various protocols are used for communication in an IoT system. Similarly as before, an IoT middleware acts as a middle layer to hide the actual protocols used for communication so that the other parts of the system do not have to manage this. However, an IoT platform is not only an IoT middleware but provides more functionalities. Whereas the main goal of the middleware is to hide heterogeneity of an IoT system, an IoT platform enables for example managing devices, storing or analysing gathered data too.

6.3.1.2 IoT Platform Capabilities

Even though the functionality of any two different IoT platforms are different to each other, they all provide similar core functionalities [30]:

- Acts as IoT middleware
- Manages IoT devices
- Triggers actions when receiving messages
- Stores data
- Analysis of gathered data
- Provides interfaces to other cloud services

Device management Besides acting as middleware, an IoT platform provides means to manage all devices from a central point. With a growing number of independent, possibly heterogeneous, devices joining a IoT ecosystem, the maintenance of these "things" becomes very complex without the help of a central service which handles this task. IoT platforms allows an end-user to easily update and control a big variety of devices without a lot of effort [30].

Rule-based Actions With the communication and device management working properly, the next core task of a IoT platform is to act according to inputs received by individual devices. Take a temperature sensor as an example: After the sensor transmits temperature to the IoT platform, certain measures must be triggered. Two sample actions could be to message a temperature control device to change the temperature accordingly or warn an end-user that the temperature has changed. Due to this, IoT platforms provide means to create rules and according actions when receiving messages from different devices. This part of an IoT platform is the backbone of the whole IoT system. It is the part which enables the system to be "smart" and acting autonomously on changes in the environment which is one of the main goals of an IoT system [15].

Data Storage and Analysis The functionality regarding storage and analysis of data are connected closely. The data storage is used to store data received by the devices and data on the devices themselves [30]. This is needed as the other parts of the IoT platform may use the data in later stages to analyse or visualize it for the end-user. As real-time monitoring of the system environment is one of the key goals of IoT, the analysis of data forms a central part of the whole system. It is the part which provides actual perceived value for the end user [15].

External Interfaces Lastly, an IoT platform must provide a big variety of interfaces for external services. Many IoT platforms vendors provide additional cloud-services which may range between better visualization up to complex prediction frameworks based on artificial intelligence which all need the data gathered from the devices by the IoT platform [4, 14]. All these before mentioned functionalities are in most cases separate modules which form together an IoT platform [30]. The typical modular structure of an IoT platform is shown in Figure 6.3.

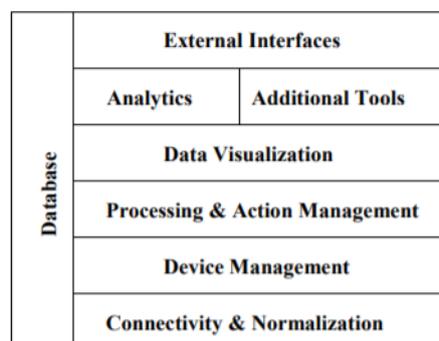


Figure 6.3: Standard modules of an IoT Platform [30]

IoT Platform Architecture Moving away from the functionalities of IoT platforms, their architecture may also vary. IoT platforms can either be centralized, semi- or fully distributed [30]. Centralized solutions are most common in the platforms provided by the big vendors as they already have an fully-established cloud service running in which the IoT platform is included [14, 4, 27]. Open-Source IoT platform solutions are more often designed in a distributed manner [29].

6.3.2 Difficulties in Comparing Different IoT Platforms

In today's market, a big variety of different IoT platforms exists. They range from small-scale solution to solution which can manage thousands of devices. These platforms can either be provided by vendors or are Open-Source projects. Hence, it is not trivial to compare different IoT platforms as they appear so differently. Nevertheless, when building an IoT system it is crucial to use the correct IoT platform as it forms the core of the whole system. For this reason, we provide three aspects which should not be neglected when comparing between different systems and finally deciding which one to use.

Functionality Functionality may be the most important factor when deciding which system to use. However, functionality forms also the part with the highest variance between the different IoT platform solutions. Thus, in order to compare different IoT platforms quantitatively and qualitatively, a comparison framework must be established. [30] already provide a framework which can be used to compare platforms as it is shown in Figure 6.4. They especially focused on the criteria of reliability, latency and scalability because they form the fundamental pillars of an IoT platform.

Security			Platform
<ul style="list-style-type: none"> - Authentication - Confidentiality - Data Ownership - Data Storing - Protection - Throat protection 	Usages	<ul style="list-style-type: none"> - performance - Correctness - Accessibility - Predictability - Usability - Comprehensibility 	<ul style="list-style-type: none"> - Testability - Scalability - Improvability - Reusability - Portability - Reliability - Flexibility - Latency of Receiving Data
	Data	<ul style="list-style-type: none"> - Visibility - Data Visualization - Data Storing - Data Processing and Data Sharing 	
	communicati	<ul style="list-style-type: none"> - Interfaces - APIs - Supporting for Heterogeneous Devices 	

Figure 6.4: Framework to Test IoT Platforms [30]

Architecture As introduced in the last section, the architecture can differ between IoT platforms. Platforms may be centralized, semi- or fully-decentralized. However, additionally also the low-level architecture may differ in means of different layers or devices. [16] tried to capture these differences by introducing a reference architecture model which is based on state-of-the-art IoT platforms which is shown in figure 6.5. They further compared Open-Source as well as commercial IoT platforms and found that all of them can be mapped to their model. However, it came apparent that the terminology in different systems may vary heavily which demonstrates again the heterogeneity in the IoT world.

Pricing Besides the functionality, the economic costs of IoT platforms are crucial for companies considering using them. During project planning, calculating the costs of different platforms is important as this may be the criteria which finally causes the selection or exclusion of an IoT platform. However, caused by the big differences in functionality and architecture, calculating final costs of an IoT platform in order to compare platforms with each other, is very difficult and requires a lot of effort. While some modules of IoT platforms are included in the pricing, other may cause additional costs when using them. Furthermore, every vendor has an individual strategy which modules are included in the standard pricing or for which modules extra costs arise [14, 4, 27].

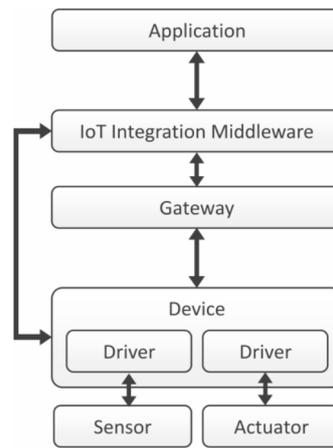


Figure 6.5: Reference Architecture for IoT Platforms [30]

Additionally, the pricing strategies of different IoT platforms vary from each other [14, 4, 27]. Whereas some price calculations are based solely on the data usage, other solutions are based on number of messages between devices or provide "flat-rate" solutions which cover all parts of the system.

Furthermore, the technology used by the devices themselves may introduce some discrepancy between the costs of IoT platforms. The selection of a communication protocol may lead to changes in the number and size of messages. For example the MQTT protocol first needs to establish a connection and then exchanges usually rather small messages but quite frequently. Communication with an HTTP client always causes a request and response which are may be rather large in comparison. Thus, this also may influence the actual accountant costs by the platform vendors if they include it in their pricing model. Hence, comparing and calculating prices before the actual usage of these services may become a difficult and complex procedure.

6.3.3 Compare IoT Platforms Qualitatively

In this section, we use the beforehand established theoretical aspects of IoT platforms and try to compare three platforms from big software companies first qualitatively and in the next section by calculating the costs for two different example use cases. For this reason, we chose the IoT platforms of Amazon, Google and Microsoft [14, 4, 27]. We shortly introduce their functionality and context before explaining their pricing model for the IoT platforms.

6.3.3.1 Amazon Web Service - IoT Core

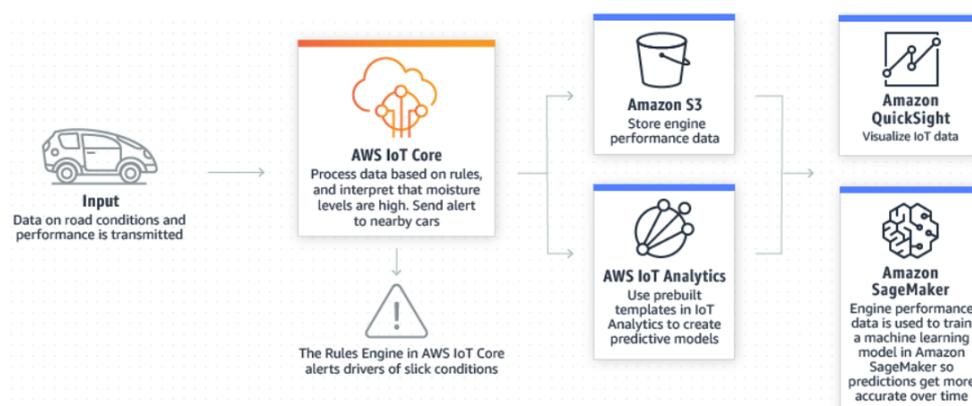


Figure 6.6: Example Configuration with AWS IoT Core [4]

Amazon Web Services (AWS) are the web services provided by Amazon. AWS IoT and especially AWS IoT Core provide the functionalities of an IoT platform which were introduced in the last section. An example configuration is shown in Figure 6.6.

AWS IoT allows the connection of a huge number of devices and can handle the messages of these devices [4]. It is fully based in the cloud and embedded tightly in the ecosystem of the web services provided by Amazon. For this reason, it offers a variety of interfaces to other cloud services, for example to analytic services or to Amazon SageMaker to train machine-learning based prediction. AWS IoT Core further provides a software development kit (SDK) which can be installed on IoT devices and eases the connection to the AWS IoT core. With this SDK installed, devices can communicate with AWS IoT core either over HTTP, MQTT or standard websocket protocols.

Pricing Strategy The total price of a system using AWS IoT Core can be split up into four subcategories [4]. Firstly, the total duration of the connectivity of the individual devices to the platform is measured. Furthermore, the number of messages between devices and the platforms billed (requests and responses). One billed message cannot exceed 5KB, otherwise this messages is billed as two messages. For example, one 8KB message is billed as two messages. The next cost factor is the number of rules triggered by messages and the number of according triggered actions. Lastly, costs arise from updating operations which access the device shadow or the device registry. The Device Shadows saves the current state of the device whereas the device registry allows naming and managing the devices connected to AWS IoT Core [4].

Hence, AWS IoT Core uses usage based pricing which is calculated by the actual traffic frequency between the IoT devices and the platform together with the number of actions which are triggered by this traffic.

6.3.3.2 Google - IoT Core

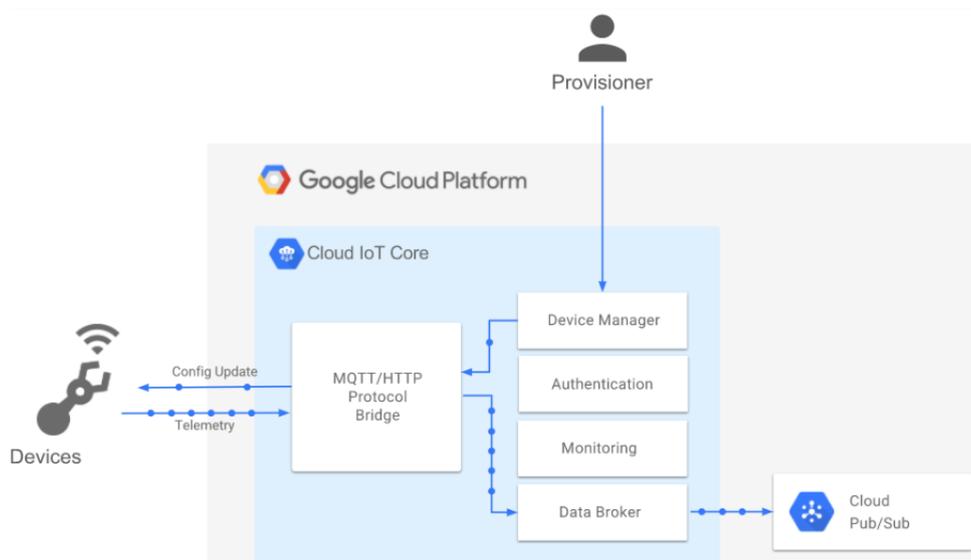


Figure 6.7: Schematic Configuration of Google IoT Core system [14]

Google IOT Core claims to be a "A fully managed service to easily and securely connect, manage, and ingest data from globally dispersed devices" [14]. Similarly to AWS IoT Core, Google IoT Core is embedded in the Google Cloud ecosystem too. This again gives the big advantage of easy connection to other cloud services of Google.

Google IoT provides the possibility of connecting, managing and delegating actions to other cloud services as shown in Figure 6.7. It allows two types of protocols to connect

to devices - HTTP and MQTT - and further handles the authentication of these devices. Besides the Google IoT Core itself, the Google Cloud Pub/Sub service (see Figure 6.7) is an important part of the IoT ecosystem as it exchanges data between the separate cloud services of Google.

Pricing The pricing in Google IoT Core is also usage based. However, instead of billing the number of messages, the price is based on the total data volume exchanged. Hence, only the number of bytes exchanged is eventually billed to the customer. However, on restriction is that the minimal message size lies at 1024 bytes. Messages below this threshold are regarded also as 1024 bytes messages. When intending to connect Google IoT Core to other service by the Cloud Pub/Sub service additional costs arise based on the number of actions triggered.

6.3.3.3 Microsoft Azure IoT Hub

Microsoft offers an IoT Platform too called "Azure IoT Hub" [27]. The IoT Hub provides similar features as the two platforms from Amazon and Google since Microsoft Azure is not only an IoT platform but also a cloud service provider too. The IoT Hub provides means to securely connect and communicate to IoT devices. The common communication protocols - HTTP, MQTT and via websockets - are offered. Furthermore, the device management is centralized on this platform and allows to connect to other cloud services without much effort. A possible integration of the Azure IoT Hub into other cloud services is shown by Figure 6.8.

Pricing The pricing strategy of the Azure IoT Hub differs from the one of Google and Amazon. It is still based on actual usage and on the number of messages exchanged between the devices and the platform [27]. However, Azure provides different types of tiers, each one with an upper limit of messages which can be sent in this tier. Furthermore, two kinds of tiers are offered: basic and standard tier. Whereas the basic tier only allows one way communication from a device to the hub, the standard tier provides communication in both directions. Lastly, the messages are calculated in 4KB steps, in which for example a 6KB messages is billed as two messages.

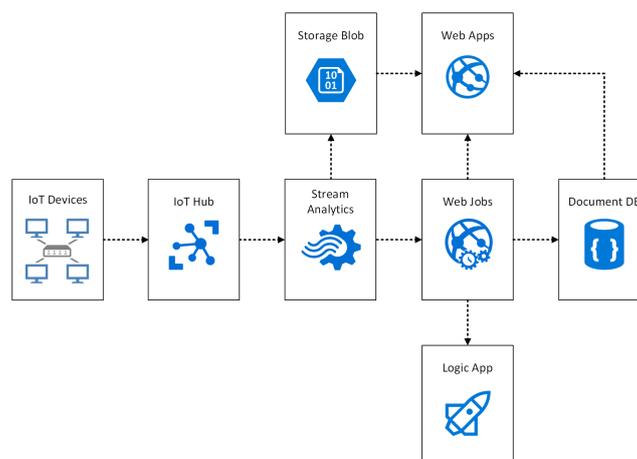


Figure 6.8: Schema of the Azure IoT Ecosystem [27]

6.3.4 Compare IoT Platforms by Use Cases

In this section, we create two different use cases and apply the pricing models from the IoT platforms introduced in the last section. We calculate the costs which arise from the

communication traffic between the platform and the devices per month and compare them with each other. The result should visualize how the different pricing strategies come into play and demonstrate their benefits and downsides.

6.3.4.1 Use Case 1

The first example use case should imitate a large scale IoT system with big number of IoT devices. Furthermore, these devices frequently communicate over HTTP with their respective platform. However, the messages and the responses by the platform do not consist of much data. This example should represent a IoT system in which small and simple sensors send continuously data to a platform. The use case consists of following specifications:

- 50'000 devices are constantly connected to the IoT platform
- Each device sends in average one message per minute and every messages gets a response by the platform
- The devices are connected over HTTP to the IoT platform
- A single message (and response) does not exceed one KB

AWS IoT Core Regarding this use case, two factors of the pricing system of AWS IoT Core come into play: the connectivity and the number of messages. Firstly, as all devices are constantly connected to the platform, 36 million hours of connection are billed. Secondly, as each device sends 1440 messages per day and triggers the same amount of responses, 4.32 billion of messages are exchanged per month in total. With costs of 0.08\$ per a million minutes of connectivity and 1\$ per million messages (.8\$ after the first billion messages), we come to a total of 3829\$ per month [4].

Google IoT Core As Google IoT Core does not count the messages, but the total data volume, we first have to calculate the total data exchanged by the use case. The use case specification leads to a total of 4.32 TB of data traffic per month. The costs per MB are structured into levels depending on the data volume already exchanged. The first 250 MB are free, up to 250 GB a MB is billed with 0.0045\$ and up to 5 TB one MB costs only 0.002\$. Hence, this leads to a total cost of 9263.875\$ per month.

Azure IoT Hub Lastly, the costs for the Azure IoT Hub are calculated. As the use case requires bi-directional communication, we must use the standard tier [27]. As the editions of the standard tier differ by the total number of messages per day, we first need this metric. From the use case, an average of 72 million of messages are sent per day which leads to a total of 144 million messages exchanged daily in total. Since it is possible to run fewer small tiers instead of only running the biggest tier, this traffic volume could also be managed by 24 S2 tiers (max. 6 million messages per day) $\tilde{\text{A}}$ 250\$. However, this is more expensive than one S3 tier (max. 300 million messages) for 2500\$.

6.3.4.2 Use Case 2

This use case symbolizes a system with only a few devices, which however send big messages. Again each device sends one message per minute. The following specification apply:

- 500 devices are constantly connected to the IoT platform

- Each device sends in average a message per minute and every messages gets a response by the platform
- The devices are connected over HTTP to the IoT platform
- A single message consists of 15KB, but the response makes up only 1KB

AWS IoT Core In difference with the first example, every messages causes three messages to be billed as the maximal size for one message is 4 KB [4]. This use case leads in total to 360'000 hours of connectivity and the exchange of 86.4 million billable messages. In total this results in only 108\$ per month in total.

Google IoT Core By considering the use case specification, the total data volume exchanged per month results in 345.6GB. This leads to a monthly bill of as much as 1314.58\$ which is more than ten times higher than the costs of AWS IoT Core [14].

Azure IoT Core Again, the maximal number of daily messages must be calculated in order to evaluate the correct standard tier edition. Additionally, the message size steps are only 4 KB, hence each message from the device is billed as four messages and one message for the response [27]. The number of messages exchanged daily lies at 3.6 million billable messages. A hub with a S1 tier costs 25\$ and has a maximum of 400'000 messages per day. The costs of one S2 hub (6 million messages daily) lie at 250\$. Hence, choosing nine S1 hubs would be slightly cheaper, however one additional device would lead to the necessity of an additional hub. Thus, choosing one S2 hub seems as the better idea when keeping the future in mind and therefore the monthly costs total at 250\$.

6.4 Pricing-Models

6.4.1 Economic Concepts Based Pricing

This section gives a short introduction to the theoretical foundations of classical economic concepts which will later be applied to various problems of IoT.

6.4.1.1 Cost Pricing

In cost-based pricing one sets the price of a product resp. of a service - as the name suggests - based on the cost of producing said product resp. service. Thereby the costs are made up of the variable costs, such as labor costs, and fixed costs, e.g. infrastructure costs. Once the cost has been determined one adds a percentage of the costs (called markup) as the desired profit.

The formula is given by

$$p = C * (1 + m) \quad (6.1)$$

where p is the profit, C the total costs and m the markup.

One of this strategy's strengths lies in its simplicity and reliance only on costs, making it very easy to model. Its downside is that it doesn't take into account external market factors such as the demand of buyers. Additionally such strategies can easily be copied by one's competitors [22].

6.4.1.2 Consumer Perceived Value

In consumer perceived value pricing the price isn't calculated according to the costs of producing a commodity but rather on what the buyer perceives to be its value. The price is therefore dependent on a set of value drivers which represent value to the buyer. As such it can be expressed as a function with different value driver inputs:

$$p_v = f(v_e, v_p, v_s, v_m, v_c) \quad (6.2)$$

In this equation are listed five value drivers identified as having a strong influence in the sensing data market [17]. They are the buyer's perception of the costs of creating the product / service (economic value v_e), the buyer's perception of the satisfaction gained from consuming a product / service (performance value v_p), the buyer's perception of the reputation or credibility of the seller (supplier value v_s), the buyer's psychological motivation for buying a product / service (v_m) and the situation context v_c [22].

The consumer-based value pricing is a trade-off between the perceived utility from gaining the information and the perceived price of acquiring said information. While it does take into account the demand of buyers it neglects market competition as well as the incomplete information conditions [22].

6.4.1.3 Supply and Demand Model

In a supply and demand market multiple sellers and buyers compete for a product, service or commodity. Assuming the buying price P can be expressed by a linear decreasing function dependent on the demand quantity Q_d and the selling price by a linear increasing function based on the supply quantity Q_s there exists an intersection where the supply and demand quantity as well as the prices match, called the market equilibrium. That quantity Q^* and price P^* are also called the clearing quantity resp. price [22].

6.4.1.4 Smart Data Pricing (SDP)

Smart data pricing is a pricing scheme where the prices are adjusted based on different factors such as time, location or activity. They aim to reduce network congestion and to avoid peak demands caused by too many users consuming scarce resources at the same time.

There are a significant quantity of different methods of smart data pricing. The two most relevant for IoT are time-dependent pricing, where the price changes over time, and usage-based pricing, where users are charged based on usage. These methods help reduce peak demand and fill up valley periods in the case of time-dependent pricing, or help create fair and efficient resource usage in the case of usage-based pricing. These schemes can lead to a situation beneficial to both providers and users [22].

6.4.1.5 Option Pricing

In option pricing a buyer can acquire an option, i.e. the right to buy or sell a particular asset at a price that is fixed at a certain date or at any time [9]. In exchange for that right the buyer must pay the seller a premium. The price of such an option is often determined using the Black-Scholes model. It models the price based on different factors such as the value lost during the duration of an option, the time to maturity of an option, the risk free rate of return and more [22].

6.4.2 auction-based Pricing

Even though the auction-based pricing is part of the game theory it is introduced in a separate subsection in this work as there are many variations of the auction-based pricing. An auction is a process where bidders, which are buyers who would like to buy an item, submit their bids to a seller who would like to sell an item. Two different prices are defined in an auction; the asking price which indicates how much money the seller wants for the item; and the bidding price which indicates the bid of the buyer. In the following we are going to introduce several different auction-based pricing models [22].

6.4.2.1 Traditional Auction Methods

To describe the differences between the auction-based pricing-models used in IoT-systems and the traditional auction methods, we first have to introduce the traditional auction methods. In the following important traditional auction methods are listed [22].

English Auction In the English auction a floor price is set, which defines the minimum price for the asset. Afterwards, the bidders, or buyers submit their bids which are higher than the defined floor price. In the end the asset is sold to the buyer who submitted the highest bid [31].

Dutch Auction The goal in the Dutch auction is to find the highest price for which all items, that are sold in the auction, can be sold. Therefore, the price to which all items can be sold is accepted. For example, two items are sold, the highest bid is four units for one item and the second highest bid is three units for one item, then the items are sold for three units each [31].

Swedish Auction In the Swedish auction the seller is not obligated to accept the highest bid. Any bid may be accepted or declined [31].

Japanese Auction In the Japanese auction the price of the item, that is to sell, increases continuously after a certain amount of time. Buyers indicate when they do not want to buy the item anymore. The last buyer who is interested in buying the item wins the auction [31].

6.4.2.2 Auctions Methods

An auction can either be based on the seller's or on the buyer's side, depending on the performed auction method. The forward auction is based on the seller's side, whereas the reverse and the double auction are based on the buyer's side [22].

Forward Auction In the forward auction the buyers bid for an item. In the end the highest price wins the auction [22].

Reverse Auction In the reverse auction the sellers with the lowest ask wins the auction. Sellers submit their asks for an item to an auctioneer until no lower ask can be submitted. If the lowest ask can be accepted by the buyer, the ask price is chosen as the price for the item [22].

Double Auction The double auction reflects the supply and demand model and is similar to the Dutch auction. Asks and bids are submitted to the auctioneer by the sellers and buyers respectively. Subsequently the auctioneer defines a price which is higher than the asking price from the sellers and lower than the bidding price from the buyers. The clearing price p is usually set as $(p_i + a_j)/2$, where p_i and a_j are bid and ask from the i th bidder and the j th seller. With the double auction an competitive equilibrium can be achieved [22].

6.4.2.3 Sealed-Bid Auction Methods

Buyers submit their bids simultaneously in the sealed-bid auction. Other than in traditional auction methods a single buyer can not see the bids of its competitors, which ensures that the bids are not influenced by bids of the competitors. It is possible to combine the sealed-bid auction methods with the auction methods described in the previous subsection. Nevertheless, it follows the forward auction method in its original form. In the following two types of sealed-bid auctions are presented [22]:

Kth-Price Sealed-Bid Auction In the k th-price sealed-bid auction the k th highest price wins the auction. The two most famous k th-price sealed-bid auctions are first-price and second-price sealed-bid auctions, where the second-price sealed-bid auction is known as Vickrey-auction. In the first-price sealed-bid auction the buyer who submitted the highest bid wins the auction and has to pay the bidding price. The buyer who submitted the highest bid wins the second-price sealed-bid auction as well but in this auction method the price is determined by the second highest bid. Due to this method the winner can not influence the price that is paid, therefore there is no motivation for a competitor to misrepresent its the submitted bid [22].

Vickrey-Clarke-Groves (VCG) Auction Following the VCG auction every sale of an item leads to a loss of the social value which has to be paid by the buyer of this item. Multiple items are sold in an VCG auction. In case a buyer wins the auction for one item (by submitting the highest price) the price, that has to be paid, is calculated by subtracting the social value of the remaining items from the social value of all items without considering the values from the winner of the auction [22].

6.4.2.4 Combinatorial Auction Method

In the combinatorial auction method bids for a whole bundle of items are submitted to the auctioneer. Subsequently the auctioneer finds the optimal allocation as well as the winner of the auction based on the bids and the item allocation from the sellers. Benefits from this auction model are the utility maximization for buyers as well as the revenue maximization for the sellers. Nevertheless, determining the winner of the combinatorial auction is a NP-complete problem. Therefore, algorithms which deliver approximated solutions are usually used [22].

6.4.3 Game Theory

Game theory studies situations where a multitude of players must make decisions that possibly have an effect on the interests of the other players. It too can be used to solve a variety of problems faced with in IoT. To understand the following models some definitions are needed. A player is what a participant in a game is called. They act rationally and thus always try to maximize their payoff, which represents how desirable an outcome is

to a player. Players usually have a set of strategies, which are complete plans of action or instructions a player can follow [22].

6.4.3.1 Non-Cooperative Game

A non-cooperative game is a simple game in which players can't form pacts or make agreements. One proposed solution to it is a so called Nash equilibrium. In such an equilibrium no player can improve their own payoff by switching strategies while the other players remain with their strategies. A non-cooperative game may also have no or multiple Nash equilibria. Furthermore it assumes that all players know the opponents' strategies and that they are revealed at the same time. This assumption doesn't however always hold in real markets in which case one can't calculate one's Nash equilibrium [22] [13].

6.4.3.2 Stackelberg Game

In a Stackelberg game a player can wait for their opponent to announce their strategy and based on that find their own optimal strategy. The player that goes first is called the leader while the other is the follower. In the case of multiple players this happens sequentially. It was found that that if the follower chooses their optimal strategy depending on the situation, the solution is an optimal one for the leader as well and they receive a payoff function that is at least as good as that of a Nash equilibrium. The reason for this being that the leader's choice of strategy imposes itself on the strategies of the followers and thus facilitates itself [22].

6.4.3.3 Bargaining Game

In a bargaining game two or more players must come to a consensus on how to distribute an object. Consider a simplified situation with one seller and one buyer. Both try to find a price p^* which maximizes their projected profit. The price they submit is considered their strategy. If the buyer's price p_b^* is greater or equal to the seller's price p_s^* a bargain is struck. The price the data is sold at is

$$p = kp_b^* + (1 - k)p_s^* \quad (6.3)$$

with $0 \leq k \leq 1$. Here too there is a Nash bargaining equilibrium which means that no player can improve their profit by switching strategies while the other players don't switch theirs [22].

6.4.4 Utility Maximization and Knapsack Problem

In this subsection we are going to introduce the utility maximization and the knapsack problem.

6.4.4.1 Utility Maximization

Buyers value different items differently. The values can be modelled by using the buyer's utility function which is often used in microeconomics [25]. As buyers are constrained by their resource one assumes that they act rational and try to get the most value for their resource. As the preferences of each buyer differ depending on the item or the service a buyer has to choose between alternative goods to maximize its utility [1, 22].

As long as an item provides more utility per resource unit than other items the buyer will purchase more of this item rather than buying another one [1].

6.4.4.2 Multi-Objective Knapsack Problem

One of the most famous types applications of the utility maximization is the knapsack problem. In the knapsack problem every item has an assigned tuple (v_i, w_i) , where v_i represents the value or the utility of the item and w_i the corresponding weight. The portability of the knapsack is defined by a constant M . The portability defines the constraint and can be seen as finite resource in this problem. The goal is to pack as much value as possible into the knapsack without exceeding M . As there are multiple items and the knapsack capacity is restricted by M , the item with the highest value will not necessarily be packed into the knapsack because of its weight. It is possible that multiple items with lower value but also lower weight lead to a higher value. The computation of the result for the knapsack problem has high costs as there are many combinations of items that can be packed together into the knapsack [22].

6.4.5 Combination of Pricing-Models

In this subsection we are going to combine pricing-model, in order to propose new conceptual pricing-model with benefits from the underlying ones.

6.4.5.1 First-Price Sealed-Bid Auction and SDP

The first-price sealed-bid auction and the SDP pricing-model could be used to fast forward urgent messages to the server or to other devices. In such a model devices which communicate with a server would act as buyers and the server itself as a seller. In case of machine to machine communication where the devices communicate with each other the devices would have to act as both buyer and seller. Normal messages would have a low bid, where urgent messages would have a high bid so that all devices know that this message is to prioritize and has to be forwarded with the shortest possible delay so that this message reaches its destination even in times the network is congested.

6.4.5.2 Stackelberg Game and Cost Pricing

A Stackelberg game can be used to find the optimal strategy based on the strategy of the participants who have gone before [22]. It can be used for example in cases where a data packet needs to be split up by a relay node, and since the relay node may determine the splitting strategy a Stackelberg game is useful in finding the optimal relay node. It could then be combined with other pricing models, such as cost based pricing to determine the price at which the data should be sold, based on the cost of gathering, splitting and relaying, and based on the profit that one would like to achieve.

6.5 Application of Economic and Pricing-Models for IoT-Systems

In this section we are going to introduce problems where the application of pricing-models can be used to solve these in a more efficient way.

6.5.1 Pricing-Models for Data Exchange and Topology Formation

This section concerns itself with how a IoT system's network is created or how it reacts to fluctuations in its topology or routes. Furthermore it addresses problems related to

how data is exchanged between IoT components. Several different pricing and economic models are then applied to solve these problems.

6.5.1.1 Data-Aggregation and Routing

In many IoT systems there are a multitude of sensors gathering data. This data must then be aggregated before relaying it further to the system. This also serves to remove redundant data which additionally reduces the energy required to relay the data. Since the data may be transferred directly or through relay nodes, one must also take into account the problem of finding the routes which minimize latency and, since these usually shorter routes are also more energy efficient, consume less energy [22].

One given fact in typical WSNs is that sensors and wireless links are constrained in regard to their energy and capacity [3]. Thus one major goal of such systems is improving the efficiency of the energy usage which also results in a greater overall network lifespan. This can however be at odds with another requirement, to provide the required Quality of Service (QoS). This latter one includes data quality, latency and packet lifetime to name a few. Different economic and pricing models can be used to address this conflict [22].

Using value-based pricing a system can adjust the resource usage based on what the buyer i.e. the requester perceives the value to be. The sink node as the seller therefore tries to maximize the buyer's utility and sets the price according to the buyer's expectation about the QoS [22]. Simulations done by [2] showed that this system outperformed mobile ad-hoc schemes in terms of lifetime, delay, scalability and more.

A different approach is to use a sealed-bid reverse auction. One possibility is to use it to maximize the energy efficiency and network lifetime. In this scenario the WSN fusion center acts as the buyer and the sensors as the sellers. The sensors' asks represent their remaining energy, i.e. a lot of remaining energy results in a low ask. These asks are privately and concurrently submitted to the fusion center which then selects the optimal subset of sensors to maximize its own utility function. Simulations of this approach did indeed improve the network lifetime of the system compared to ones using other strategies [22].

A sealed-bid reverse auction might be used to minimize latency and maximize the network throughput, i.e. to solve the routing problem as well. In that case a router acts as the buyer which must choose a neighbor (the sellers) to relay the data. The neighbors' asks consist of a path price which also depends on the projected relay time to the destination. The router then selects the neighbor which has the lowest path price, leading to a system which decreases its overall delay [22].

The problems discussed in this section also affect another kind of IoT system, that of participatory sensing and crowdsensing networks. The main issue here is to incentivize users to collect and share data, since this costs them energy and bandwidth. For its solution one can again look to economic models.

Similarly to above one can use a sealed-bid reverse auction. Here it is used to encourage users, who represent sellers, to share the data they sensed (gathered). The server (buyer) broadcasts the description of the sensing task upon which interested users can accept and carry out the task. They then submit their asks which include the price and gathered data, from which the server selects a subset. For this scheme to work there needs to be a large enough number of participants so that the prices can be kept competitive. Thus this model is often paired with other schemes. One such scheme is the Virtual Participation Credit (VPC), which encourages losers of an auction to keep participating in the network. Using VPC a loser of a previous round has a better chance of winning a future round. Another scheme is that of ReCruiting (RC), which is often paired with the sealed-bid

reverse auction and VPC. It addresses the problem of winners of an auction not being satisfied with their reward, making it more likely that quit the network after the next round. Using RC, dropped out users are sent the maximum reward awarded to a winner in a previous round in an attempt to entice them to rejoin [22].

The strategy elaborated upon above however fails to take into account other factors when determining the price, namely QoS factors such as data quality. Therefore a multi-attribute sealed-bid reverse auction can be used which incorporates weighted attributes such as data quality or the location accuracy into the asks. Simulations of this method done by [20] showed that it improved the utility to the service provider since they have more choices to choose the optimal seller from. Similar models have been created using coupon rewards rather than monetary rewards. That way a system can reward users with a good history of participation and reliable data, e.g. a high credibility with more coupons. Or it can reward coupons according to user preferences and thus keep the reward value up [22].

A different approach is a combination of two models: That of an all-pay forward auction (an all-play auction is one in which all buyers including the losers must pay the bid) and a posted price mechanism. The system (buyer) divides up a given task with a deadline into smaller periods with their own corresponding budget. Phone users (sellers) can then find the periods best suited to them and submit their sensing data. Using the all-pay auction the system chooses a subset of users as optimal winners. For the following period it then sets a threshold price so that when a new seller arrives according to the posted price mechanism the seller can compare the cost for the task execution with the threshold and accept it if the former is smaller than the latter. One benefit of this system is that it increased participation in networks with increased budgets [22].

A simpler approach is to once again use value-based pricing. Using it the system as the buyer determines the compensation for the sellers based on different value drivers which increase the utility to the system. In this approach there are two different kinds of sellers, the users that collect and share data and helpers: Users who share their cellular connection to help relay the data. One thing of note is that in the value of the collected data degrades over time, as the data becomes more and more outdated. Thus once some particular data has been forwarded to the service provider a payout is generated based on the initial value and according to the freshness of it. That payout is then given to both helpers and data sources [22].

Another economic concept, that of demand and supply, was used in [26] to find the equilibrium in which a fusion center's demand matches the data bits supplied by sensors. Unlike in the traditional demand and supply model however in this method the buyer (the fusion center) adjusts the price rather than the seller. The seller broadcasts a vector of prices it is willing to pay for sensing data bits. Based on this the buyers (the sensors) calculate a demand vector that maximizes their profits. This happens continuously and iteratively until the equilibrium is found and supply and demand line up [22].

Using a multi-objective Knapsack problem one can model a system with a smaller budget and thus fewer selected users while still guaranteeing the QoS. Thus it's a optimization problem with the goal of low payment and high data quality, with the limited number of participants being the weight constraint. The service provider represents the buyer and the participants the sellers. In the eventuality of the amount of selected participants not being equal to the weight constraint a crowding distance criterion can be used. It measures the distance from participants to one another and using it one can remove or add new participants. While this scheme doesn't have the smallest payments it does outperform other models in terms of the ratio of payments to QoS [22].

6.5.1.2 Opportunistic Transmission and Neighbour Discovery

This part concerns itself with how data gathered in crowdsensing and participatory sensing is sent to server. Since cellular networks can be both expensive and limited in regards to capacity they aren't economically feasible. Thus one can use economic models to minimize transmission costs and maximize energy efficiency.

One such approach uses cost-based pricing in order to minimize the global system cost. In it a source phone user ,i.e. the seller, creates a list of available relay neighbors reachable in one hop including the associated relay cost. This table is therefore called a one-hop neighbor table. The source then chooses the neighbor that maximizes their profit, i.e. has the smallest cost. Since short-range communication such as WiFi is usually cheaper than long range communication this method leads to systems that prefer the former [22].

6.5.1.3 Relay Selection

Collected data from sensors must be relayed to its destination, e.g. a sink node. This often happens through relay nodes. The question of which routes to take is the topic of this section. There are different possibilities such as the least energy consumption route or the quickest route. Since relay nodes can however try to incentivize their own selection. They can therefore be thought of as selfish and rational actors, which leads to interesting applications of pricing strategies [22].

One such strategy is a first-price sealed-bid reverse auction. It is applied here with the goal of minimizing the energy consumption and as a result thereof to increase the network lifetime. It models the source node as the buyer who purchases the data relaying service from several relay nodes in the neighborhood, i.e. the sellers. Similarly to the problem discussed in the previous section the source node builds a table containing all neighbors and their link quality. The link quality in turn depends on the residual energy of the node and hop count from it to the destination (i.e. the sink node). Here the model differs a bit from a typical auction in that both sellers and buyers submit their asks at the same time. A deal between a source and a relay node is struck if the asking price from a relay node is smaller than that of the source node. Should this be the case for multiple relay nodes the one with the lowest ask is chosen [22].

A related approach uses a Dutch reverse auction. It aims at creating a highly reliable system, i.e. one with minimal packet delivery failure. Here again the source node is the buyer and the relay node the sellers. The source provides the relay nodes with forwarding regions which have different priorities as the asking price. Naturally regions closer to the destination have a higher priority. A seller can then accept the offer price and relay the data, with the first node to accept the task being awarded it [22].

There may however be rational interactions between nodes. One example would be a relay node that rather than relaying an entire packet only agrees to relay part of it. Since the packet splitting strategy now depends on the independent decisions of the relay node rather than on that of the source node a Stackelberg game can be used. The relay nodes represent the leaders that choose their own pricing strategy, i.e. set the relay prices. The source node as the follower can then find its best splitting strategy corresponding to the relay prices [22]. As a matter of fact [18] modeled the system as a two-stage Stackelberg game with the relay nodes being able to re-evaluate their prices at this stage. They found that it resulted in a more balanced energy dissipation of relay nodes.

6.5.1.4 Congestion Management

Congestion in this context is the situation in which multiple packets arrive at a node at the same time. In typical WSNs this may happen when multiple source nodes transmit data to the sink node, resulting in congestion in the relay nodes. Congestion leads to network delays which also makes the information less valuable. Here too pricing strategies can be used to solve the problem.

Several different second-price sealed-bid auctions have been proposed as a solution in order to prioritize which data is to be forwarded. In the first one the node that is congested represents the seller and the packets act as buyers that compete for the current slot for transmission. The winner of which gets only the current transmission slot rather than all the node's resources, hence the second-price auction. The value of a packet is determined by the utility of the loss of information and the node selects the packets with the highest one as the winner. To ensure a minimum of the total network utility loss the losers of an auction receive a fund for the next rounds proportional to the information utility loss sustained due to the delay of not being chosen. One problem with this approach is that congested nodes only take into account their local information and not information from other nodes. Thus the loss of information utility can be calculated falsely. Therefore the aforementioned strategy can be extended with a traveling auction scheme. In it nodes in the network share information such as packet creation time and delay [22].

Alternatively value-based pricing can be used. One method is to use it to determine which packets to drop in case of congestion. The sink node acts as buyer in this case sets the price for packets from the seller (the sensors). It sets these prices based on the predicted value of packets, namely packets with more important information. Since this can induce unfairness for the selection of packets the scheme is usually paired with the Jain's fairness index. It makes it possible to calculate the likelihood that a packet is accepted by the sink node and can be applied to create a more fairly distributed opportunity for selection [22].

6.5.2 Application for Resource and Task Allocation

This section describes problems which arise from components of WSNs having limited resources such as energy, bandwidth, processing power, storage capacity and more. Thus these resources have to be allocated and used in an optimized fashion, which can be modeled by various economic and pricing models. Furthermore the section describes schemes to allocate tasks to sensors in the network in a way that minimizes delay and achieves a fair energy distribution.

6.5.2.1 Resource Allocation

As previously mentioned a system has limited resources. Furthermore how many available resources a system has over time isn't constant but rather fluctuates. This can additionally lead to situations where the supply of resources doesn't match the demand. By creating an artificial market between the sensors using economic and pricing models they can dynamically trade resources with each other [22].

Since sensor data can often be complementary such as data from multiple sensors which together increase the accuracy of the measured item it makes sense to bundle this data and sell it together. Thus a combinatorial auction can be used. The requesters act as the buyers which submit a task as their bid. The sellers of the system, the sensors and transmission channels, act as the sellers and it is a combination of them that will eventually solve the task. The sensor manager as the auctioneer, once receiving the bid, iterates through all possible allocations of the task. Thus the winner of this auction is in fact an

optimal resource allocation. Naturally one can use different algorithms to find determine the optimal allocation. One such algorithm is CABOB which aims to maximize the revenue of the auctioneer while allocating each resource to at maximum one bidder [22].

As mentioned previously, the resources demanded by a component might not always match the supply. Thus certain sensors might have more resources than needed for a task while others have too little. By facilitating the exchange of such resources between sensors the global performance can be maximized. In this case a double-sided auction can be used, where the tasks are both buyers and sellers of resources. The asks respectively the bids form a demand and supply curve. Now sensors continuously and at any time submit asks and bids for resources to the auctioneer. In the Continuous Double Auction Parameter Selection (CDAPS) the auctioneer chooses the bid with lowest price and the ask with the highest price - so long as the asking price is higher than the buying price - as a valid transaction with the clearing price being the average of the two. Then the selected buyer receives the resources and the seller the payment. This procedure is repeated to create pairs for every buyer and seller [22].

6.5.2.2 Task Allocation

Task allocation concerns itself with how tasks should dynamically be assigned to particular sensors. They too serve to optimize resource usage and can be modeled by a variety of pricing and economic models [22].

Several first-price sealed-bid auctions have been proposed. One such model, called a Just-In-Time market, sees the sensors as sellers and any system relying on collected data as buyers. An auctioneer is responsible for assigning tasks [32]. The buyers submit their requested tasks including prices and each sensor determines its own utility according to the resources it has. Tasks are then assigned if the task can be executed. Should a predetermined deadline be reached, the auction can be aborted to make the system more adaptable to changes [22].

A related model has the sensors act as sellers (and auctioneer) and sell a task should it not be able to perform it itself. Other nearby sensors act as buyers and can acquire the task if they can complete it and have enough energy. In this model there is a price threshold that the buying sensors' bids have to overbid. Additionally the prices are evaluated after a certain time has passed. The model is supposed to maintain an energy balance, as the prices reflect the energy of sensors [22].

A double auction method can be used as it is more fair and efficient in regards to resource allocation [34]. Whereas in single-sided auctions the buyer or seller holds the power over a transaction, in a double auction it is distributed over both seller and buyer and the model follows the supply and demand model. Thus a method proposed for task allocation models sensing tasks from users as the buyers of the system and smartphone users as sellers. The commodity in this case are sensing times. The tasks (buyers) submit price and demand of sensing tasks and smartphones submit their prices and supply. A server acting as auctioneer searches for the first smartphone that has enough sensing time for the tasks with the highest price, so long that no profit is lost [22].

6.5.3 Pricing-Models for Sensing Coverage and Target Tracking

The sensing coverage measures how well sensors observe the environment in which they are located and is a crucial part of a WSN. There are classical approaches which deliver the optimum global result of how to achieve the maximum sensing coverage and target tracking. Nevertheless those classical approaches have high computational costs and may lead to an communication overhead. Applying pricing-models on this problem provides

similar results to the optimized solution but with significantly lower computational cost. In this subsection we are going to introduce the application of pricing-models on the three main types of sensing coverage as well as on target tracking [22].

6.5.3.1 Area Coverage

The area coverage measures the covered sensing area in a sensing field. Beside static sensors which are installed in the sensing field mobile sensors can be used to cover the holes in the field. To maximize the area coverage the mobile sensors have to be placed at the optimal position [22]. In [38] the first-price sealed-bid auction is used to find the optimal position for the mobile sensors. The static sensors act as buyers where the mobile sensors act as sellers. To detect the local holes a Voronoi polygon is formed by each static sensor with respect to the position of its neighbors, as illustrated in figure 6.9. The static sensor S_1 covers the grey area which surrounds it. To avoid coverage overlapping of the static and the mobile sensors the farthest vertex of the Voronoi polygon is selected as the desired position of the mobile sensor [38].

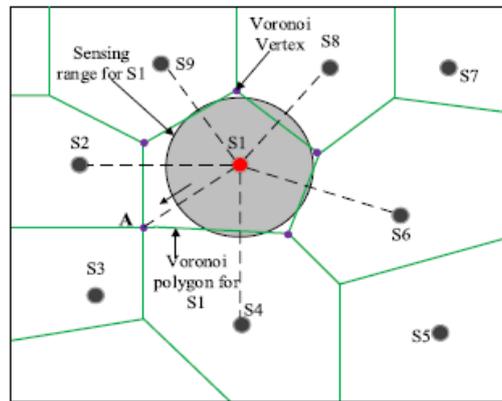


Figure 6.9: Voronoi Diagram for the Area Coverage Problem [22]

The static sensors calculate their bids according to their detected local hole. The mobile sensors simultaneously calculate their ask price which is based on the hole they cover at their current position. After the bids from the static sensors are submitted it is checked whether the highest bid is greater than the mobile sensor's asking price. If the bid is lower the mobile sensor is not moved to the new position, as it covers a larger hole at its current position. In case the highest bid is greater than the mobile sensors' price the mobile sensor is moved to the new desired position and takes the highest bid as its new price since this is the new hole covered by it [38].

This process is repeated until no price of any mobile sensor is lower than any bid of the static sensors. The process has a lower computational cost than classical approaches, nevertheless, it is very energy consuming. To save energy the proxy-based bidding protocol as it is proposed in [37] could be used. In the proxy-based bidding protocol messages are exchanged before a sensor moves to its final position, which ensures that the mobile sensor only moves when the final position is found.

6.5.3.2 Target Coverage

The target coverage deals with the problem of reaching defined interesting targets with known location with minimal energy consumption. In [28] an application of the utility maximization is proposed where the energy consumption is the price that is to pay, and the utility is described by the individual target. The sensing range is only increased to cover the target if and only if the utility of covering a certain target is higher than the

price of the energy consumption. Nevertheless, this method only works if the network knows the utility function of every sensor [22].

6.5.3.3 Barrier Coverage

The barrier coverage deals with the issue to recognize intruders that want to cross a sensing field. It is important that the sensing areas of the sensors overlap so that no holes may be misused by an intruder to cross a sensor field without being recognized. In case a gap is recognized mobile sensors have to be moved to close this gap, which is energy consuming. By applying pricing-models the energy consumption for the movement can be minimized [22].

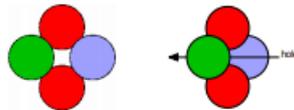


Figure 6.10: Example of a Barrier [5]

In [5] an approach for a underwater-3D-barrier, which minimizes the energy consumption, is proposed. Figure 6.10 shows such a 3D-barrier. The figure shows a hole in the barrier which has to be covered. To determine which mobile sensor should be moved to the grid point, the first-price sealed-bid auction may be used. The mobile sensors submit a bid which is inversely proportional to their distance to the grid point. The mobile sensor with the highest bidding price wins the auction and is moved to the grid point. The first-price sealed-bid auction ensures that always the nearest mobile sensor is moved, and therefore, the energy consumption is minimized [5].

6.5.3.4 Target Tracking

Pricing-models can be applied to predict future locations of a target, and therefore support the object tracking task. To support the object tracking task the authors of [8] proposed an approach which uses the first-price sealed-bid auction. When a sensor detects the target in its sensing area and there is no other leader in the sensing field it takes the leader role and acts as the seller. Subsequently the leader sensor informs other sensors in the sensing field, which respond with a bid. After receiving the bids the leader determines which sensors are selected to perform the task and send their sensing data. The number of selected sensors is predefined and the buyers which submitted the highest bids are selected. This approach does not require a sensor to know its neighbors, additionally less messages between buyers and the sellers are exchanged what leads to a reduce of the energy consumption as well as a reduce of the used storage and the used computing resources [22].

6.5.4 Pricing-Models for Improving Privacy Security

Since privacy is becoming increasingly important, it is also critically questioned in IoT-systems. To find out how much the privacy of individual users is worth pricing-models can be used. In this subsection we are going to introduce how pricing-model may be applied to evaluate the worth of private data.

6.5.4.1 Privacy Evaluation

To evaluate the privacy the authors of [10] performed an experiment with volunteers where they used the second-price sealed-bid reverse auction. In their experiment they aimed to

collect data about the current location of the participants from their mobile phones. The participants define the asking price for their locations and submit them in encrypted form to the sellers. Subsequently a set with a predefined size of the participants who submitted the lowest asking prices are selected as winners. The winners get the lowest asking price of the remaining participants, which are not included in the set, as reward. By applying this pricing-model it is ensured that participants are remunerated fairly and do not get prices which are lower than they worth their private data [22].

6.5.4.2 Privacy-Preserving Mechanisms

The authors in [39] proposed an approach for privacy preserving. Their approach is similar to the approach in [10] the asking prices are encrypted and signed by the user before they are sent to the auctioneer. Furthermore pseudonyms are used to replace the users real identity which preserves the privacy of every single user. At the end a set of users which submitted an asking price that is lower than a predefined threshold are selected to send their right data. With this approach it is ensured that users which value their private data more than the buyer is ready to pay stay anonymous [22].

6.5.5 Pricing-Models in M2M Communications

There are two types of M2M communications. The devices can either communicated directly with each other without human interaction or multiple devices can communicate with one server. Pricing-models can be applied to optimize the communication between the machines. The SDP pricing-model can be used to ensure a stable communication and at the same time reduce the congestion in a network and reduce the communication costs [22].

The authors in [24] propose an example for an online charging service for the communication between the devices and the application, which is the server at the same time. The model they provide consist of multiple machines which perform a monitoring task and the application to which the data from the monitoring task is sent. By applying the SPD pricing-model, charging the machines higher in times when the network is congested, and enable the machines to postpone normal messages to a later time, it is possible to reduce the congestion in the network and the communication costs. In case an emergency has to be reported the data package can be marked with an emergency flag. Packages with an emergency flag are forwarded to the server with the shortest possible delay. Machines which would like to send normal messages about their monitoring task will postpone the communication to a time where the network is not congested, and therefore be charged less than in time of congestion. The model ensures that emergencies are reported as fast as possible, and that a congested network is not cluttered with even more traffic, allowing the network to recover more quickly [24].

6.5.6 Applicability on Business Sectors

In this subsection we are going to find examples where pricing-models can be applied in different business sectors.

6.5.6.1 First-Price Sealed-Bid Auction and Sensing Coverage

In sectors where all the data has to be collected always one could use the first-price sealed-bid auction to faster compute the position of a mobile sensor in case on of the sensors breaks down. Example sectors for such an application could be the health care, or also the heat control in public vehicles where the break down of a sensor can impact human life.

6.5.6.2 First-Price Sealed-Bid Auction and Target Tracking

In sectors where the tracking of a target can be very useful the first-price sealed-bid auction may be used to support the target tracking task [22]. Examples for such sectors could be the law enforcement (i.e. police, or military). Examples for target tracking could be the pursuit of a fugitive, or the monitoring of former convicts.

6.5.6.3 Smart Data Pricing and M2M Communication

In the retail sector the smart data pricing model may be used to support the communication between the logistics devices and the controlling server to combat congestion and to ensure a stable communication [22]. An example could be a large company with multiple warehouses. Each of the warehouses has logistics devices that monitor the stock in the warehouse and communicate to one centralized server. In case a big change in the stock happens, the message is marked with a flag which ensures that the messages is fast forwarded to the server. During normal activities the messages about the stocks are only sent in times when the network is not congested.

6.6 Challenges and Limitations

6.6.1 Technical

IoT faces several technical challenges in the future. As IoT-systems are expected to consist of millions of devices, the scalability of these systems is a major challenge. Although such systems will be organized in hierarchical sub-domains, the number of interconnected objects and therefore also the exchange of data will increase. Therefore, scalable systems and the possibility to put processing mechanisms 'to the edge' are possible solutions [36]. In addition, even though IoT devices focus on low-energy and low-bandwidth consumption, the internet bandwidth in which the devices communicate can get saturated with data traffic quickly which would result in system-wide performance problems.

The most prominent technical issue in IoT are the security challenges. Terrible stories about hacked drug infusion pumps, cameras intensify the need for security in IoT. IoT devices often lack the computational power, operating system and storage capacity to be able to deploy security mechanisms. In terms privacy of data the encryption could erase the privacy problem. Nevertheless, the encryption of data potentially increases the message size and forces the actors in the network to actively encrypt and decrypt data when sending or receiving it. As IoT aims to provide data in real-time, the encryption and decryption delay the transmission of data which would therefore in a higher latency [36].

To update devices and check the status of each device, a larger protocol than MQTT has to be used (e.g LWM2M). Although through such a protocol the devices can be managed, the detection of faulty hardware is still expected to be hard. Moreover, the longevity of hardware is a challenge the manufacturers face, as a sensor or actuator that is placed in a network should be able to work for a long time.

Another factor that is challenging for future IoT-systems and their interoperability is the current lack of standardization. By now, many protocols and IoT-platforms are used to create systems. This variety of protocols and platforms create constraints in which independent systems face the challenge to communicate with each other while using a different protocol.

6.6.2 Business

For businesses, IoT may potentially be identified as possibility to enhance their products and services through the insights that are generated for the data. Nevertheless, an IoT-system also brings risks when implemented. If a wrong pricing model for the enhanced services or products is chosen, the selling of those may result in a loss. Since it is hard to estimate the additional value generated for a customer, the estimation of the total cost of ownership is difficult and the generated value may also be poorly estimated which could furthermore result in a loss for a company [21].

6.6.3 Social

As technology keeps advancing at a fast pace, the law often lags behind with regulatory and legal issues. This could lead to a set of problems as soon as new laws for IoT have been enforced. Another factor that impacts IoT are that the usage for IoT may change over time. Lastly, the privacy aspect of the gathered data in sensor networks is a challenge to overcome as well. How do such systems ensure that only the data is captured that people consented to? Is it the responsibility of the individual to take action if his/her privacy is violated or do these systems filter out any conflicting information?

6.7 Discussion

In this paper, we reviewed the problem of pricing models in the Internet of Things from different perspectives. In a first step, we introduced IoT platforms as they form a core part of a functional IoT system. We compared three different IoT platforms in regard to their functionality and pricing strategies and found big differences. Furthermore, comparing IoT platforms is a difficult task as they are not standardized and consist of many heterogeneous sub parts but with the help of testing frameworks this task can be simplified.

We introduced economic concepts which are used in economic everyday life. These concepts can not only be used to evaluate the prices for IoT-systems but can also be applied to problems in IoT-systems to reduce computational costs, minimize the energy consumption, reduce congestion, and evaluate the worth of sensible data. Nevertheless, most of the introduced applications are conceptual. Additionally, the calculation of the prices often ignores the costs for hardware and the infrastructure. For example in the application of the second-price sealed-bid reverse auction for the privacy preserving mechanism it is not defined how to define the threshold for the price. This is essential as the selection of the winners is based on this threshold. To conclude we can say that pricing-models can be applied well on problem in IoT-systems, but nevertheless, much more research has to be done on this subject.

Bibliography

- [1] Utility maximization model.
- [2] Ashraf E Al-Fagih, Fadi M Al-Turjman, Waleed M Alsalih, and Hossam S Hassanein. A priced public sensing framework for heterogeneous iot architectures. *IEEE Transactions on Emerging Topics in Computing*, 1(1):133–147, 2013.
- [3] Zapata AlSkaifm and Bellalta. Game theory for energy efficiency in wireless networks: Latest trends. *Journal of Network and Computer Applications*, 54:33–61, 2015.
- [4] Inc Amazon Web Services. Amazon web service: Iot core.
- [5] Stanley Barr, Benyuan Liu, and Jie Wang. Underwater sensor barriers with auction algorithms. In *2009 Proceedings of 18th International Conference on Computer Communications and Networks*, pages 1–6. IEEE, 03.08.09 - 06.08.09.
- [6] Carsten Bormann, Angelo P. Castellani, and Zach Shelby. Coap: An application protocol for billions of tiny internet nodes. *IEEE Internet Computing*, 16(2):62–67, 2012.
- [7] Chebudie, Abiy Biru & Minerva, Roberto & Rotondi, Domenico. Towards a definition of the internet of things (iot). In *IEEE Tech. Rep.*
- [8] Jianxia Chen, Chuanzhi Zang, Wei Liang, and Haibin Yu. Auction-based dynamic coalition for single target tracking in wireless sensor networks. In *2006 6th World Congress on Intelligent Control and Automation*, pages 94–98. IEEE.
- [9] John C Cox, Stephen A Ross, and Mark Rubinstein. Option pricing: A simplified approach. *Journal of Financial Economics*, 7(3):229–263, 1979.
- [10] George Danezis, Stephen Lewis, and Ross Anderson. How much is location privacy worth. In *In Proceedings of the Workshop on the Economics of Information Security Series (WEIS, 2005)*.
- [11] Amir Vahid Dastjerdi and Rajkumar Buyya. Fog computing: Helping the internet of things realize its potential. *Computer*, 49(8):112–116, 2016.
- [12] Tharam Dillon, Chen Wu, and Elizabeth Chang. Cloud computing: Issues and challenges. In Elizabeth Chang, editor, *24th IEEE International Conference on Advanced Information Networking and Applications (AINA), 2010*, pages 27–33, Piscataway, NJ, 2010. IEEE.
- [13] James W. Friedman. A non-cooperative equilibrium for supergames. *The Review of Economic Studies*, 38(1):1–12, 1971.
- [14] Google LLC. Google cloud iot core.

- [15] Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, and Marimuthu Palaniswami. Internet of things (iot): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7):1645–1660, 2013.
- [16] Jasmin Guth, Uwe Breitenbucher, Michael Falkenthal, Frank Leymann, and Lukas Reinfurt. Comparison of iot platform architectures: A field study based on a reference architecture. In *2016 Cloudification of the Internet of Things*, pages 1–6, Piscataway, NJ, 2016. IEEE.
- [17] R Harmon, H Demirkan, B Hefley, and N Auseklis. Pricing strategies for information technology services: A value-based approach. In *2009 42nd Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2009.
- [18] Eirini Karapistoli and Anastasios A Economides. Routing in wireless sensor networks: An approach using stackelberg games. In *2014 IEEE 10th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 322–327. IEEE, 2014.
- [19] Jaewoo Kim, Jaiyong Lee, Jaeho Kim, and Jaeseok Yun. M2m service platforms: Survey, issues, and enabling technologies. *IEEE Communications Surveys & Tutorials*, 16(1):61–76, 2014.
- [20] Ioannis Krontiris and Andreas Albers. Monetary incentives in participatory sensing using multi-attributive auctions. *International Journal of Parallel, Emergent and Distributed Systems*, 27(4):317–336, 2012.
- [21] In Lee and Kyoochun Lee. The internet of things (iot): Applications, investments, and challenges for enterprises. *Business Horizons*, 58(4):431–440, 2015.
- [22] Nguyen Cong Luong, Dinh Thai Hoang, Ping Wang, Dusit Niyato, Dong in Kim, and Zhu Han. Data collection and wireless communication in internet of things (iot) using economic analysis and pricing models: A survey. *IEEE Communications Surveys & Tutorials*, 18(4):2546–2590, 2016.
- [23] Somayya Madakam, R. Ramaswamy, and Siddharth Tripathi. Internet of things (iot): A literature review. *Journal of Computer and Communications*, 03(05):164–173, 2015.
- [24] Ranko Maric, Tomislav Grgic, Maja Matijasevic, and Ignac Lovrek. Online charging based on machine context for m2m communication in lte. In Mari Carmen Aguayo-Torres, Gerardo Gómez, and Javier Poncela, editors, *Wired/Wireless Internet Communications*, volume 9071 of *Lecture Notes in Computer Science*, pages 18–31. Springer International Publishing, Cham, 2015.
- [25] Andreu Mas-Colell, Michael Dennis Whinston, and Jerry R. Green. *Microeconomic theory*. Oxford Univ. Press, New York, NY, 1995.
- [26] Engin Masazade and Pramod K Varshney. A market based dynamic bit allocation scheme for target tracking in wireless sensor networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4207–4211. IEEE, 2013.
- [27] Microsoft Cooperation. Azure iot core.
- [28] Marjan Naderan, Mehdi Dehghan, and Hossein Pedram. A distributed dual-based algorithm for multi-target coverage in wireless sensor networks. In *2011 International Symposium on Computer Networks and Distributed Systems (CNDS)*, pages 204–209. IEEE, 23.02.11 - 24.02.11.

- [29] Bhumi Nakhuva and Tushar Champaneria. Study of various internet of things platforms. *International Journal of Computer Science & Engineering Survey*, 6(6):61–74, 2015.
- [30] Amirfardad Salami and Alireza Yari. A framework for comparing quantitative and qualitative criteria of iot platforms. In *2018 4th International Conference on Web Research (ICWR)*, pages 34–39. IEEE, 25.04.2018 - 26.04.2018.
- [31] Sanjay Bulaki Borad. English auction, 2018.
- [32] D. Schrage, C. Farnham, and P.G. Gonsalves. A market-based optimization approach to sensor and resource management. volume 6229, 2006.
- [33] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5):637–646, 2016.
- [34] Yu-e Sun, Jianying Zheng, He Huang, Kai Xing, Zhili Chen, Hongli Xu, and Liusheng Huang. Sprite: a novel strategy-proof multi-unit double auction scheme for spectrum allocation in ubiquitous communications. *Personal and Ubiquitous Computing*, 18(4):1–12, 2013.
- [35] Dinesh Thangavel, Xiaoping Ma, Alvin Valera, Hwee-Xian Tan, and Colin Keng-Yan Tan. Performance evaluation of mqtt and coap via a common middleware. In *2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, pages 1–6, Piscataway, NJ, 2014. IEEE.
- [36] Rob van Kranenburg and Alex Bassi. Iot challenges. *Communications in Mobile Computing*, 1(1):75, 2012.
- [37] Guiling Wang, Guohong Cao, Piotr Berman, and Thomas F. La Porta. Bidding protocols for deploying mobile sensors. *IEEE Transactions on Mobile Computing*, 6(5):563–576, 2007.
- [38] Guiling Wang, Guohong Cao, and T. LaPorta. A bidding protocol for deploying mobile sensors. In *11th IEEE International Conference on Network Protocols, 2003. Proceedings*, pages 315–324. IEEE Comput. Soc, 4-7 Nov. 2003.
- [39] Ying Hu, Yingjie Wang, Yingshu Li, Xiangrong Tong. An incentive mechanism in mobile crowdsourcing based on multi-attribute reverse auctions, 2018.

Chapter 7

Survey and Analysis of Existing Cloud SLA Compensation Processes and Values

Fan Feng, Ruike Wang, Yue Ding and Yuang Cheng

This report introduces concepts of Quality of Service (QoS) and Service Level Agreements (SLAs), then investigates and compares SLAs of cloud computing services from seven major Service Providers in the world. SLAs and corresponding compensation policies are evaluated from both customers' and providers' perspectives. To address problems in current SLA compensation implementations, an automated approach based on blockchain technology is introduced.

Contents

7.1	Introduction	185
7.2	Service Level Agreement	186
7.2.1	Service Level Agreements (SLA)	186
7.2.2	Service Level Objectives (SLO)	186
7.2.3	SLA Lifecycle	188
7.2.4	Compensation Process	188
7.3	Survey On Current Service Providers	190
7.3.1	Amazon Web Service	191
7.3.2	IBM Cloud	192
7.3.3	Google Cloud	194
7.3.4	Microsoft Azure	195
7.3.5	Oracle Cloud	197
7.3.6	Huawei Cloud	198
7.3.7	Alibaba Cloud	201
7.3.8	Comparison	204
7.4	State-of-art:Automated Compensation Mechanisms	206
7.4.1	Blockchain and Smart Contract	206
7.4.2	The implementation of SLA based on blockchain	206
7.4.3	Cons and Pros	207
7.5	Discussion	208
7.5.1	Concerns of privacy and data security	208
7.5.2	The Form of Compensation	209
7.5.3	Ethical Concerns	209
7.6	Summary	210

7.1 Introduction

Cloud computing, as one of the most popular and well-adopted technologies, has made great progress in recent years. The rapid growth of public cloud computing services makes it a great challenge to monitor Quality of Service (QoS) and establish the trust between Service Providers (SP) and customers. Service Level Agreement (SLA) is an agreement signed by both the SPs and the client [1], which clarifies the commitment between the SPs and the customers.

In cloud computing, grouped computer system resources are managed and provided as online services, which can be accessible to users via Internet. Thus users do not need to maintain the resources that they use, and can determine the valid duration of contract and service level based on their need with flexibility.

Clouds may be confined within an organization and managed internally, *i.e.* private clouds; or open for the public use, where the resources are shared among customers. In the following, this report mainly focus on the public cloud SPs and their SLAs.

There are several service models of cloud computing according to definition from National Institute of Standards and Technologies (NIST), Infrastructure as a Service (IaaS), Software as a Service (SaaS), Platform as a Service (PaaS) [2]. IaaS enables customers to deploy and run software without managing the underlying infrastructure, SaaS provides cloud applications that are accessible to users from local devices, and PaaS allows consumers to develop applications on the cloud infrastructures using development tools supported by providers. Another emerging service type is Storage as a Service, and it provides storage spaces to individuals or enterprises [4].

Cloud computing provides an economic alternative to computing resource consumers, where they can find pricing modes varying between SPs and even among services. For example, Amazon Web Services (AWS) offers a general pay-as-you-go approach which provides more flexibility than on-premise mode as shown in Figure 7.1, however, customers can choose to invest in reserved instances to save budget, and the price of services goes down as customers use more [5].

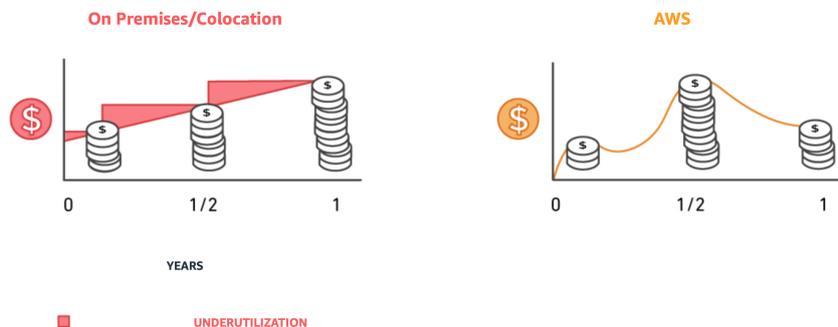


Figure 7.1: On-Premise vs. On-Demand [5]

This report is structured as follows. Section 7.2 defines SLAs and its management. Then, Section 7.3 presents a survey of current SPs, their SLAs, and compensation process. An overview of the state-of-the-art in SLA compensation is presented in Section 7.4. Further, the report is discussed in Section 7.5. Finally, Section 7.6 summarizes the report.

7.2 Service Level Agreement

7.2.1 Service Level Agreements (SLA)

SLAs in cloud computing are the crucial part to manage the cloud resources and ensure the service quality. The goal of a service level agreement is to clarify and limit the scope of service and commitment of service vendors to users, and to guarantee the rights of all involved parties. If a service provider fails to meet the agreed service level, then the SLA serves as the baseline in the following compensation process [1]. Therefore, SLAs build trust between users and SPs, and more importantly, promote the development of cloud services and wider adoption.

SLA is an agreement with regards to QoS that is established between a customer and a service provider. An SLA mainly defines agreed service, detailed characteristics of service level, metrics to evaluate the quality of the service, as well as actions to be performed in case of service violations.

The typical structure of an SLA contains the following elements [3]:

- Involved parties
- Service level parameters
- Metrics to evaluate service level
- Algorithms to compute service level
- Actions to take in case of SLA violation
- Valid time, cost of service, and location

7.2.2 Service Level Objectives (SLO)

SLOs are the targets that the cloud service want to achieve, which can be evaluated by measuring and computing metrics of service level parameters. Since SLA metrics in cloud computing architectures vary in different use cases, the corresponding SLOs rely on the metrics and clients' needs.

SLOs in cloud computing services are the specific and required parameters that agreed by both service users and SPs [6]:

- **Throughput:** the volumn of data to be delivered within a time unit
- **Security:** the integrity, encryption and authentication
- **Availability:** the quality of being able to be used or obtained
- **Response time:** time delay between the request and the process
- **Monitoring:** methods and tools of monitoring
- **Privacy:** the confidentiality related to customers and services
- **Billing:** cost and pricing mode of the service subscription

To have a deeper look at the specific SLA parameters, four types of services are selected as examples. First, if a cloud user is interested in IaaS, then the metrics to be compared may include [1]:

- **CPU capacity:** the processing ability of CPU
- **Memory size:** cache memory size of virtual machine
- **Storage:** space to store data
- **Response time:** time to process the request and respond to users
- **Boot time:** time for virtual machine to be ready for work
- **Availability:** uptime ratio of service
- **Scale up:** maximum available virtual machines for one user
- **Scale down:** minimum available virtual machines for one user

In PaaS, developers are able to develop their applications on cloud infrastructures using development tools supported by providers. The key parameters that providers usually guarantee most in PaaS are [6]:

- **Scalability:** the ability to allow a large number of online users
- **Environment of deployment:** tools and environment for developing
- **Network:** bandwidth for uploading and downloading
- **Integration:** the ability to integrate services from other platforms
- **Servers:** virtual cloud computing resources allocated to customers' application
- **Browsers:** supported browsers

Service level parameters considered in SaaS are used to evaluate the performance of applications supported by providers in cloud service [1]:

- **Usability:** easy to learn and to use
- **Reliability:** the ability to support constant operating
- **Scalability:** number of users
- **Availability:** uptime of cloud applications
- **Customization:** adaptable to users' specific needs

Another important service based on cloud resources is the Storage-as-a-Service, in which large storage space is offered to users. The SLA parameters to be guaranteed in this type service are [1]:

- **Storage Space:** available storage units to store data
- **Scalability:** change scope of storage space
- **Security:** cryptography, authentication, and authorization
- **Privacy:** confidentiality of data
- **Backup:** replicates of stored data
- **Throughput:** amount of data can be retrieved from the system in unit of time
- **Location:** the geographical place to store data

7.2.3 SLA Lifecycle

A complete SLA life cycle consists of multiple stages as proposed in the European Commission report, which are *Service Use*, *Service Modeling*, *SLA Template Definition*, *SLA Management*, *SLA Enforcement*, and *SLA Conclusion* [7], and they are depicted in Figure 7.2.

According to definition in [7], the service use phase reveals the service usage information of cloud resource consumers, and aims to obtain services. The modeling phase is to design the features and structure of services that will be provided as cloud computing resources. Next phase is to create and refine the service level templates based on business modeling process. The SLA management is the core part of the whole lifecycle, and it handles with multiple issues in SLA runtime, including instantiation, negotiation, dynamic renegotiation and so on. In the SLA Enforcement process, service monitoring data are collected and evaluated. During the final stage, SLA stops functioning and comes to an end completely.

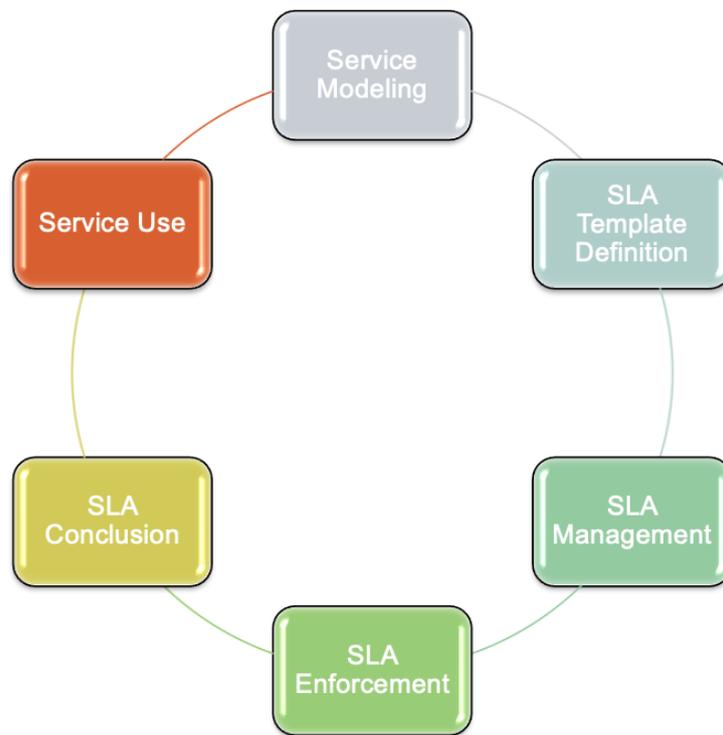


Figure 7.2: SLA lifecycle

Among these phases, the SLA management phase is the key stage that closely relates to cloud users, especially when a customer identifies a breach of SLA. According to legal terms of use of current SPs, it is required that customers submit report of the detailed description about the issue and claim for compensation within certain limited days, as well as provide sufficient evidence as support to acquire the compensation from SPs.

7.2.4 Compensation Process

During the validity, the QoS is monitored and evaluated regularly. The monitoring methods are defined in SLA in advance, and this process can be conducted by providers themselves as they provide cloud service as the same time, an example in place is Amazon CloudWatch [8], which enables observation across applications and infrastructures on a single platform, and it is mainly used to collect operational data of AWS resources. Obvi-

ously the self-monitoring relies heavily on the trust between SPs and cloud resource users, and the impartiality of the monitoring mechanism from SPs.

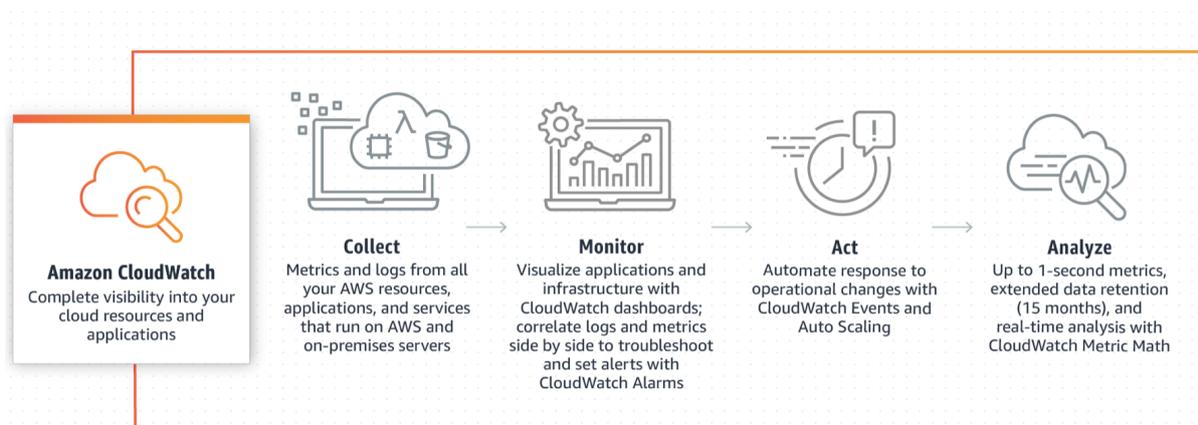


Figure 7.3: Workflow of Amazon CloudWatch [8]

The alternative to self-monitoring is outsourcing the regular inspection tasks to Trusted Third Party (TTP). However, the TTP as a monitoring agency still needs trust from cloud customers and SPs, and outsourcing service will increase the budget.

In the case that the SP does not meet the QoS defined in SLA, *e.g.*, downtime exceeding the pre-defined time, generally most SLA breaches can be identified by monitoring mechanism. And once a service provider fails to achieve the SLOs, the compensation process has to be enforced [1]. It is handled jointly by both the SLA Management and SLA Enforcement phases as mentioned in SLA lifecycle, since the threshold of compensation and redemption values are predefined during the negotiation of SLAs, and the legal commitments of providers are enforced during the SLA enforcement process.

In modern cloud services, the compensation process still requires great manual efforts to complete, *i.e.*, in the case of SLA violations, customers are supposed to report the breach of cloud service in valid time with detailed support by themselves, sometimes the procedure can be cumbersome and complex to users. If the request is approved by service provider, then they issue certain amount of service credit [16], which usually could only be spent on same cloud service consumption in future use, instead of refund or payment in cash.

Service credits are calculated as a percentage of the service fee paid by a customer, and the amount of service credit as compensation depends on the compliance degree to SLOs of providers.

7.3 Survey On Current Service Providers

Nowadays, cloud computing has become a huge business throughout the whole world due to its soaring demand from governments, companies and individuals. According to Adroit Market Research, the global cloud computing market size is estimated to reach USD 319.80 billion from USD 696.25 billion with a maximum CAGR of 10.2% over 2019 and 2025 [9]. Among all the deployments, private cloud keeps a leading role while public cloud and hybrid cloud are becoming more and more important parts. In terms of region, North America is considered a mature market in the cloud computing sector [10], owing to an outsized presence of organization with the availability of technical expertise and advanced IT infrastructure. The US and Canada are the highest contributory countries to the expansion of the cloud computing market in North America. However, it should be specially noted that Asia-Pacific countries (APAC) exist as huge potential markets and forceful development entities. That’s why this report not only investigated 4 world leading cloud SPs but also 2 Chinese cloud providers which are extending their influences over the years. Figure 7.4, Figure 7.5, and Figure 7.6 present an overview of the mentioned points [11].

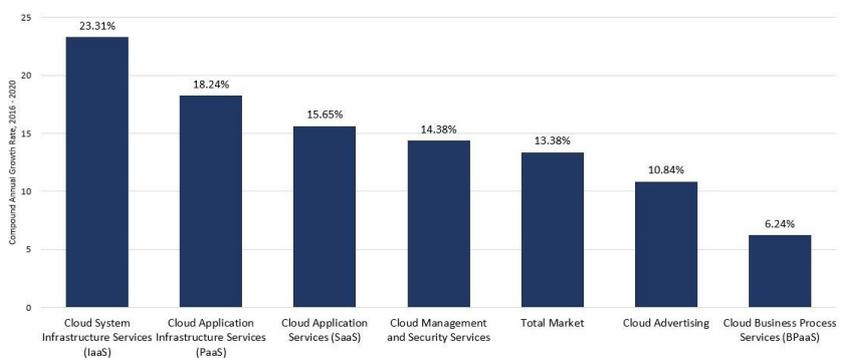


Figure 7.4: Compound Annual Growth Rates (CAGRs) of each Cloud Service Area [11]



Figure 7.5: Worldwide Public Cloud Forecast [11]

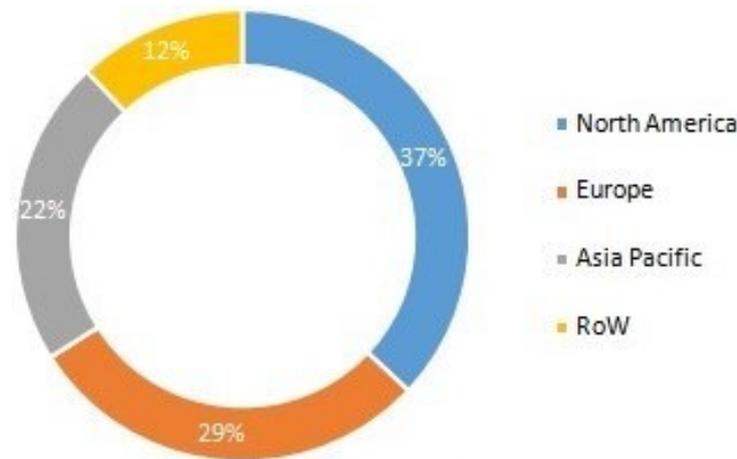


Figure 7.6: Global Cloud Computing Market by Region [12]

7.3.1 Amazon Web Service

Amazon Web Service(AWS) [13] is one of the worldwide leading cloud platforms, which offers comprehensive services varying from cloud computing to data storage, as well as many other products for different application scenarios. Among those most broadly adopted web services, Amazon Elastic Compute Cloud (EC2) provides secure, resizable compute capacity in the cloud. Moreover, AWS offers the largest global footprint in the market. No other cloud provider offers as many regions with multiple Availability Zones, with 69 Availability Zones (AZs) within 22 geographic regions around the world, and announced plans for 13 more AZs and four more AWS Regions in Indonesia, Italy, South Africa, and Spain. Figure 7.7 depicts an overall expansion of regions covered by AWS, where blue circle stands for regions already covered by AWS, and yellow circle means regions AWS is planning to cover).



Figure 7.7: AWS Global Footprint [14]

7.3.1.1 Instances Types and Pricing

Amazon EC2 defines various instance types to fit different use cases, within which CPU, memory, storage and networking capacity are combined in an optimized way, and it also allows customers to flexibly scale resources to the target workload. The provided instances are classified into several categories: general purpose, compute optimized, memory optimized, accelerated computing, and storage optimized. There are multiple ways to pay for Amazon EC2 instance: On-Demand, Reserved Instances, Dedicated Hosts, and Spot Instances [15]. It is noteworthy that EC2 accepts per-second billing based on actual usage of resources.

7.3.1.2 SLA Compensation Plan

According to Amazon Service Level Agreement, AWS is responsible to keep the Monthly Uptime Percentage of at least 99.99% [16]. In any case of failure to meet the Service Commitment, clients are eligible to be compensated with service credits. Table 7.1 shows AWS's entire compensation plan.

Table 7.1: Amazon Compute Service Compensation Plan

Monthly Uptime Percentage	Service Credit Percentage
less than 99.99% but $\geq 99.0\%$	10%
less than 99.0% but $\geq 95.0\%$	30%
less than 95.0%	100%

7.3.1.3 Credit Claim Procedure

It is customers' obligation to submit a credit request by the end of second billing cycle, otherwise no service credit will be redeemed. Moreover, AWS requests sufficient information from customers including:

1. The words "SLA Credit Request" in the subject line;
2. The dates, times, and affected AWS region of each Unavailability incident that you are claiming;
3. The resource IDs for the affected Included Service ;
4. Your request logs that document the errors and corroborate your claimed outage (any confidential or sensitive information in these logs should be removed or replaced with asterisks).

Apart from all the requirements, the Service Commitment and Hourly Commitment do not apply to any unavailability, suspension or termination an Included Service, or any other Included Service performance issues caused by factors outside of our reasonable control, or that result from any actions or inactions of you or any third party.

7.3.2 IBM Cloud

As a reliable alternative, people choose IBM Cloud [17] out of various reasons, such as high performance cloud servers, globally distributed data centers, and availability to the PaaS. IBM Cloud Services allow clients to select and configure services to meet technical requirements via IBM Cloud User Interface. The highlight of IBM Cloud is their hybrid multicloud approach, which maximizes Return on Investment (ROI) by enabling a consistent, standardized approach across all your cloud vendors.

Our approach	Other vendors
Hybrid multicloud	Hybrid monocloud
Hybrid cloud software that can run in your data center, or on any public cloud platform.	A hybrid cloud stack or appliance that runs in your data center and connects to a vendor's public cloud.
<ul style="list-style-type: none"> – Open-standards based – Any cloud infrastructure or service – Hybrid or Private (not tethered) – Multivendor portability – Multicloud management 	<ul style="list-style-type: none"> – Proprietary – Same software and hardware – Tethered hybrid only – Single-vendor portability – Single-vendor management

Figure 7.8: IBM Cloud Service Highlight [18]

7.3.2.1 SLA Compensation Plan

IBM provides SLAs for two kinds of configurations, High Availability and Non High Availability, which also corresponds to different downtime measurements.

Table 7.2: Monthly Uptime Percentage

High Availability	Non-high Availability	Service Credit Percentage
less than 99.99% but \geq 99.90%	less than 99.9% but \geq 99.0%	10%
less than 99.90%	less than 99.0%	25%

7.3.2.2 Infrastructure Hardware Replacement and Upgrade SLA

To be noted, the downtime of hardware is also an important component of IBM Cloud SLA, and there exists a compensation plan for excessive long time periods of hardware replacement and upgrade. This is a plan that greatly benefits customers, and IBM is the only company offers infrastructure hardware replacement and upgrade compensation among all the cloud service providers are introduced in this report.

Table 7.3: Infrastructure Hardware Replacement and Upgrade SLA

Service Level Time Period	Credit Percentage
less than 2 hours	none
2 hours	20%
6 hours	40%
10 hours	60%
14 hours	80%
18 hours	100%

7.3.3.1 Services by Google Cloud Platform

Google provides a huge range of cloud services include but not limited to Cloud Filestore, Cloud Datastore, Cloud IoT Core, Cloud Natural Language, and Cloud SQL.

7.3.3.2 SLA and Compensation Plan

After looked into several services available on Google cloud platform, now we can put our eyes on the services level agreements of Google cloud platform. Services providers are supposed to provide cloud services in good quality to satisfy daily or commercial usage, and they also need to have clear compensation policy to deal with breakdown situations. Services level agreements are important for giving users a picture about the services quality and how compensation works. These agreements describe the services qualitatively and quantitatively to make the services seem trustworthy to stakeholders. Google determined a term Services Level Objective (SLO) to describe how well should a service be. They also explained several terms to make the agreements more unambiguous to readers for each service. For example, what Covered Services mean differs from services to services. For Cloud Filestore provided by Google, the services level objective is to provide a monthly uptime percentage to customer of at least 99.9%. Fail to meet the services level objective can trigger the compensation process. The compensation policy describes three compensation level for three different situation and clearly states that what would its customers get for reimbursement is financial credits. The compensation level is decided by the following method: if the monthly uptime percentage is lower than 99.9% but higher than 90.0%, the customer gets credits which equals to 10% of his/her monthly bill. The mentioned policies are listed in the following table.

Table 7.4: Cloud Filestore SLA [20]

Monthly Uptime Percentage	Service Credit Percentage
less than 99.9% but \geq 99.0%	10%
less than 99.0% but \geq 95.0%	25%
less than 95%	50%

7.3.3.3 Credit Claim Procedure

The compensation process of Google cloud platform is manually. The Google technical support must be informed by the customer in order to deal with the compensation. Besides, customer must also provide Google with log files showing Downtime Periods and the date and time they occurred. The aggregate maximum number of Financial Credits to be issued by Google to Customer for any and all Downtime Periods that occur in a single billing month will not exceed 50% of the amount due from Customer for the Covered Service for the applicable month.

7.3.4 Microsoft Azure

Microsoft Azure provides customers over 90 compliance offerings which make it the largest portfolio in the industry. Azure also claims that USD 1 billion is invested per year in security to protect customers data from cyberthreats.

Besides, Azure provides flexible purchasing and pricing options for all possible cloud scenarios, such as the Azure Hybrid Benefit, and offers extensive tools to help customers manage their cloud spends. According to Azure price calculator, the cost in most countries

are below 4 thousand dollars a month for 48 cores, 192 GB RAM and 1200GB Temporary Storage, but the same service for Switzerland is slightly expensive, over 4 thousand. There are almost 80 instances for all regions to choose, except South Africa, where only 2 choices are available. While, there are no services available in other parts of Africa. Customers could choose to pay as they go or reserve one or three months, of course the latter is slightly cheaper. Apart from above, Microsoft Azure also has price advantage comparing to AWS, including but not limited to the costs of running Windows Server virtual machines (VMs), SQL Server running as a PaaS service, and SQL Server running on a virtual machine (IaaS).

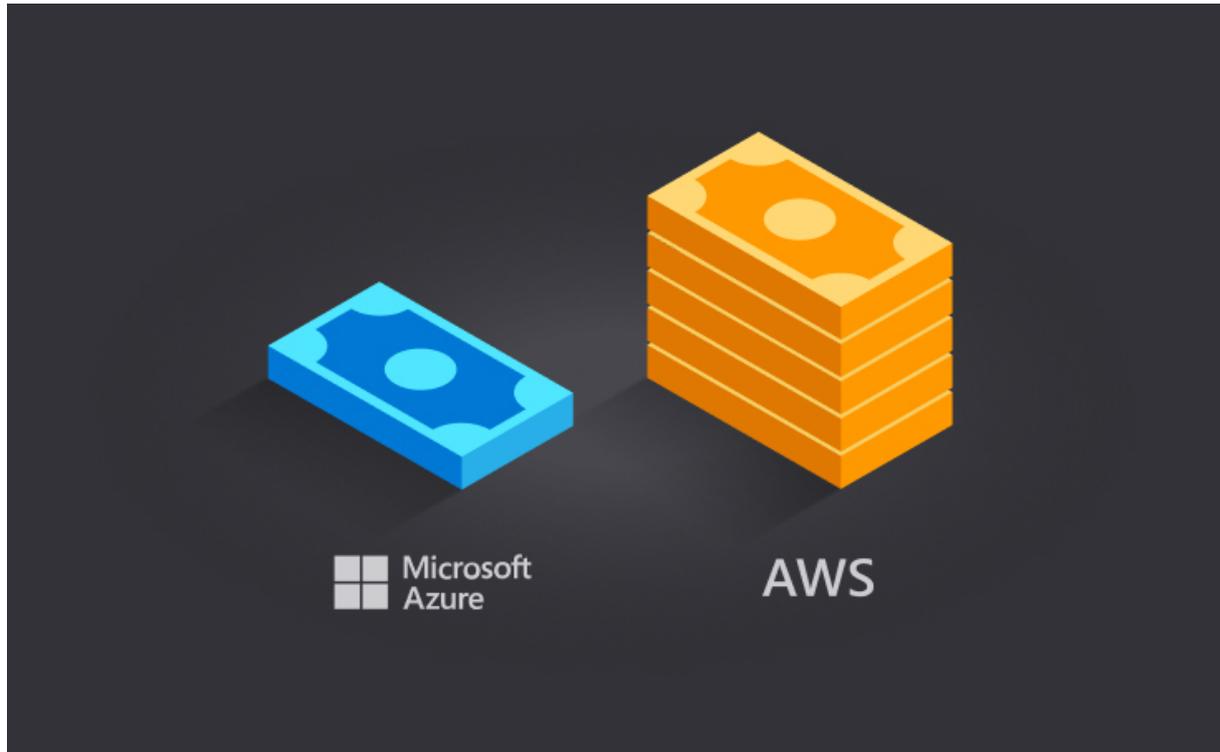


Figure 7.10: Pricing policies of Azure and AWS [21]

7.3.4.1 Marketplace

Microsoft Azure also works as a platform, helping sell the Apps developed by other companies which are using Microsoft's services, such as DELL and CHECK POINT. Besides, there are also various paid consulting services, mainly in forms of workshops and implementation. Any customer can contact the third parties directly.

7.3.4.2 SLA Compensation Plan

Microsoft Azure offers varieties of services compared to other SPs, such as Blockchain, Databases, Integration, Internet of Things, Storage. We will have a deep look at computing, specifically cloud services. The compensation for monthly uptime percentage is shown as below.

Table 7.5: SLA Compensation Plan

Monthly Uptime Percentage	Service Credit Percentage
less than 99.95% but $\geq 99.0\%$	10%
less than 99.0%	25%

7.3.4.3 Credit Claim Procedure

In order for Microsoft to consider a claim, customers must submit the claim to customer support at Microsoft Corporation including all information necessary for Microsoft to validate the claim, including but not limited to:

- A detailed description of the Incident;
- Information regarding the time and duration of the Downtime;
- The number and location(s) of affected users (if applicable); and
- Descriptions of your attempts to resolve the Incident at the time of occurrence.

7.3.5 Oracle Cloud

Oracle Cloud provides PaaS and IaaS public cloud services. There are three physical hierarchies in Oracle Cloud infrastructure: Region, Availability Domain(AD) and Fault Domain(FD). Region refers to a localized geographic area, such as Zurich, Frankfurt. AD is data centers located in a region and FD is a grouping of infrastructure within an AD. These definitions will be helpful in illustration of SLA [22].

7.3.5.1 Overall Policy

The service of Oracle Cloud is measured in three dimensions: availability, manageability and performance. Should the target SLA is not met, customers will get the compensation in the form of Oracle Service Credit, which is calculated as a percentage of the net fees that customers have paid. The compensation ranges from 10% to 25%.

7.3.5.2 Details and Services

Availability. Uptime or Availability Level is measured over the immediate preceding month for every calendar month. As shown in the table different measurement is applied according to the type of service.

Table 7.6: Availability SLA

Service	Measurement	SLA Uptime	Credit
Compute	Region	less than 99.99% but $\geq 99\%$	10%
		less than 99%	25%
Compute	AD	less than 99.95% but $\geq 99\%$	10%
		less than 99%	25%
Block Volumes	Region	less than 99.99% but $\geq 99\%$	10%
		less than 99%	25%
Object Storage	API error rate	less than 99.9% but $\geq 99\%$	10%
		less than 99%	25%
Fast Connect	Private Connectivity	less than 99.9% but $\geq 99\%$	10%
		less than 99%	25%

Manageability. Manageability is measured as API error rate. The SLA ensures customers to manage and monitor resources.

Table 7.7: Manageability SLA

Service	SLA Uptime	Credit is SLA breached
Compute	less than 99.9% but $\geq 99\%$	10%
	less than 99%	25%
Block Volumes	less than 99.95% but $\geq 99\%$	10%
	less than 99%	25%
Database	less than 99.9% but $\geq 99\%$	10%
	less than 99%	25%

Performance. Oracle Cloud is the first cloud vendor to guarantee performance in SLA. It is measured as disk IOPS or network performance.

Table 7.8: Performance SLA

Service	Measurement	Performance SLA in 99% of time	Credit if SLA breached
Compute local NVMe drive	disk IOPS	less than 99.9% but $\geq 99\%$	10%
		less than 99%	25%
Block Volumes	disk IOPS	less than 99.9% but $\geq 99\%$	10%
		less than 99%	25%
Network	network performance	less than 99.9% but $\geq 99\%$	10%
		less than 99%	25%

7.3.5.3 Credit Claim Procedure

The compensation process is not automated. Customers are required to manually file the claim with supporting evidences provided. The claim must be submitted within 30 days from when the SLA is breached. The Service Credit is only valid in the month when it is granted and will be expired after this calendar month.

7.3.6 Huawei Cloud

Huawei is a leading world-class communication enterprise based in China. Apart from the basic services and products, for instance, mobile communication technologies and smart-

phones, it also provide cloud services whose bases located at different cities across the globe. The data centers of Huawei Cloud distribute as the figure 7.11 [23] illustrates.



Figure 7.11: Huawei Cloud data centers [23]

7.3.6.1 Services and Prices

Cloud services provided by Huawei included but not limited to compute, storage, security. Table 7.9 shows example products and corresponding content of these products.

Table 7.9: Huawei Cloud Services

Cloud Products	Content
Compute	Auto Scaling, FunctionGraph, etc.
Storage	Object Storage Services, Scalable File Services, etc.
Security	Anti-DDos, Web Application Firewall, etc.

Prices for Huawei Cloud products varies. Different services, different chosen data centers, different operation systems from customer's side yield different costs. For each service, several subscription options for prospective customers are listed on the website. It can be observed that generally, farther chosen data center from China has higher price. For same data center selected, the most flexible but most expensive option is to purchase Huawei Cloud services by hours. Customers can be benefit from discounts if longer subscriptions are selected or higher storage capabilities are purchased. Sample prices are listed in the table 7.10.

Table 7.10: Huawei Cloud Price Samples

	Sample 1	Sample 2	Sample 3
Services	EC *	EC *	Advanced Anti-DDos**
Data Center	Singapore	Hong Kong	Hong Kong
Price per hour (USD)	0.032	0.02	Not Provided
Price per month (USD)	22.34	10.7	8'000
Price per year (USD)	223.38	107	96'000

* EC: Elastic Cloud. 1 vCPU and 1GB Memory

** Website service with basic Protection Bandwidth (50Gbit/s) and 5 domain names protected

7.3.6.2 SLA Compensation

The service level agreement of Huawei is clearly stated on Huawei website.

In SLA, crucial and relevant terminologies are defined, and situations not covered by SLA, which means customer would not get compensation under those situations, are listed. Here are a few instances. The term 'Service Cycle' is defined as 'a natural month' and the term 'Total Time of Service Cycle' is described as 'the total minutes during every Service Cycle' [24]. Issues caused by normal maintenance, by the customer's side or the third parties' side, and by Art of God are not excluded from SLA. The only compensation that customers would receive for any issues listed in SLA is Service Credit. Unfortunately, the definition of Service Credit is not clearly showed on the website.

Most interested parts of the SLA, however, are the conditions and policies about compensation. To receive remedy, first the related customers should make sure that they apply after their bills for the Service Cycle are settled and within the required period. The deadline of such application is 'two months after the end of the Service Cycle in which the incident that's the subject of the claim occurred [24]'. Then the application can be submitted online via the link provided on the Huawei SLA website. After the Huawei user center received the application, the services availability rate with respect to total minutes in service circle will be computed. The compensation will be sent to customers after the service credit is calculated according to specific policies. The computation criteria of services are described below in table 7.11.

Table 7.11: Service Credit for Content Delivery Network

Service Uptime	Service Credit
less than 99.99%	Service Credit = (Service Unavailability Duration * the Monthly Guaranteed Minimum Bandwidth Usage Fee by the Contract / the total minutes in a Service Circle) * 2 [24]

The policy for Content Delivery Network is quite flexible as it does not simply divide the remedy into several levels but calculate it by real unavailability rate. Note that the service credit cannot exceed the monthly usage fee.

Table 7.12: Two Levels compensation

Levels	Uptime Percentage	Service Credit
Level 1	less than 99.9% but $\geq 99\%$	10%
	less than 99.5% but $\geq 99\%$	
	less than 99.9% but $\geq 99\%$	
	less than 99% but $\geq 98\%$	
	less than 95% but $\geq 90\%$	
Level 2	less than 99%	30%
	less than 99%	25%
	less than 98%	25%
	less than 90%	30%

Table 7.13: Three Levels compensation

Levels	Uptime Percentage	Service Credit
Level 1	less than 99.9% but $\geq 99.9\%$	10%
Level 2	less than 99.8% but $\geq 99.5\%$	20%
Level 3	less than 99.5%	50%

For other services, the compensation is divided into two or three levels. showed in table 7.12 and table 7.13. Most of the service have the two levels compensation policy, for example, Object Storage Service, Relational Database Service, and Cloud Stream Service. Only two services hold the three levels compensation policy, which are Bare Metal Service and Advanced Anti-DDos.

7.3.7 Alibaba Cloud

Alibaba is a world-class internet commerce, retail and technology multinational conglomerate company. It is famous for its online C2C platform, Taobao, also known as Aliexpress, which based on Alibaba's own cloud services from Alibaba Cloud.

Alibaba Cloud, also named Aliyun, is a leading cloud service vendor based in China and is now expanding its business in SEA and Europe. Alibaba Cloud covers a wide range of high-quality services with reasonable price. On Alibaba Cloud website, they highlight that they provide cloud solutions with end-to-end security and boundaryless connection. Figure 7.12 [25] from its website shows how the cloud system works. Examples of Alibaba Cloud services and prices will be showed in the following section.

7.3.7.1 Services and Prices

Alibaba Cloud provides more than 100 types of cloud services divided into several categories, which is not possible to cover them all in this section. Table 7.14 shows examples of categories and services provided by Alibaba Cloud.

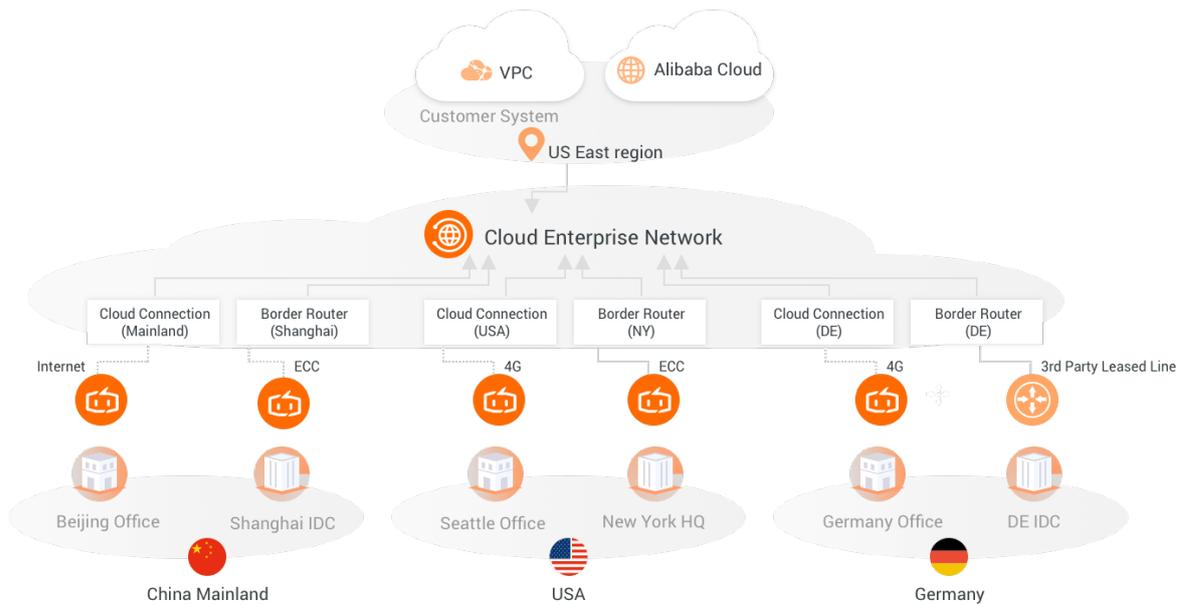


Figure 7.12: Alibaba Cloud System [25]

Table 7.14: Alibaba Cloud Services

Categories	Services
Elastic Computing	Elastic Compute, Elastic GPU, etc.
Database Services	Data Transmission, Time Series Database, etc.
Networking	Virtual Private Cloud, Elastic IP, etc.

The pricing of Alibaba Cloud is listed on website in details but it is quite complicated. For each service, Alibaba Cloud offers different subscriptions. Monthly subscription and purchase by hour are available and each option has two level of services, the entry-level service and the enterprise-level service. Yearly reserve option does not distinguish entry-level and enterprise-level. Customers would benefit from longer subscriptions by higher discounts. The price of one Alibaba Cloud service is consisted of instance fee, storage fee, and network traffic fee, etc. and for each fee it also varies. For example, the instance fee of a monthly subscribed entry-level service is calculated by different instance type chose by the customer. Table 7.15 shows an example of service price. Next section will be introduction and analysis of Alibaba Cloud service level agreement.

Table 7.15: Alibaba Cloud Price Example

Service	Fees*				
	Instance Fee	Level	Type	Hourly	Monthly
Elastic Compute		Entry-Level	Burstable**	0.05	13.27
	Storage Fee	100GB	ESSD Cloud Disk PL1	0.032	15.30
	Network Traffic Fee	0.11/GB			

* In USD

** Concretely, instance type: ecs.t5-c1m1.large; with 2 CPU and 2GB Memory.

7.3.7.2 SLA Compensation

Alibaba does not offer a general description about service level agreement but service level agreement of services respectively.

In each SLA, pronouns, scope and definitions are stated carefully. For services that not only include single instance, multi-zone service is defined. In Elastic Compute Service (ECS) SLA, instance unavailable is described as ‘the ECS instance is deemed unavailable if the disconnection between an ECS instance configured with access permitted rules and any IP address over TCP or UDP in the inbound and outbound directions lasts for more than one minute. [26]’ And multi-zone service unavailable is described as ‘if you have deployed ECS instances in at least two zones in one region, and any zone (“Unavailable Zone”) experiences Instance Unavailable for all ECS instances, and any ECS instance in any other zone(s) in the same region (“Other Zone(s)”) also experiences Instance Unavailable, then such Instance Unavailable in Other Zone(s) is called the instance(s) encounter Multi-zone Service Unavailable. [26]’ Exclusions that customers would not get paid if the services are down are also stated, for example, natural disasters, and payment delay.

Table 7.16: Alibaba Cloud Product SLA

Service	Uptime Percentage	Service Credit
ECS/Instance	less than 99.95% but $\geq 99\%$	10%
	less than 99% but $\geq 95\%$	25%
	less than 95%	100%
ECS/Multi-zone	less than 99.99% but $\geq 99\%$	10%
	less than 99% but $\geq 95\%$	25%
	less than 95%	100%
ApsaraDB for RDS	less than 99.99% but $\geq 99\%$	10%
	less than 99%	25%
Anti-DDoS Premium	less than 99.99% but $\geq 99.98\%$	10%
	less than 99.98% but $\geq 99.95\%$	25%
	less than 99.95%	50%
Blockchain	less than 99.98% but $\geq 99\%$	10%
	less than 99%	25%
MaxCompute	less than 99.99% but $\geq 99\%$	15%
	less than 99% but $\geq 95\%$	30%
	less than 95%	100%
AI Face Recognition	less than 99.99% but $\geq 99.5\%$	15%
	less than 99.5 but $\geq 95\%$	30%
	less than 95%	100%

The compensation process required customers to submit a detailed application within the deadline and in a particular period of a month. Concretely, in ECS SLA, the deadline is sixty days after the problem month ends, and the submission period is ‘starting from the sixth working days of the following calendar month of occurrence of the event giving rise to the claim [26]’. In ECS SLA, it is noticeable that multi-zone unavailable time can be summed up but the final compensation per instance is calculated respectively and the remedy cannot exceed the monthly service fee per instance.

Table 7.16 presents examples of compensation policies in the Alibaba cloud.

7.3.8 Comparison

This section compares the selected cloud SPs in terms of offered services and SLA compensation processes.

7.3.8.1 Services Comparison

Services offered by different providers are actually quite similar in terms of their content. In general, all providers offer cloud service that can be categorized into Compute, Storage, Security, and Networking. These world-class enterprise all have data centers across the globe and thus, customers can choose a nearby data center to improve access speed and avoid the shutdown caused by submarine cables damages. The price of their services varies and it is hard to compare without knowing their actual difference and without similar pricing systems. The pricing systems, however, is comparable, and it is clear that Alibaba holds a complicated system which one cannot have a basic picture in mind about its service fee by just skimming the website.

7.3.8.2 SLA Compensation Comparison

The compensation policies of these providers are unsurprisingly different. One obvious difference is that they have different compensation levels and accordance compensation rate. Some providers, for example, Alibaba, offer full compensation that customers can receive compensation that equals to their monthly service fee. However, they also hold few same rules. For instance, the remedy is calculated based on monthly service time and the remedy for customer cannot exceed their monthly service fee. Commonly, the remedy is paid by credits, not real money. Besides, all cloud service companies describe some exclusions that do not trigger compensation in their SLA. These exclusions are mostly the same, including force majeure, issues caused by customer's side or third parties' side, etc.

The compensation procedures of these cloud service vendors, however, are similar. Statements of time limit are constantly included in the SIAs. Application for compensation is required to submit within the deadline by all these vendors. Another similarity is, compensation process is handled manually.

Figure 7.13 depicts a general compensation procedure.

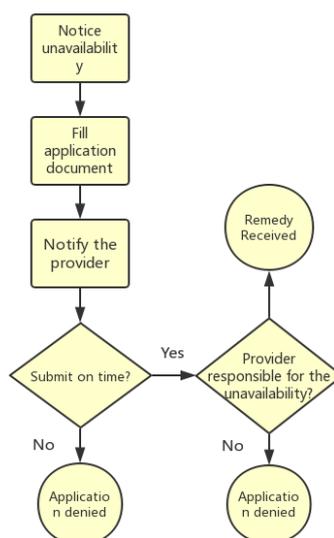


Fig.sub.1

(a) Customer Side

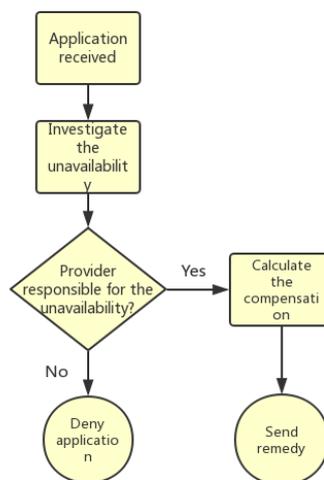


Fig.sub.2

(b) Provider Side

Figure 7.13: Compensation Procedure

The manual compensation process is complicated. To perform such process can not only waste time but also waste money. Therefore, solutions to deal with these problems are needed. Two current methods that present alternatives to manage the compensation process are introduced in the next section.

7.4 State-of-art:Automated Compensation Mechanisms

The compensation policies and processes regulated in SLA of major SPs do not differ a lot from each other. Customers are required to file claims for compensation manually. And in such model, SP and customer trust a third-party(*e.g.* a bank) and their chosen monitoring solutions. This method makes compensation process troublesome because one party may not trust the monitoring solution trusted by the other party.

Eder J. Schied proposed an implementation of automated compensation [27] based on Ethereum blockchain and Smart Contract. This approach makes the compensation process fully automated and brings a new trust relationship between SP and customer(see figure 7.14). A single monitoring solution that is trusted by both SP and customer acts as a third-party and injects monitored information of service into SC.

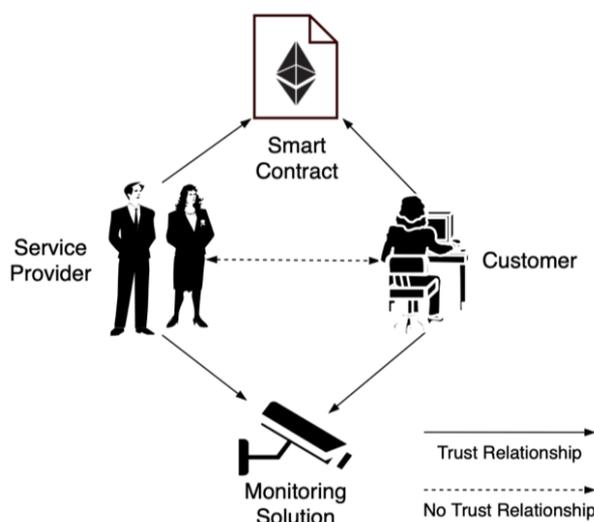


Figure 7.14: Relationship in Proposed Implementation [27]

7.4.1 Blockchain and Smart Contract

Blockchain is a distributed ledger or in other words a decentralized peer-to-peer data storage. It stores transactions and other data which is broadcasted to every stakeholder. In a decentralized system with high Byzantine fault tolerance that nodes can act maliciously and that a complete failure is possible to take place, a consensus mechanism has to be claimed in order to make every node reach the same state. And it is the consensus mechanism that ensures the framework of blockchain and data stored on it to be permanent and tamper-proof. In Bitcoin and Ethereum, Proof-of-work serves as the consensus algorithm. The block whose hash value is lower than the hash value of previous block is the next block of this chain.

In Ethereum, Smart Contracts(SC) are codes run on Ethereum Virtual Machine(EVM). Since the SC itself can not access data outside the blockchain therefore an Oracle SC should be deployed in order to introduce data from outside world into SC. A commonly trusted monitoring solution serves as an Oracle in this method.

7.4.2 The implementation of SLA based on blockchain

Only Cloud service provider, customer and a monitoring solution trusted by both parties are authorized to interact with the SC. The SC checks whether the monitoring solutions

proposed by the SP and the customer respectively are the same. The monitoring solution collects information of cloud service and inject this data into SC. The compensation value is calculated based on data provided by this commonly trusted monitoring party and predefined compensation policy.

To deploy a new SC, the configuration of SLA has to be stated. The critical terms and parameters of SLA like validity, service price, compensation value and addressed of stakeholders have to be translated into variables in SC. Given that SP and customer require some sensitive data to be undisclosed, the data stored on chain can be encrypted using public key.

The business model in the implemented SLA is different from what is applied in conventional SLA model used by major SPs. Figure 7.15 shows how the payment and compensation work. Customer pays a deposit as the service fee to contract's address and thereafter cloud service is initiated. However, this amount of money (in the form of digital currency) will not be transferred to SP until the end of the service. The whole service is divided into several monitoring sessions. In each session SLA is monitored by the monitoring party and data of objectives (*e.g.* service uptime) is injected into the SC. The value of compensation is calculated in each session and money will be transferred from contract's address back to customer's address. In sum SP will receive what the customer pays at the beginning of the service less total compensation value calculated in each session.

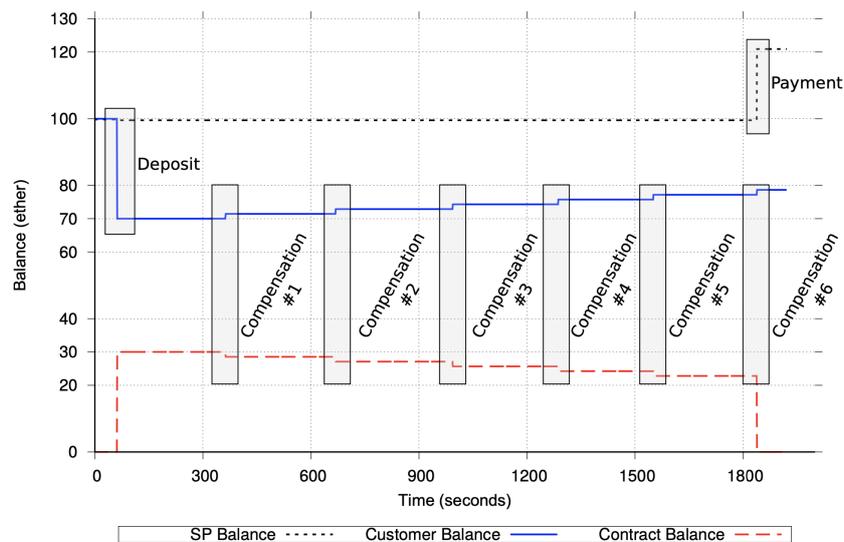


Figure 7.15: SLA payment and compensation [27]

7.4.3 Cons and Pros

The major advantage of implementation of SLA on blockchain is automated processing. Customers do not have to file claims for compensation manually and SPs don't have to employ human effort to check applications. The payment, monitoring and compensation processes are fully automated. Additionally, with the help of other automated implementation of SLA, the whole stages of SLA including negotiation [28] can proceed automated.

The decentralized and tamper-proof nature of blockchain ensures the absolute neutrality and security of this method. Once the SLA policies and parameters are defined in the SC, they can no longer be altered by any party. The possibility of misconduct is minimized. Therefore the blockchain serves as a mechanism and as a data storage that can be trusted by anyone. SC itself can possess an account which enables SC to serve as

a perfect third-party with absolute impartiality.

The proposed implementation requires a third-party monitoring solution that is trusted both by SP and customer. The process will be more convenient and less complicated. However, there is potential problem when malfunctioning of monitoring solution takes places. It will lead to anomalies in monitoring and compensation process of SLA.

The value of digital currency is extremely unstable. For instance, the price of ETH drops almost 50% from 328 CHF to 189 CHF in less than 4 months(see figure 7.16). The high volatility undermines the usability of decentralized applications that are closely connected with financial and legal affairs in real world. Moreover, it is hard to persuade SPs to accept digital currency as a method of payment. Companies may prefer cash to digital currency. The liquidity of the latter is incomparable with that of cash.

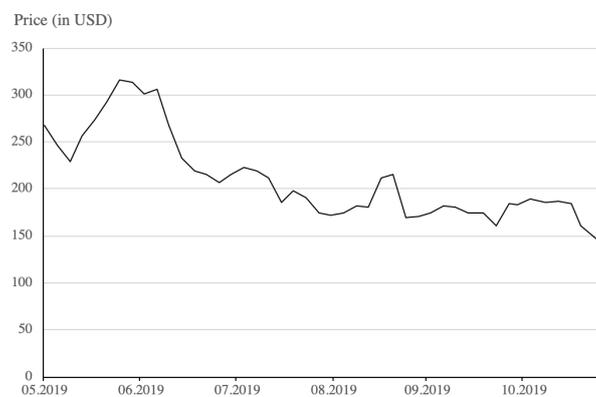


Figure 7.16: Price of Ether *datasource:coinbase* [29]

Finally, transaction fee(gas in Ethereum) should not be neglected in this decentralized application. Stakeholders have to pay gas when they transfer money or run functions in a SC. The amount of transaction fee rises when operations are complex. From SP's perspective, a possible solution is to include predicted gas value in total service fee.

7.5 Discussion

7.5.1 Concerns of privacy and data security

Data stored on blockchain is broadcasted to all nodes. Therefore the privacy of data should be taken into consideration when SLA is implemented based on blockchain. SPs and customers wouldn't like to see their sensitive business information to be disclosed to competitors. The proposed method to deal with data privacy in blockchain implementation employs encryption using public key.

It is a more general concern to stakeholders that whether we should trust a third-party monitoring solution and that whether it is secure to hand sensitive data to third-party. Especially when the mutually agreed General Data Protection Regulation (GDPR) [30] came into force on May 25, 2018. It is public knowledge that multiple companies have already been fined in various countries under the scope of the GDPR regulation. Depending on the nature and extent of infringement, the potential fines are divided into two tiers. Namely, the lower level with the higher of 10m Euro or 2% of worldwide annual turnover, and the upper level higher of 20m Euro or 4% of worldwide annual turnover. Take Microsoft as an example, with a record annual turnover of \$110.4 billion in 2018,

potential breaches related to GDPR could hypothetically lead to a maximum penalty of \$4.4 billion. So customer privacy and data security are taken seriously. Actually even SPs themselves are careful about customers Personally Identifiable Information (PII), let alone any third party (It is stated in Amazon Cloud Service SLA that 'any confidential or sensitive information in these logs should be removed or replaced with asterisks').

7.5.2 The Form of Compensation

According to SLAs of major cloud SPs introduced in this paper when SLA violations take place customers can get Service Credit as compensation instead of cash, and these service credit can be used to pay service fee. This form of payment is indeed convenient for long-term users who are continuously using cloud service provided by the same SP. However, it is not stated in SLAs how will SPs proceed with compensation when customers want to end the service. Whether customers in such condition can withdraw from balance or Service Credit in their service accounts is not clearly stated.

7.5.3 Ethical Concerns

A common concern on the Internet is that automation would lead to unemployment. The smart contract introduced in this report is also a sort of automation technique. It is more like a social science problem but still worth to take into consideration as technology is highly related to daily life. One should carefully think about the consequence while developing new technique although it might be difficult in most case. Luckily, for automation, now it is widely agreed that it would not cause severe unemployment in long term. Although it might make people lose their jobs in short period, the market always create new opportunities and most of these people would shift to new positions.

7.6 Summary

A service-level agreement(SLA) is a commitment between SPs and customers. In the context of cloud-computing. SLAs clarify the scope of services and guarantee the quality of service. Service-level objective(SLO) is a target that a cloud service should reach. SLOs are evaluated by measuring service-level parameters including availability, reliability, manageability and scalability. SLA is usually consist of several SLOs.

In case that SLAs are violated, customers can get compensation depending on compensation policies of SPs. SLAs and compensation policies of top cloud service vendors are introduced in this paper. Despite the difference in compensation values, there are few variations in compensation processes. Customers are required to file claims for compensation manually whatever SP they choose. Service Credit is paid to customers as compensation, which can be used to pay service fee. This type of compensation process is complex.

An implementation of SLA based on Ethereum blockchain and Smart Contracts proposed by Eder J. Schied [27] can overcome this shortcoming. Thanks to the decentralized and tamper-proof nature of blockchain, this method brings high security and neutrality. Moreover, it makes compensation of SLAs fully automated. However, there are still unsolved problems such as the extreme unstability of digital currency.

Bibliography

- [1] Mohammed Alhamad, Tharam Dillon, and Elizabeth Chang. Conceptual SLA Framework for Cloud Computing. In *4th IEEE International Conference on Digital Ecosystems and Technologies*, pages 606–610. IEEE, 2010.
- [2] Peter Mell, Tim Grance, et al. The NIST Definition of Cloud Computing. 2011.
- [3] Keller, Alexander and Ludwig, Heiko. The WSLA framework: Specifying and monitoring service level agreements for web services. *Journal of Network and Systems Management*, 11(1):57-81, 2003.
- [4] Gurudatt Kulkarni, Ramesh Sutar, and Jayant Gambhir. Cloud Computing-storage as Service. *International Journal of Engineering Research and Applications (IJERA)*, 2(1):945–950, 2012.
- [5] Amazon Web Service. How does AWS pricing work? Available at "https://aws.amazon.com/pricing/?nc2=h_q1_pr_ln#".
- [6] Chenkang Wu, Yonghua Zhu, and Shunhong Pan. The SLA Evaluation Model for Cloud Computing. In *International Conference on Computer, Networks and Communication Engineering (ICCNCE 2013)*. Atlantis Press, 2013.
- [7] Dimosthenis Kyriazis. Cloud Computing Service Level Agreements Exploitation of Research Results. *European Commission Directorate General Communications Networks Content and Technology Unit, Tech. Rep.*, 5:29, 2013.
- [8] Amazon CloudWatch. Observability of Your AWS Resources and Applications on AWS and On-premises. Available at "<https://aws.amazon.com/cloudwatch/>".
- [9] Adroit MarketResearch. Global Cloud Computing Market Size 2018 by Deployment (Public, Private, Hybrid Cloud), By Application (IT Telecom, BFSI, Manufacturing, Aerospace Defense, Retail, Consumer Electronics), By Product (IaaS, PaaS, SaaS), By Region and Forecast 2019 to 2025. Available at "<https://www.adroitmarketresearch.com/industry-reports/cloud-computing-market>".
- [10] Marketers Media. Cloud Computing Market Share 2019 Overview by Types of Services Deployment Models, Service Providers, Benefits, Technology, Security Challenges, Applications Forecast 2025. Available at "<https://marketersmedia.com/cloud-computing-market-share-2019-overview-by-types-of-services-deployment-models-88893602>".
- [11] Forbes. Cloud Computing Market Projected To Reach 411B By 2020. Available at "<https://www.forbes.com/sites/louiscolombus/2017/10/18/cloud-computing-market-projected-to-reach-411b-by-2020/#2816368678f2>".

- [12] Variant Market Research. Global Cloud Computing Market by Region. Available at "<https://www.variantmarketresearch.com/report-categories/information-communication-technology/edge-computing-market>".
- [13] AWS. Amazon Web Service. Available at "<https://aws.amazon.com/>".
- [14] Amazon. Amazonpic. Available at "<https://aws.amazon.com>".
- [15] Amazon Web Service. Amazon Elastic Compute Cloud Pricing. Available at "<https://aws.amazon.com/ec2/pricing/>".
- [16] Amazon Web Service. Amazon Compute Service Level Agreement. Available at "<https://aws.amazon.com/compute/sla/>".
- [17] IBM Cloud. Ibm Cloud Service Description. Available at "[https://www-03.ibm.com/software/sla/sladb.nsf/pdf/6605-19/\\\$file/i126-6605-19_10-2019_en_US.pdf](https://www-03.ibm.com/software/sla/sladb.nsf/pdf/6605-19/\$file/i126-6605-19_10-2019_en_US.pdf)".
- [18] IBM. IbmePIC. Available at "<https://www.ibm.com/cloud/why-ibm>".
- [19] Googlepic. Available at "<http://www.rubelmahmud.com/useful-google-products-services-you-probably-dont-introduce-yet>".
- [20] Google Cloud Platform. Google Cloud Platform Service Level Agreements. Available at "<https://cloud.google.com/terms/sla/>".
- [21] Microsoft. Micropic. Available at "<https://azure.microsoft.com/en-us/overview/azure-vs-aws>".
- [22] Oracle Cloud. Oracle Cloud Infrastructure Service Level Agreement. Available at "<https://www.oracle.com/cloud/iaas/sla.html>".
- [23] Huawei. Huawei Cloud Information. Available at "<https://intl.huaweicloud.com/en-us/global/>".
- [24] Huawei. Huawei Cloud SLA. Available at "<https://www.huaweicloud.com/en-us/declaration/sla.html/>".
- [25] Alibaba. Alibaba Cloud Information. Available at "https://www.alibabacloud.com/solutions/networking?utm_content=se_1003136172&gclid=EAIaIQobChMIs5_m9Z7o5QIVGKaaCh3GJQq_EAAYASAAEgL7BvD_BwE".
- [26] Alibaba. Alibaba ECS SLA. Available at "<https://www.alibabacloud.com/help/doc-detail/42436.htm>".
- [27] Eder J. Scheid, Bruno B. Rodrigues, Lisandro Z. Granville, and Burkhard Stiller. Enabling Dynamic SLA Compensation Using Blockchain-based Smart Contracts. *IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, 2019.
- [28] Linlin Wu, Saurabh Kumar Garg, Rajkumar Buyya, Chao Chen, and Steve Versteeg. Automated SLA Negotiation Framework for Cloud Computing. *IEEE/ACM International Symposium*, 2013.
- [29] Coinbase. Ethereum Price Chart. Available at "<https://www.coinbase.com/price/ethereum>".
- [30] European Union. The General Data Protection Regulation. Available at "<https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1528874672298&uri=CELEX%3A32016R0679>".

Chapter 8

Commercializing Blockchain: Transformation and Emergence of Web 3.0 Business Models

Clive C. Javara, Naël M. H. Prélaz, Syed S. Ahmed, Alphonse Mariyagnanaseelan

Abstract

The intended purpose of this report is to understand the dynamics of the Web 2.0 and Web 3.0 principles, their comparison, and their place in this era. The secondary purpose of this review is to understand the different business models of Web 2.0 and Web 3.0 and their implications. Moreover, it also entails a detailed analysis of the economic side of blockchain technology by discussing it from an institutional perspective rather than from a core technical point of view. It also discusses briefly the trust mechanism behind blockchains and looking at the sustainability paradigm of blockchains from the lens of the Fourth Industrial Revolution. Lastly, it acknowledges the limits of blockchain technology from an economical perspective and provides weight through various real-world commercial examples.

Contents

8.1	Introduction	3
8.2	The Evolution of the Web	3
8.2.1	The Initial Vision for the Web	3
8.2.2	Web 1.0	4
8.2.3	Web 2.0	5
8.2.4	Web 2.0 Business Models	6
8.2.5	The prospects of the future internet: Web 2.0 Ambitions	7
8.2.6	Introducing Blockchain	7
8.2.7	Web 3.0	8
8.2.8	Web 3.0 Landscape	9
8.2.9	Web 3.0 Business Models	11
8.3	Economics of Blockchain	12
8.3.1	As an Institutional Technology	14
8.3.2	Harnessing Trust	15
8.3.3	Decentralization of Economies	17
8.3.4	Implications in the Fourth Industrial Revolution	18
8.3.5	Institutional Economic Evolution	19
8.4	Economic Challenges and Limitations with Blockchain	20
8.4.1	Scalability	21
8.4.2	Privacy	21
8.4.3	Security	22
8.4.4	Issues with Regard to Government Regulation	22
8.4.5	Challenges Affecting the Business Model	23
8.4.6	Interoperability	23
8.5	Conclusion	24

8.1 Introduction

The World Wide Web started out as an idea for a decentralized network, where people could communicate and collaborate freely, without any entity being able to prevent people from doing so. Over time, major parts of the Web became centralized - centralized in the sense that few central authorities take large portions of the profits, while being able to predict the behavior of the general public, and, to some extent, even influence it in a controllable manner. [3] That's the Web as we know it today. Web 3.0 is a new wave of networking technologies, that tries to reclaim some of the power to the users of the Web, by using decentralized protocols and applications, such as blockchains for example [3].

To understand the potential placement of blockchains in our modern world it is necessary to understand some parts of capitalism, our current *system of the world*, as one of many possible mechanisms aiming to achieve an efficient state of economics. By exploring the main incentives of the capitalist machine, hence profit extraction from knowledge accumulation and the improvement of learning curves, we uncover some inherent limitations that may deem the current capitalist model and the pool of possible business models derived from it unsustainable in the long-run. As we'll try to explain, blockchains may allow to push beyond those limitations by means of alternative information scarcity, information ownership and business model structures that allow for the formation of fundamentally different economic organisations, better tuned to face the societal, environmental and economical challenges of our times.

The key to understand blockchains is not just through an information, communication and technology perspective but through an economic institution perspective which allows for a new type of coordination mechanism. If we follow through a neoclassical perspective on blockchain, we might consider it as a general purpose technology and will model it through a shift in total factor productivity curve. However, in this paper, we extend Davidson's [29] argument that we need to think about blockchain through an institutional economics perspective because blockchain is not just another technology; it offers a new type of decentralized coordination mechanism that competes with markets, firms, and relational contracting. Moreover, we also briefly discuss as to how this decentralized immutable peer to peer network can help fight challenges like climate change and air pollution.

But blockchain is not bullet proof and like any other technology, it faces issues and challenges for example: scalability, interoperability, and privacy etc [43]. These problems not only appear from a technical background but can also be viewed from an economical standpoint.

8.2 The Evolution of the Web

In this section we attempt to timeline the various transitions the web has gone through as well as the different prospects the web has in store for its future. Fundamentally we look at the Web 1.0, Web 2.0 and Web 3.0 from an architectural perspective and a business model perspective.

8.2.1 The Initial Vision for the Web

When working as a software engineer at CERN, Tim Berners-Lee noticed the difficulties the scientists had, to get access to information that was stored on a different computer. They usually had to log on to the computer and, since people didn't use standardized software and protocols, they had to learn to use different programs on those computers [54]. Berners-Lee designed a solution for that problem, which would build on the internet - that solution is now known as the World Wide Web. He developed HTML (Hyper Text

Markup Language), URI (Uniform Resource Identifier) and HTTP (Hypertext Transfer Protocol) in order to make his solution work. Further, he developed a web browser to view the documents, and a web server to distribute the documents.

In order to make the web unleash its full potential, Berners-Lee decided to make the technology available for free and allowed using it without permission. Berners-Lee points out that one cannot propose something that is supposedly a universal space, a public utility, but keep control of it at the same time [56].

The web is fully decentralized, there is no central controlling node, no single point of failure and there is no way to shut down the web. No one needs permission from a central authority to publish something on the web. These properties are supposed to prevent arbitrary censorship and surveillance. Another idea that emerged early was that, since the internet is considered a common carrier, ISPs (Internet Service Providers) must treat any content equally. That concept is today known as Net Neutrality.

Another concept that evolved early is consensus. It describes the fact that all computers, that want to interact in a network, must agree to use a common language and to adhere to common standards. This does not matter on the hardware, neither does it depend on the location, nor on political and cultural beliefs of the people.

8.2.2 Web 1.0

Initially, the web was a collection of static HTML pages connected together through hyperlinks without any interactive content. Most of the original web wasn't even indexed by search engines, such as Google, and websites had to be found through open directories. One major technology that emerged in Web 1.0 was the electronic mail (E-Mail). The number of peer to peer interactions was low, websites were static and mostly used for unidirectional communication and delivery of information, e.g. articles, home pages and ads. For this reason, Web 1.0 is sometimes referred to as the "Read-Only Web". Content was produced by a minority of web users, since publishing any kind of content on the web meant having to maintain a web server.

Websites were hosted on servers, which were mostly provided as dedicated services by companies, called hosts. The user usually got a directory, where he could put the files for his website, which the host would then serve to the visitors. In the early stages of the web, hosts didn't support server-side scripting, which is why the web necessarily consisted only of static pages. Since the content could not be adapted according to the person who is viewing the website, the web was mostly informational. Websites in the early 2000s, web browsers evolved to support client side scripting using JavaScript, which was one major step towards the web as we know it today.

In the early times, instead of forms, which you could fill out on the web page directly, usually you could find a link with an E-Mail address instead, which would open your E-Mail client once you clicked it. As the web evolved and hosts started to support server-side scripting, more and more interactive elements started to appear. Besides contact forms, an example for such an interactive element of a website are "Guestbooks". A guestbook was one distinct page of a website, where viewers could leave a comment for the owner of the website. This evolved over time to the "Comment Section", that is usually placed beneath content nowadays.

This evolution, where users would become more involved and could interact with owners of websites, as well as with other users, lead to the Web 2.0, the interactive web. However, Tim Berners-Lee, the inventor of the web, does not agree with the distinction of Web 1.0 and Web 2.0 based on interactivity, since collaboration and interaction were the main goals of the web in the first place [57]. We still use that distinction, since it has become widespread.

8.2.3 Web 2.0

The web that we know today, Web 2.0, is the outcome of an evolution of the Web 1.0 to an even more interactive network. The rise of services and social media platforms like YouTube, MySpace, Facebook and Instagram increased the interaction of peers dramatically, through a phenomenon known as network effects. Web users can now upload videos to the internet, that can be viewed by other web users. They can write comments and have discussions, and they can even have personal pages without having their own web server. Supported by the development of dynamic and interactive web programming languages like JavaScript, websites are no longer just static. As a consequence, even interaction between web services and humans was made possible.

Web 2.0 is a collection of technological standards, that describe the web as being the “web-as-a-platform” [68]. This is because prior to Web 2.0, applications were developed to target the underlying operating systems of end users [69] instead of being ubiquitous by ways of utilizing the modern standardized web-browser.

What is the role of the web as a component of our modern world? We could begin by understanding the current state of the web, the Web 2.0, together with its underlying computer hardware and software architecture and network infrastructure as a piece of the engine that powers our system of the world [37], which is capitalism.

It is important to separate capitalism from economics. Economics has persisted across time as the age-old school of thought concerned with studying how we humans deal with the need to measure and distribute resources on a finite planet. Persisted across time because although the subject has become more precisely mathematicised, its core tenants have remained the same. That is to say the measure and distribution of scarcity. Not to mention the nomenclature of economics has changed somewhat over the years. Capitalism on the other hand is arguably one of many applied mechanisms, or systems employed to achieve an efficient state of economics.

Previous systems of the world have existed in the past. For example, in the eighteenth century, Isaac Newton introduced calculus and physics that rendered the physical world predictable and measurable [38] along with another system, the “[...] gold standard, which made economic valuations as calculable and reliable as the physical dimensions of the items in trade” [38]. Newtons gold standard expanded the possibility of achieving a more efficient state of economics. For the first time, long-term financial commitments in the form of bonds, loans, investments, mortgages, insurance policies, contracts, ocean voyages, infrastructural projects and new technologies could proliferate without fearing inflation fuelled by counterfeited money. [38]

In capitalism, the factors of production are privately owned [80]. The stakeholders attempt to extract a profit from the operations that their machines do. Taking into account the collapsing prices of computer storage [71], bandwidth [72] and processing power [73] within an economic order where the main role of the machine is to produce and the main role of people is to supervise them, it becomes clear that the main productive force is information, knowledge and organization as opposed to the work of making and running the machines [65].

This is due to on the one hand, as mentioned previously, in the falling costs of producing and maintaining the machine. But on the other, it is because the essence of growth, specifically but not restricted to economic growth, is learning and knowing [39]. Hence, the acquisition of knowledge is growth. As renowned economist George Gilder puts it, “information, not the management of processes, creates economic growth” [66].

So what is the role of the computer and the web in capitalism? By understanding from previous arguments that knowledge is the precursor to economic growth and productive growth in capitalism, we can assume that the role of the computer is to help quantify knowledge and cluster information beyond the limitations of paper bookkeeping, but

more specifically the limitations of our brains. A main contribution of the computer is the amelioration of learning curves, while that of the network is amelioration of communication of knowledge. Consider how the price of ferromagnetic memory cores for computers reduced from 5 cents per bit in 1965 to less than a half cent in 1973 due to learning curve improvements [67].

We don't need to search very far to validate the rationale that the accumulation of information as a precursor to improving learning curves is the primary force for economic growth. Currently spearheading our capitalism system is Google. What does Google do? In their own words: "Our mission is to organize the world's information and make it universally accessible and useful" [74]. Google acquires and structures the world's knowledge. Facebook does the same thing, only that it specializes in people and communities. But contrary to popular belief, these tech giants don't operate for free. In fact they are powered by capitalistic business models that may be becoming inappropriate to a world full of individual minds as we'll attempt to explain in the next section.

8.2.4 Web 2.0 Business Models

The purpose of this section is to describe some of the most successful business models currently in use within the web 2.0 landscape. We analyse the most important ones and assume that other models, differentiate only marginally and generate lower revenue. This is because of the fundamental notion that the harvesting of information is the golden snitch for powering business growth. Simply put, whoever holds the majority data.

The likes of Google and Facebook operate like any other capitalist business. They have costs to cover as well as the need to create growth in the form of monetary profits for their stakeholders. Their main method in which they achieve profits is by selling advertising real-estate to third parties [75]. Their ad placement campaigns have near perfect relevance to viewers because these tech giants have access to knowledge that most competitors don't. They exclusively generate and learn from enormous amounts of private user behaviour and user preference data sets from the free use of their digital products [76]. Due to the nature of these technological solutions in place, hence the current architecture of the web today, these tech giants have a monopoly on the world's information.

This monopoly on the world's information enables these institutions to sell for a profit because under their private ownership, this information, specifically user-centric information, is considered scarce and it can subsequently be commoditized. The underlying principle for this mechanism is the use of information scarcity to exploit market information inefficiencies in the form of information arbitrage [77]. Suppose all parties in a market knew the future preferences and future tendencies of a populace, the market would be efficient and information arbitration opportunities wouldn't exist. If a driver knew directly to satisfy a traveller's need to travel, the driver nor the traveller would need Uber.

The premise for information scarcity is that information is expensive to produce, to store and to transmit. However, this has become less and less the case. Consider the evolution of information technologies that render information less scarce: the printing press in 1450, the encyclopedia in 1728, the typewriter of 1867, the hyperlink in 1968, the Apple computer in 1976, the Ethernet network in 1979, HTTP, the graphical web browser, Google (formally Backrub) and Wikipedia [78]. We have continuously evolved towards information abundance, but one is propelled to ask: at what cost comes this transition? Is it the case that we've made a trade-off between information authenticity for information abundance? For quantity over quality? It would be an understatement to say that we're in a state of information overload as well as a state of dis-information. The quantity of unauthentic information swimming on the internet is a cause for concern. For example, a hoax circulated online in 2017 that the Ethereum founder had died in a car crash. Subsequently, the Ethereum market cap dropped by 4 billion dollars. [88] Blockchains

may play a role in both evolving the information scarcity model without compromising information authenticity.

8.2.5 The prospects of the future internet: Web 2.0 Ambitions

Naturally, the intuition for the current establishment is for the ball to keep on rolling the way it is. That is to say that the perpetuation of data collection into privatized data silos via a myriad of free and highly addictive digital products, regardless of social welfare, mental health and information authenticity. Stemming in this direction together with the incentive forces of improving learning curves as the principal driver of economic growth, institutions of the current web are naturally propelled to hail big data, machine learning and AI as the next big evolution web services [38]. The term *machine learning* fits hand-in-glove with the previous notion expressed, that learning is the essence of capitalist economic growth. It is intuition to assume that this is the direction to take for computing and the future of the web while still confined to a capitalist society.

Due to the exponential increase of computers of different types and for different purposes, there has been an equally important explosion of data sources.

“These are data sets whose size expand beyond traditional relational databases to capture, manage and process the data with low latency.” [90]

Machine learning is nothing new. It has been around since the 1950s when the model was created by Donald Hebb in his book titled *The Organization of Behaviour* in which he expressed the theories on neuron excitement and communication. [89] These theories helped the development of machine learning algorithms that today paired with the massive amount of data sources make the sector a feasible breeding ground for value added services in business and throughout the economy. However there is some compelling criticism against big data, machine learning and the rise of artificial intelligence.

In the 1930s, Kurt Gödel at the age of 23 demonstrated that there exists inherent limitations on every formal axiomatic system, necessarily, dependencies that exist outside of the system [79]. Proving this, he indirectly laid the foundations for computer software. Alan Turing used Gödel’s theorem to establish the universal computer architecture, hence Turing machines, which form the basis for all modern computing architectures today. Unbeknownst at the time, Gödel’s theorem proved that every computer system under the current computing paradigm is dependent on an outside oracle, the human programmer [38]. Necessarily, computers are not excluded from this understanding. They will always be dependant on the human programmer, waiting to receive instructions.

From a societal-economic prospective, the implementation of these types of systems raise ethical concerns that threaten our democracy. In her book *Weapons of math destruction* [48], Cathy O’Neil puts into question how the use of big data and algorithms reinforce preexisting discrimination and inequalities. When algorithms are increasingly replacing humans in decision-making, who is to be made responsible if the models are incorrect or if the data fed into the models are biased to begin with?

8.2.6 Introducing Blockchain

The blockchain was introduced as a solution to a class of problems in distributed systems, namely the byzantine generals problem. It’s first and most famous use case is the enabling of digital currencies that don’t require a centralized intermediary to settle transactions [44]. But it’s proponents suggest that the blockchain has far greater reaching consequences that could shake the very foundation of the current economic establishment [81]. Specifically, blockchain can serve as a mechanism to re-decentralize the web, laying the foundation for a new web security architecture in which security is wrapped around the individual user in the form of private-public key cryptography, as opposed to security being

externalized and centralized [38]. Additionally, the blockchain enables the creation of a new class of economic institutions, decentralized organizations and governance structures. As a mechanism to re-decentralize the web? The web has become too centralized. In a centralized web structure, information is easy to censor, easy to hack, easy to manipulate and it's easier to conduct surveillance. It also poses a single point of failure. There is growing misalignment with the web's original design to be a free and open set of protocols [82]. Blockchains provide the ability to track ownership of data securely in an environment that doesn't need a centralized party. Two projects pioneering this idea is Blockstack [83] and Solid [84] (See current landscape of Web 3.0 projects).

Using the blockchain as a new foundation layer for value exchange, the creation of new types of economic institutions and the second business model of the internet is possible. Blockchains introduce the general adoption of the era of peer-to-peer. Although peer-to-peer is not new, blockchains enable the ability to launch fully-fledged peer-to-peer networks as a new breed of economic organizations that are spontaneous and self-governing [86]. In these peer-to-peer economic organisations, wealth creation is communal and production shifts to a commons-based approach, hence commons-based peer-production [40]. The essence is a shift from extractive to generative business models [40]. When we consider a business as being a "nexus of contracts" [41], we could hypothetically create an entirely decentralized business with smart contracts deployed to a blockchain ecosystem. The general name for this concept is a decentralized autonomous organization or DAO. This movement towards decentralized systems is what is known as the Web 3.0.

Another forefront actively using blockchain technology is DeFi. The Decentralized Finance Movement proposes an alternative to the global financial system. It considers that the centralized nature of the currently implemented financial system has created an imbalance of wealth distribution. Populations that are closer to the world's leading financial hubs have better opportunities to become richer and the *trickle-down* mechanisms of distributing wealth and resources is not actualizing itself, especially in developing economies. This is at the expense of financially marginalized populations, taken into consideration that the current financial system is a zero-sum game. The DeFi Movement offers borderless inclusiveness through permissionless, decentralized, trustless, transparent, censorship resistant computer networks.

8.2.7 Web 3.0

The problems we see on the internet resemble the problems we are facing in society. Bad structures, like monopolies and totalitarian states and systems, are reflected on the web - even more vigorously than in the real world, since the digital world has less boundaries and resistance by being connected through the internet. We rely on powerful intermediaries like banks and corporations in the real world, but also on powerful intermediaries that provide services online, such as Google, Facebook and Amazon. Projects usually start with good intentions and great missions, but over time, enterprises tend to become centralized and they need to be profitable to survive.

Trust is needed where there is uncertainty. If something is unconditionally certain, there is no need for trust - it just becomes a fact. Since we don't have the means to ensure corporations follow the rules, we have to rely on the promises they make, and we have to trust them. Trust is not a bad thing. We might for example trust Google and Facebook on their ability to provide availability and on their ability to make precise predictions, but we might not trust them, that they won't try to influence our decision making. In today's world, trust is too holistic, we need mechanisms to break down the required trust into more granular pieces.

One of the major issues that we are facing, is the fact that the tools we use for communication, interaction and decision making are flawed, since we use social media platforms

like Facebook, communication services like Whatsapp, and search engines like Google to reduce the uncertainty. Those services are provided by the very entities, that we are trying to inspect. When asking ourselves, why that is the case, one answer might be that these services are free, another might be because of network effects.

The architecture of the web was not made in a way that makes the individuals of the network stay in control. Security is not built from the ground up, in other words from the individual outwards. Today, most parts of the web have become centralized are closed, and thus we do not have insights on what is happening with the data we are producing. Data in this context is a broad term, and it can consist of the messages we are exchanging, the information we are searching, our preferences and our behaviors. Service providers collect this data and store it, in order to analyze and predict the behavior of their users. The best tool we have to oppose this is regulation, with which we can try to steer corporates in a certain direction. One recent way of regulation involved the use of open protocols, through which emerging enterprises and the general public gain the ability to participate in established environments. An example is the banking sector, where the European Commission put in place the Payment Services Directive (PSD and PSD2), with the purpose of increasing participation of non-bank companies in the payment industry [53]. Web 3.0 is a movement that tries to solve some of these problems of society by using technology. If technology could enforce certain behaviours through mechanism design, we could have less regulation and we could get by with less trust in corporations, since there would be more certainty in the ways they operate and hence we would trust them more. An example for such technologies are blockchains. Although blockchains at present are mainly used as a payment mechanism, they have much more potential. They could be used to build a platform, where rules can be hard-coded and enforced. On such a platform, no single entity could change the rules arbitrarily, as it is possible today with textual *terms of service* agreements.

Think of the Bitcoin blockchain as a simple calculator on a distributed ledger. The ledger with details of the accounts, i.e. owner and balance, is stored on each participating node. The participants execute movements between accounts and agree on a new state of the ledger by using a consensus mechanism. The goal of Web 3.0 is to have a distributed computer, that can do much more sophisticated operations than just addition and subtraction. That distributed computer would act in a transparent way, every participant can inspect how the computer works, but at the same time no single participant would be able to change the way it works without the consent of the other participants.

The Web 3.0 should be fully decentralized, where the endpoints should stay in control of data and computation, and where contracts are written in code to ensure precision and enforcement. The divide between users and service providers should be decreased drastically, and everyone should operate on a joint infrastructure. Users should not have to put as much trust in contracts, instead contracts should become predictable and immutable, so that trust becomes more granular. The web should be turned into what it was supposed to be initially: The web should serve humanity as a public utility.

8.2.8 Web 3.0 Landscape

“Web3 is a broad movement and an inclusive set of protocols aiming to make the web and the internet more decentralized, verifiable, and secure. Web3 is the vision of the serverless internet, the decentralized web. An internet where users are in control of their own data, identity and destiny.” - Web3 Foundation [52]

The fundamental adherent of web 3.0 is to counter existing centralized internet corporations that are considered “too big to fail”. The idea is destined to harmonize the current power balance of the web, bringing it back to its original intentions as a free and open set of protocols that empower the end user. That means building software and services that

shouldn't be centralized in nature or privately owned, but rather serve as a decentralized public utility.

“This movement isn't just about blockchains (not everything needs a blockchain!) - it is about architecting a web that protects individual property and privacy through a range of p2p technologies.” [3]

In the previous chapters we attempted to introduce the notion that the blockchain presents a new way to structure economic institutions. However, this notion necessarily needs to be understood from a perspective other than from a capitalist standpoint because blockchains allow the automation and challenge a lot of the main principles of capitalism. In this section we outline the initiatives pioneering new types of economic institutions as well as a class of institutions that are attempting to migrate existing digital services.

The initiatives below are concerned with implementing new and pioneer web architectures.

8.2.8.1 BackFeed

BackFeed [85] develops a distributed governance system for blockchain-based applications allowing for the collaborative creation and distribution of value in spontaneously emerging networks of peers.

Inspired by these stigmergic principles found in nature, Backfeed develops a decentralized protocol that relies on blockchain technologies to provide an indirect coordination mechanism for people to collaborate and cooperate, without the need for any intermediary authority or centralized agency.

8.2.8.2 DAOstack

DAOstack [86] is an operating system to build Decentralized Autonomous Organizations that require efficient governance of self-organizing collectives. The basis is that historically humans have always needed to collaborate in large groups, after all we're social creatures. Top-down, hierarchical structures have almost always been the norm. It's easier to point the ship with one leader, one captain or a small group of navigators to make decisions for the many, as it is visibly the case in most corporations, governments and militaries across civilizations today. The problem with top-down structure is the tendency for single points of failure, bad acting, simply bad judgement or interest misalignment with the many being herded. As the excitement around decentralized organization arises, DAOstack aims to provide a technical implementation of these ideas.

8.2.8.3 Polkadot

Polkadot [87] is an architecture solution to allow distinct blockchains to become interoperable with the help of specialized sidechains. One of the first problems to arise in the blockchain ecosystem is the communication between distinct private, public, consortium, permissionless and permissioned blockchains. It would be infeasible to migrate the internet to blockchains that are distinct and separated. Polkadot aims to increase blockchain seamlessness.

8.2.8.4 Blockstack

Blockstack [83] proposes a blockchain based decentralized computing network in which users entirely control their data. Through the use of private-public key (asymmetric) encryption users rely on a unique identity that entitles them to the data they generate when using Blockstack based decentralized applications.

8.2.8.5 Solid

Solid is a project of Tim Berners-Lee, the inventor of the Web, and it is similar to Blockstack in that users own their data generated by the use of decentralized apps and dedicate their storage to securely specified decentralized repositories on the network or to a device owned by the user [84].

The idea is to restructure the way applications are constructed and services are provided to the user. In today's world, the data generated by users is stored by the service providers. They use this data to improve their services, or they use the data to provide entirely new services, that depend on the analysis of the data. However, the data can be used to exploit users too. The concept of *Data Ownership* aims to counteract this vulnerability by changing the architecture.

The proposed solution is to store the data near the user, which means, the user should be able to decide how and where his data should be stored. Next, there will be an interface to interact with the user's data: Applications can have access to parts of the data or the entire data of the user. However, the applications must not be able to communicate with the outside, and this has to be ensured by the infrastructure. Such applications are called *Decentralized Applications*, or in short *DApps*.

For application developers and service providers this new infrastructure would pose both opportunities and challenges. Since data is stored at the user and since the data can never leave the storage, we could allow applications to have access to a lot more data than today. This would allow to much more precise analytics, while also eliminating concerns regarding privacy. One major drawback for service providers is the fact that aggregations over populations is no longer possible, since data can never leave the user's storage.

8.2.9 Web 3.0 Business Models

During its evolution the web went from being a "Read-Only" network, used to distribute information and advertisement, in fact an extension of physical shop windows, to an interactive, "Read-Write" system, where people can collaborate, as well as create and publish content. However, these existing services and business models come with some disadvantages, namely, we become so dependent on powerful companies, that we cannot even check, if they are acting maliciously. As mentioned before, this is due to the fact that the tools we use to inspect such powerful entities, are produced and offered by those entities. As a response to this trend, some Web 3.0 initiatives try to hand the power back to the user.

8.2.9.1 Web Browser

Brave is an open source browser, which is based on Chromium and aims to change the way we interact with the web. It wants to replace current browsers such as Chrome, Safari and Firefox, that were built for Web 2.0. As a main feature, it blocks ads and website trackers, and proposes a business model regarding ads: Users should get paid in order to surf [58]. They have created an Ethereum based token called BAT (Basic Attention Token), which the users of Brave could earn while watching ads. Users would be able to tip BATs to content producers, while ad buyers could buy views for BATs. To reduce tracking, in order to increase privacy, ad targeting is done on the client-side [59].

Brave earned critique and was declared to be a "double dip" [60]. The issue the critics are raising is the fact that Brave blocks ads, but then replaces them with their own ads to generate revenue. This is deemed to be immoral and is being discussed controversially.

8.2.9.2 Storage

P2P file sharing protocols have been around for some time, and BitTorrent is a prime example thereof. BitTorrent can be used to reduce the load on the server and the network while distributing large files. This is achieved by splitting a file into small chunks and distributing those chunks. Once a peer obtains a chunk, it can too start distributing the chunk.

IPFS (InterPlanetary File System) is a peer-to-peer network protocol that describes a distributed file system, which can be used to store and share data. It aims to replace the HTTP (HyperText Transport Protocol) and centralized servers, which host for example HTML files, pictures and movies, by storing files in a decentral manner. Peers will then distribute these files in a similar fashion as in the BitTorrent network. IPFS is designed to be an immutable storage solution, that keeps track of versions too [61].

However, IPFS can be used in a simpler way too to replace cloud storages like Dropbox, Google Drive, OneDrive or Box. Another competing project in this realm is called *Tardigrade* (former name was *Storj*). Tardigrade has a less ambitious goal, as it only aims to provide a distributed cloud storage solution, which can be used as a drop-in replacement for Amazon S3. As a user, it's also possible to earn money by offering unused disk space to the network [62].

8.2.9.3 Messaging

Instant messaging is another field where we can expect innovation through Web 3.0 technologies. One pioneer regarding this domain is *Status*, which aims to replace Whatsapp, WeChat, LINE, Viber and co. Status combines messaging, crypto wallet and Web 3.0 browser to one application. Their mission statement claims that they want to provide a secure and privacy preserving communication tool, so that we can preserve freedom of speech and human rights, while preventing censorship and surveillance. The messenger uses a peer-to-peer protocol called *Whisper* to route the messages, and end-to-end encryption to ensure confidentiality. Since the project is open source, users have the option to verify their claims [63].

8.2.9.4 Social Media

Similar to the domain of messaging, there are new projects that want to disrupt the way social media works. One such project is *Steemit*, which wants to replace Facebook, Twitter, Instagram and the like. Steemit believes that users of a social network should be rewarded for their attention, voting, sharing their voice through comments and producing good content. In contrast, in current social media networks, the profits go to shareholders. Steemit is based on the Steem blockchain, and the rewards for the users are paid in STEEM, the corresponding digital token [64].

8.3 Economics of Blockchain

If one tends to wander outside the technical debate and the crypto sphere of Blockchains, with Williamson's (1985) [33] 'economic institutions of capitalism' - viz, markets, hierarchies and relational contracting - in mind, he / she will immediately come to the conclusion that this crypto mechanism with no hierarchical model adds onto Williamson's list. This new type of economic order which guarantees a decentralized immutable and distributed ledger, has implications far beyond the current cryptocurrency fad.

The purpose of this rather elaborate section is to add flesh to the economics of blockchain and not tackle the information communication and technology (ICT) perspective. It is

to enable us to comprehend it through the lens of evolution of institutions, governance and organizations, and its role in the Fourth Industrial Revolution. Economic theory can be used to examine the invention, adoption and the use of this new technology. From an analysis of a technology adoption process, blockchain is in the early disruptive phase of Schumpeterian process of creative destruction that will likely unfold along a logistic adoption-diffusion trajectory [30]. The primary nature of blockchain technology is different from the conventional way of how economists model new technologies, namely as a shift in the aggregate production function, that translates into multifactor productivity growth. However, in the institutional/evolutionary approach, technologies aren't just limited to ICT's but also include 'social technologies', as rules of coordinating among people. In this social technology approach, technological change is a change in institutional efficiency as it opens up new and efficient methods of governance and coordination systems which lowers transaction costs. However, in the neoclassical approach, technological change lowers production costs.

The benefit of introducing key technologies in economies shifts the total factor productivity and reduces marginal costs. People adopt the new technology because of these marginal productivity efficiency gains. If we follow through with the neoclassical approach and categorize blockchains as a General Purpose Technology (GPT) then we see that it makes existing factors of production more productive by reducing the cost of production. It economises on scarce resources [29] But following the new/institutional approach, we can as well argue that blockchain is a social technology which gives rise to new organizational and institutional forms of economic governance. So, the most favourable framework to look at blockchains would be the new institutional economics (NIE) also known as Transaction cost economics (TCE). NIE originates from Ronald Coase's work on transaction costs where he compares the neoclassical and institutional ideology by stating that while economizing on production costs lead to an efficient allocation of resources, economizing on transaction costs however, lead to an efficient institutional structure of economic organization and governance.

Ronald Coase's [15] [16] in his explanation behind the existence of firms and the existence of the law argued as to why some transactions happen in firms (hierarchies) rather than in markets? The answer to this question is that in order to deal with uncertainty, asset specificity, and frequency of dealings, some transactions are conducted more efficiently in hierarchies rather than markets (Williamson 1979, 1985) [32] [33]. Transaction costs thus determine the efficiency of different government structures. This basic insight to TCE can bring blockchains and ask the same but now extended question: Why do (might) some transactions occur in blockchains, rather than in firms or markets?

The key is to understand that the existence of hierarchical organization is to control opportunism and rent seeking which ultimately gives rise to transaction cost. The valuable prospect of blockchain (as smart contracts and DAOs) is precisely to eliminate opportunism through crypto-economic mechanisms which enable a trusted market to carry forward transactions in perpetuity. A fair lot of Davidson's [29] research revolves around the discussion carried forward from Williamson that blockchains are a new type of coordination mechanism which compete with hierarchies, relational contracting and markets. The following subsections will carry forward this discussion of (1) blockchains as an institutional technology (2) how blockchains are trust machines rather than a trustless mechanism (3) the mechanics behind decentralized structures achieved through blockchains (4) how it affects the fourth industrial revolution (5) blockchains and the impact on institutional economic evolution.

8.3.1 As an Institutional Technology

The transaction cost approach to blockchain leads us to new institutional economics framework where we see that, organisational form is shaped by the need to control opportunism [33] (Pages: 64-7). There are three ways to think about the blockchain technology; one is to look at it through the neoclassical lens and think of it as a general purpose technology (GPT), second is to take the market enhancing view and think of it as an exchange technology, and lastly, is to look at it through the new institutional economics lens and think of it as an institutional technology.

The first is that blockchain is a general purpose technology, meaning that it is expected to have broad transformative application across many sectors of the economy and contribute to multifactor productivity growth [8] [14]. This perspective, whether stated implicitly or explicitly, underpins the case for hype surrounding the prospects of blockchain technology as an ‘engine of growth’. A second perspective places a different emphasis on the way in which the arrival of blockchain technology might impact the economy by viewing it through a Coasian, rather than a Schumpeterian, lens. Along this line, Catalini and Gans [9] portray the ‘simple economics of blockchain’ as the analysis of a new technology that lowers transaction costs through costless verification and without the need for costly intermediation, which they suggest will improve the efficiency and scope of markets, moving them closer to a direct peer-to-peer ideal. This distinction comes down to whether the blockchain is understood to contribute to production technology (the general purpose technology view) or to exchange technology (the market-enhancing view). We follow through with Davidson [29] in agreeing that blockchain is neither a production, nor an exchange technology, but is better understood from the economic perspective as an institutional technology.

Blockchain is the technology that underpins Bitcoin, the first successful cryptocurrency which was ever started. The breakthrough was the creation of a distributed ledger, such that each node in the network has a copy of the ledger, and there is a mechanism - a cryptographically secure and crypto-economically incentivized mechanism - to ensure consensus about the true state of the ledger without the need to trust a centralised node or authority [29]. This has been the revolutionary innovation in how we keep our ledgers. These ledgers have not greatly changed since double entry bookkeeping was developed in the Venetian Republic in the 15th century. By the late 20th century they have been digitized, but until the blockchain, invented in 2008, they always remained centralized. The ledger is a technology of accounting, of keeping track of who owns what, and is instrumental to modern capitalism (Nussbaum 1933, Yamey 1949) [21] [36]. But so too is trust in the ledger, which is most effective when it is centralized and strong, and so centralized ledgers for property titling, contracts, money, etc, are also critical in connecting government to modern capitalism. Centralized solutions are expensive in the same way that governments are expensive, and have many problems, particularly in relation to problems of trust and its abuse. Yet until very recently no effective decentralized solution has existed. In a recent lead article on blockchain - which they dubbed ‘The trust machine’ - The Economist (2015) [55] explained that:

‘Ledgers that no longer need to be maintained by a company - or a government - may in time spur new changes in how companies and governments work, in what is expected of them and in what can be done without them.’

The control of centralized ledgers undoubtedly bring in the facet of opportunism and rent seeking behavior which gives rise to transaction costs. Wiles argued that these decentralized ledgers which are now technically possible will eventually compete on cost with the centralized ones because they run down three exponential cost curves: (1) Moore’s law (cost of processing digital information, i.e. speed, halves every 18 months); (2) Kryder’s Law (cost of storing digital information, i.e. memory, halves every 12 months); and

(3) Nielsen’s Law (cost of shipping digital information, i.e. bandwidth, halves every 24 months) [31].

We learned from Williamson that because of asset specificity, frequency of transactions, and uncertainty, transactions happen in markets and firms, but one critical aspect to understand over here is that this economic efficiency of hierarchies (such as firms), rest on a nexus of contracts, but specifically as a nexus of incomplete contracts [12] [13] [33]. Blockchains refer to a particular class of economic system that Coase [15] taught us to see through the lens of contracts: namely a blockchain is an economic world of complete contracts.

If blockchains can eliminate opportunism, then they will outcompete traditional organizational hierarchies and relational contracts. (How do blockchains eliminate opportunism? In essence, by radical public transparency coupled with crypto-consensus mechanisms, executed automatically with smart contracts.) Blockchain enabled smart contracts facilitated transactions should face less of the efficiency problems of information asymmetries - moral hazard and adverse selection. These machine readable contracts are bound to reduce transaction costs. This assault of authority (through centralized governance) can be diluted through blockchains if we can achieve the transition from incomplete contracts to complete contracts. Thus, the institutional lens of Ronald Coase will favour us in setting this new form of institutional technology as a decentralized governing structure.

8.3.2 Harnessing Trust

We can start from the inception of this technology, which surprisingly was just over a decade ago when Nakamoto challenged the third party trust mechanism in merely eight pages [44]. In Nakamoto’s own words - in the conclusion section of his paper, he said: ‘We have proposed a system for electronic transactions without relying on trust’. His absolute focus was to solve the Byzantine problem without the mediation of a trusted third party in financial transactions. This can clearly be achieved as long as honest nodes hold more CPU power than any cooperating group of attacker nodes [44].

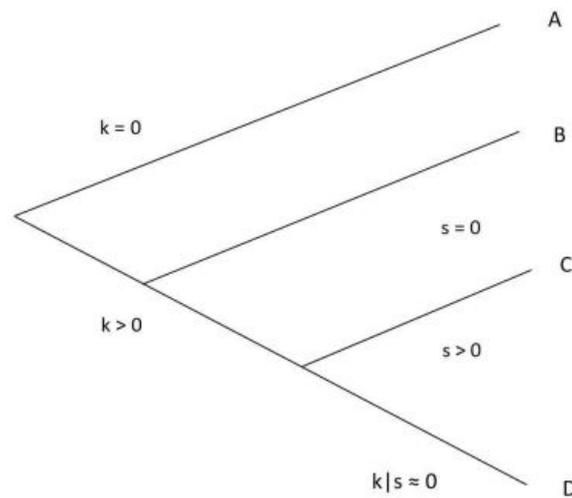
So, how do we harness trust from a system that was originally made to break trust. Under what conditions and in which circumstances can we harness trust from this system? Why is it that trusting a centralized authority is usually not preferred if such an option exists? Or better yet, trusting who? Do we feel safer if we divide trust among masses, rather than a centralized authority?

Blockchain is a decentralised computation technology for coordinating activity in a distributed economy [5]. This view follows in the transaction school tradition of Nobel laureates Ronald Coase and Oliver Williamson and sees the blockchain as a new type of economic institution that enhances (and competes with) the existing economic institutions of capitalism: firms, markets, commons, relational contracting, and governments [5].

Williamson has specified two behavioural assumptions that drive the contracting process; bounded rationality and opportunism. Bounded rationality relates to the fact that there are limits to human rationality. Opportunism is self-seeking with guile [5]. As Williamson (1985: 47) [33] writes, opportunism includes, “calculated efforts to mislead, distort, disguise, obfuscate, or otherwise confuse”, and as a result, “promises to behave responsibly that are unsupported by credible commitments will not, therefore, be reliably discharged”. Williamson [35] makes the argument that if parties to a contract promise engage in cooperative behaviour and those contracts were self-enforcing then promise is an efficient mechanism to facilitate trade. That sounds very much like what the blockchain and smart contracts (algorithmic contracts maintained and resolved on blockchains) can offer. Indeed, this could be what is meant when blockchain is described as being “trustless”.

We agree with Berg, Davidson and Potts that Blockchains are not trustless but trust machines. [5]. Berg applies Williamson's transaction cost analysis to the blockchain consensus mechanism and illustrates his argument regarding opportunism and the control of opportunism using a diagram similar to Williamson's figure below. In the first instance consider modes A, B, and C. In the diagram k represents an investment hazard associated with opportunism. If there were no opportunism then $k = 0$. In that instance contracts can be organised by what Williamson describes as being "competition" - contractual performance is easily observed and non-compliance easily corrected. In those instances where $k \neq 0$ then the question of contractual safeguards (s) becomes important. Consider the well-known market for lemons problem, in the instance that a used car salesman cannot adequately signal ($s = 0$) their trustworthiness (i.e. credibly commit to not defrauding the buyer) the transaction may not occur at all, or if it does occur will do so at a deep discount to true value. Of course, we well know that various mechanisms to safeguard transactions evolve ($s > 0$) ensuring that transactions do occur. These mechanisms, however, are costly and impose that cost on the parties to the transaction [5].

Figure 8.1: Opportunism vs Trust



One thing to note over here is that trust is usually described by economists as a mechanism to overcome opportunism. In the figure above, 's' is the solution to the problem 'k' and not the absence of it. What is rather interesting is that Blockchains are best understood through mode D. Now in this case we have a transaction which is associated with investment hazards due to opportunism and could most definitely take place in a non-blockchain environment. But this transaction could also take place in a blockchain environment. The blockchain technology incorporating proof of work (or proof of stake) implies that for the parties to the transaction $k | s \approx 0$. It is not that s overcomes problem k at some cost to the parties, but that blockchains suppress k at a cost to a third party (miners). This implies that employing a blockchain would be a preferred transaction technology to both B and C. This is the mechanism whereby blockchains can and will disrupt existing business models. The condition $k | s \approx 0$ is not an externality - miners are paid to validate and record transactions although they themselves are not party to the transaction [5].

The employability of Blockchains on any network where the rise in transaction cost is facilitated by the lack of trust, can greatly impact on how we do and think about businesses. The existence of opportunism can not be denied in any sphere of relational contracting or markets, but it can always be controlled at the expense of trust mechanisms. Blockchains guarantee with proof of work (or proof of stake) that the condition $k | s \approx 0$ will hold if honest nodes have more than 50% of computing power. This mechanism validates the

argument that the decentralized structure of blockchain distributes trust in a way that there is no centralized node with complete control or authority over the network.

8.3.3 Decentralization of Economies

As a general technology, blockchains facilitate decentralization. Gavin Wood and Vitalik Buterin from Ethereum describe blockchain as a technology that is converging on being a ‘world computer’, as a global singleton [51].

Contractual incompleteness is the origin of the study of economic organizations and governance because in a world with zero transaction costs, all contracts would be complete and all economic transactions would be market transactions. Incomplete contracting models [28] usually invoke transaction costs arising from: (1) uncertainty, or unforeseen contingencies; (2) costs of writing contracts; (3) costs of enforcing contracts. Uncertainty refers to information problems [30].

Now on the face of it, blockchain has a revolutionary implication because it undermines the strong case for the economic efficiency of hierarchies (which exploits incomplete contracts) and relational contracting (which requires trust between parties) over markets. As emphasized in the conclusion of the last section, if blockchains can eliminate opportunism, then they will most certainly outcompete traditional organizational hierarchies and relational contracts on the basis of distributed trust and lowered transaction cost. (How do blockchains eliminate opportunism? In essence, by radical public transparency coupled with crypto-consensus mechanisms, executed automatically with smart contracts.) But the most obvious problem is that blockchains only work on complete contracts, whereas most firms (cf. DAOs) are largely (entirely?) revolving around incomplete contracts [11]. We argued in the previous sections that this decentralized structure of blockchain can be viewed through the lens of new institutional economics and it should be modelled through not just a GPT but as an institutional technology. With radical transparency and crypto consensus mechanisms, this technology has a self governing architecture where trust is distributed amongst the participants of the chain rather than any one single intermediary or node. All these components combined provide the basic building blocks for the deployment of so-called Decentralized Collaborative Organizations (DCO), organizations that are not controlled by any given entity, but rather consist of a large number of individuals contributing out of their own free will to a common (collaborative) project. It is open to argument that such spontaneous and distributed collaboration already exists in the realm of open source software, where many developers collaborate towards the achievement of a common goal in a coordinated but decentralized manner. Perhaps, but open source software represents only a small part of modern society. A proper model for DCOs should enable decentralized large scale and systematic collaboration in potentially every sector of activity: from content creation to online gaming and networked communications, from fundraising to financial transactions, from corporate management to organizational matters, etc [30].

One such engine which is heavily discussed in the domain of blockchain and propagates the movement of decentralized collaboration is Backfeed. At its core it implements a Social Operating System for decentralized organizations, which enables humongous open-sourced collaboration without any form of centrality. As opposed to any open source model, in the case of a Backfeed enabled DCO, decentralized cooperation can be achieved in a way that is sustainable and effective over time. Contributions in a DCO are motivated by a specific system of economic and reputational incentives, and the resulting value produced by every contribution is shared among all collaborators through a specific evaluation protocol which lies at the core of the Backfeed protocol.

Imagine, for instance, a significant proportion of people writing books and publishing them in a decentralized fashion without any publisher or middleman; millions of people insuring

each other, without the absolute need of any centralized insurance companies; thousands of freelancers collaborating information together in a decentralized crowd based journalism organization, thousands of citizens coming together to form a decentralized real-time ride-sharing or park-sharing network; and millions of internet users contributing to a decentralized social search-engine. By combining blockchain infrastructure with Backfeed's distributed governance model, this vision is now beginning to unfold - eventually leading to a revolution in the way people work and organize themselves today [30].

While this all seems promising, we should remind ourselves that this is only possible in cases of complete contracts. The primary nature for the existence of firms is because transactions which carry high asset specificity, frequency, and uncertainty are often executed through incomplete contracts. The primary nature of blockchain is to work on complete contracts. While there is a progression in terms of transitioning from contracts that are incomplete to pure completeness, the velocity of such progress doesn't seem too promising. Nonetheless, the decentralizing power through DCOs DAOs and smart contracts open a new field of governing structure which embodies trust, is something that wasn't imaginable a couple of decades ago.

8.3.4 Implications in the Fourth Industrial Revolution

From an anthropocentric perspective, the past century (particularly in the past few decades) of human existence has marked a very successful period of population and economic growth [49]. As the Fourth Industrial Revolution gathers pace, with emerging technologies like the Internet of Things (IoT), virtual reality, Artificial intelligence (AI), and Blockchains, we see a basaltic convergence in the domains of economies, values, biodiversity, identities, and possibilities for our future generations.

The "great acceleration" [6] in human progression has delivered exponential economic growth. Real output grew five-fold in the four centuries leading to 1900, before accelerating more than 20-fold in the 20th century [7]. The follow on effects of this acceleration has recorded massive improvements in human welfare as the number of people living on \$1.25 a day has been cut by one-half since 1990 and more than 700 million people have moved into the global middle class [50].

Yet, the researchers from the World Economic Forum (WEF) identified that from the Earth systems perspective, the human success story is dismal. On the 40th anniversary of the first world climate conference on 5th-Nov-19, 11,000 scientists from 153 nations unequivocally stated that there will be untold suffering due to climate change unless there are major transformations to global society. Just over a year ago, in september 2018 WEF laid down six pressing environmental challenges where there is an opportunity to harness blockchain: climate change, natural disasters, biodiversity loss, ocean-health deterioration, air pollution, and water scarcity. While these challenges are grave and exceptional, they collapse with an era of unprecedented innovation, technical change and global connectivity - the Fourth Industrial Revolution.

Distributed computing and cryptography have both existed for decades but it was only in 2009 when these ideas merged in the form of Bitcoin; the enamoured cryptocurrency networks. As people delved deeper into the technology underpinning Bitcoin, they realized that it has the power to cut intermediary and reconciliation costs and revolutionize manual, frequently disjointed, opaque processes to increase their efficiency [47]. A brilliant example of this innovation is inception of Ethereum in 2015 (now a \$19 billion cryptonetwork) which showcased that blockchain is more than just a niche technology for the financial industry, but also offered a new, decentralized, trusted, and transparent platform that could benefit a much wider range of industries [47]

With the increase in traffic of blockchain coders, startups and industries have started to invest considerably in solutions which encapsulate blockchain as the underpinning tech-

nology infrastructure. This trend has clearly convinced investors, speculators and entrepreneurs to take an interest in this possibly disruptive technology. Advancements of such degree has led to the evolution of an environment which is conducive to blockchain technology. With added convenience of our smartphones it has made it increasingly relevant to carry our digital wallets. It is only because of such enabling global environment that there are various blockchain applications being launched.

Climate change has been one of the most crucial topics of discussion and debate in this generation. Through blockchains, clean power can be managed through peer-to-peer renewable energy-trading systems, optimized distributed grid management and decentralized authentication of renewable energy certificates. Smart transport systems can be initiated through managing a data ledger for optimized transport logistics, blockchain-based decentralized delivery networks, peer-to-peer vehicle sharing, smart parking system for optimized mobility management. To tackle air pollution through blockchains, four areas have to be targeted namely: clean air, monitoring and prevention of CO₂, early warning mechanisms, and clean fuel monitoring. Blockchains can help us in all four areas through automated air-quality monitoring system, early detection of toxic chemical leaks, cryptocurrency payments for EV public charging, and air pollutant data collation from distributed sources.

Similarly, various mechanisms powered by blockchains exist to preserve our biodiversity, ensure water security, to keep our oceans healthy and to build resilience against harsh weather. What is however disappointing are the recent investment efforts on these blockchain powered projects. For example, in the first quarter of 2018, 412 blockchain projects raised more than \$3.3 billion through ICOs. However, less than 1% were in the energy and utilities sector [46]. Nonetheless, the potential of blockchain too help solve these environmental challenges can be amplified exponentially when it is combined with other emerging Fourth Industrial Revolution technologies such as AI, IoT, drones, 3D printing and biotechnologies. When it is applied this way - as a “cocktail mixer“ for other emerging technologies - blockchain starts to become a truly game-changing technology [47].

8.3.5 Institutional Economic Evolution

When we think about blockchains being the building blocks of a new type of order then we automatically arrive in the domain of a decentralized collaborative organization (DCO). But before we delve into this topic, we need to ask a pressing question: What margin upon which blockchain institutions compete with alternative modes of economic coordination - markets, hierarchies and relational contracting [32] [34], as well as clubs, commons and government [23] [24] [22]. To answer that question, we need to follow through the institutional economics approach and think of blockchain as an institutional innovation. The relevant margin of analysis is therefore not the total factor production and growth, but rather the substitute mechanisms of economic coordination and governance that blockchain provides. Following through Williamson approach only, we have to understand the comparative institutional advantage of blockchains and the co-evolutionary dynamics with other institutions of market capitalism [29].

Blockchain-based distributed ledger technology adds an additional category to the suite of Williamson’s [33] ‘economic institutions of capitalism’ - viz. markets, hierarchies and relational contracting - with a new type of economic order: a decentralised collaborative organisation (DCO). A DCO is a self-governing organisation with the coordination properties of a market, the governance properties of a commons and the constitutional, legal and monetary properties of a nation state.

It is an organization which has the token systems (Backfeed) that coordinate distributed action, but it is neither hierarchical, nor it is a market because the predominant activity

is production, not exchange. And it has the unanimous constitutional properties of a rule-of-law governed nation state, by complicit agreement of all ‘citizens’ who opt in to such a decentralised collaborative organisation, and the automatic execution of the rules of that DCO through smart contract enforcement [4].

Thinking of blockchain as just an ICT actually misrepresents its nature as a technology. From a coordination perspective, its significance is as an evolutionary development in the institutions of market capitalism [19]. One path by which the institutions of market capitalism may adapt to blockchain technologies is through the substitution of economic governance from firms, markets, and relational contracts with blockchains. The same economic activity is institutionally reallocated. Currency transactions or settlement of financial trades move ‘to the blockchain’ for instance. But another path is that blockchains-based coordination may enable new types of economic activity that were previously not able to be governed by firms, markets or governments because the transaction costs were too high to justify the expected benefits. For example, Backfeed, [85] a social protocol that builds upon blockchain based infrastructure and the smart-contract platform provided by Ethereum, implements an alternative and more generic consensus algorithm called proof-of value that relies on human evaluation to discover the value of every contribution as perceived according to the distinctive value system of each individual network. Steem, [45] a blockchain-based social media organisation, performs a similar function though community-voting using its native cryptocurrency. Individual members of a community or organisation evaluate the contributions of others, who will be rewarded (according to the value they bring to the community) with economic tokens (transferable) and a reputation score (non-transferable) that indicates the influence they hold within the organisation.

The Ethereum blockchain-based examples of Backfeed and Steem discussed above illustrate that bringing economic coordination and governance institutions to spaces that currently are either served poorly or served not at all by extant coordination mechanisms of markets, hierarchies and governments. In other words, the impact of blockchain technology may be less to improve the efficiency of existing economic orders (for example dis-intermediating payments and finance) than to expand the scope and depth of economic governance through the evolution of new types of coordinating institutions that are native to blockchains [29]

The evolutionary character of modern institutional economic analysis is Veblenian and Darwinian [18] [20] or game theoretic [27]. Institutions are better understood as coordinating rules, rather than as disruptive new technologies. What is rather interesting about blockchain, and discussed elaborately in preceding sections is that the current mix of hype and scepticism about its status as an ICT or a GPT has largely overlooked its status as an institutional technology. As Davidson [29] states that new technologies of governance are relatively rare but it is important to identify them because unlike most GPTs, where the main dynamic effect is diffuse productivity gains, an institutional technology introduces a new mode of economic coordination and governance.

8.4 Economic Challenges and Limitations with Blockchain

Blockchain at the moment faces few big challenges, some of the most striking issues are: Data Privacy, Scalability, and Security which are pervasive over all applications. Further sticking points are government regulation, and interoperability.

8.4.1 Scalability

Currently the technology runs into a number of challenges which needs to be solved before it can be extended, adopted and be treated as a mainstream technology. One of the most evident challenge blockchains are currently ringing with is scalability. This very issue can make the difference between extensive adaptability in various sectors such as health, government and finance or just limited private use in a consortium.

As mentioned previously, the blockchain technology is not a panacea to all of the worlds problems and it is not without limit. In other words, yes, the blockchain has a scalability problem.

There are several concerns that are important to consider. Firstly, Blockchain is a high energy-consuming technology. While attempts to reduce the energy costs already exist, Blockchain will always require servers and computers to process transactions. Therefore, in countries where the Internet is frequently shut down, where there is poor energy infrastructure, and where brownouts are common, the distributed ledger technology rapidly reaches its limits of scalability [42].

In essence a blockchain is a sequence of linked blocks of a specific size and each containing some information such as transactions. Now one of the reasons the scalability problem arises is due to the very size of a block. If we for instance take the bitcoin blockchain, the block size there is currently limited to 1MB which means that it can commit to 4.6 transactions per second(TPS) whereas in contrast businesses like visa can process 1736 TPS. Now obviously if we just read this fact it is easy to come up with the idea of just increasing the block size which certainly would fit more transactions. However being that said there are a lot of arguments which speak against this approach for example if we assume to increase the block size we would also require more computing power, more storage and more network bandwidth which is not easy to come up with on a public blockchain. Meanwhile these limitations don't apply to private blockchains owing to the fact that we can make sure that each individual node on the network is of high quality and compute with high bandwidth internet connection [92] [93].

Furthermore, in today's business world any type of organisation is dealing with some sort of transaction and some companies like visa have to deal with humongous amount of transactions day by day, when these transactions become constrained in some way, so does the business. Now when one tends to look at the scalability issues that blockchain brings along with it many businesses will have to think twice from a very critical point of view if adapting their business to such a technology is efficient.

This challenge out of several to be discussed can make the difference between extensive adaptability or just limited private use [1] [43].

8.4.2 Privacy

One of the main pronouncement blockchain promises with its technology is Privacy. In simple terms privacy deals with the competence of an individual or any organisation to decide whether data or information should be revealed to third parties and also how they should make use of such data.

Blockchain technology can be mainly divided into 3 types, i.e.: public, private and consortium blockchains. The challenge of privacy is predominantly present in the public blockchain due to several reasons. Mainly due to the fact that transactions on a public blockchain are transparent and globally published which makes the application for organisations where privacy is of topmost importance, such as finance, health or in the government, a reasonable challenge for adopting or extending their business. Moreover, public blockchains transparency enables anyone to make use of the information stored in the blockchain for personal gain. Even though it is possible to encrypt the data stored

on a blockchain, this would also expose the user to several risks. For instance, in the case where the owner of the data loses its Private key to decrypt, the information then can not be recovered accurately anymore or also the risk of the Private key being stolen and being published is a real possibility, resulting in all the sensitive data being forever be decrypted in the blockchain since the data cannot be altered [94].

Therefore it's understandable when people don't feel comfortable adapting to blockchain storing personal or sensitive data on the chain when the danger of theft still exists [1] [43].

8.4.3 Security

Security and privacy are strongly linked together in the digital world. In other words, with the level of privacy a technology provides to its stakeholders will ultimately affect how secure the data or information can be organised for the end user.

Even though blockchain technology is frequently praised how secure it is through the use of asymmetric and symmetric cryptography, many people would be surprised when they knew the number of issues the distributed ledger technology (DLT) on which blockchains are implemented on could be affected by, such as the 51% attack, when a single entity achieves to hold 51% of the hashing power it can take control of the blockchain in that case being able to modify transactions or even halt the block verification process. Additionally, most of the famous blockchains like bitcoin or ethereum have demonstrated their robustness against all types of attacks, however applications which are built on top of these are still prone attacks because of minor software flaws in the application [43] [95].

Furthermore, the access point into the blockchain can be the interface a blockchain is exposed to most security threats. For example a person or a business who accesses their blockchain based server to perform a transaction will most likely do that on a computer, laptop or any other device. Depending on how secure these devices are, for hackers this is the point on the chain which is most reachable for them to get the private key. Once the private key gets stolen or hacked from the computer the business or individual will have to fear that everything on their blockchain will be in serious danger and any unauthorized parties will have access to it [96].

It just takes a small security hole which could trigger massive or even irrecoverable damage to an individual or business. Although security strategies and tactics are getting better and more secure, we need to keep in mind that criminals also update their tricks to breach these security walls [96].

8.4.4 Issues with Regard to Government Regulation

When the subject of blockchain is raised, we instantly think of Bitcoin and other cryptocurrencies, where it has proven its huge potential. Its technology is looked upon as revolution in the digital sphere where transactions between parties take place in a transparent secure and trustless environment thereby completely cutting out all third parties, middleman and any commonly involved fees.

When taking the above view into consideration one might feel to adopt blockchain in their own organisation. However, although blockchain technology has many perks and definitely is a cutting edge technology, many business leaders are reluctant to adopt this mechanics in their own business environment. One of the few reasons which can be elaborated upon is the worry of future government regulations which might affect their business, such that costly and complicated changes have to be undertaken in order to comply with the government [97].

We can see a more concrete instance in the United States, especially in New York where a Bitlicense has been released which has unfriendly terms and regulation which makes

it especially challenging for small businesses to establish themselves, or China's government which has started a war against bitcoin and other digital currencies by banning fundraising through initial coin offerings and shutting down all mainland digital currency exchanges. These limitations could set the tone for worldwide regulations and slow down the mainstream development of the blockchain industry. Based on how different governments perceive and adapt to blockchain, it can make a huge difference for the end-user [98].

Countries like Great Britain or Australia have issued high-level government science reports on the prospects of the technology or cities such as Dubai or cantons like Zug in Switzerland are trying to move many aspects of government services to the blockchain, or to create special crypto-economic zones. If the government alters their regulatory measures for instance to comply with the blockchain economy, it could actually widen the adaptability of blockchain and help the blockchain revolution along [98].

Moreover through a regulated blockchain it will narrow down illicit undertakings on the blockchain such as money laundering, tax evasion and even terrorist financing activities. The blockchain stands behind the philosophy to provide a decentralized system abstracting away from all third parties and therefore creating an immutable and anonymous network for individuals or any business to interact but if the network is so optimized, that it provides a platform for criminals to engage in fraudulent operation many sincere and law abiding citizens would feel uncomfortable using such a platform [91].

8.4.5 Challenges Affecting the Business Model

Not only does the distributed ledger system come with issues and challenges to be solved but has also achieved to cause a disarray in the business sphere. In order to explain how blockchain will affect current businesses and undertakings when trying to adapt to this technology, we need to analyse as to how it is going to affect their business model.

In this day an age most transactions, trades or purchases are carried out through a middleman, also known as the intermediary that serves as a link between parties that are trying to facilitate a business deal. Now this activity is generally done for a certain amount of fees or commission. Bringing blockchain into the picture we notice that the role as intermediaries is getting less and less appropriate as the blockchain assures to entirely cut out the middleman. Cutting out the middleman sounds great, unless you are the middleman and there are a lot of middlemen and intermediaries, whole industries such as payments, settlements or securities clearing have evolved to rely them because they have created trust in a place where it is strictly speaking not necessary. Moreover intermediaries know and trust each other and their business model is built on it, that being the case we can say that they have far more to lose than to gain from breaching that trust. Therefore the distributed ledger has the potential to disrupt the entire ecology of intermediaries and hence poses a challenge to them.

Additionally, also a problem businesses face when trying to adapt to blockchains is, when they already have invested a large amount of money in existing technologies over the years. Furthermore, according to a report from "World Economic Forum and Accenture" a survey of 55 people across 13 industries revealed that on average they expected 24% return on investment (ROI) on early blockchain projects but saw only 10% on average. Therefore it makes sense for organisation to carefully think about the adoption of blockchain in their business and not just because of the hype it is creating[101].

8.4.6 Interoperability

Blockchain technology has now been around for almost over a decade and has since then become a rapidly expanding industry with thousands of undertakings making use of

the distributed ledger technology. With so many different networks, the problem arises when these individual networks desire to communicate with one another. Most of these Blockchains work in silos and are unfit for any communication or sending or receiving information from other networks. In simple words, one blockchain has no awareness of possible existence or available information in another blockchain. For example the bitcoin blockchain has no information of data contained in the ethereum blockchain and vice versa [99].

According to a Deloitte report, the lack of interoperability “grants blockchain coders and developers freedom, and can give IT departments headaches as they discover that platforms can’t communicate without translation help.” The report highlights that on GitHub, over 6,500 projects are leveraging a variety of blockchain platforms with different protocols, coding languages, consensus mechanisms, and privacy measures. “Standardization could help enterprises collaborate on application development, validate proofs of concept, and share blockchain solutions as well as making it easier to integrate with existing systems,” as stated in the Deloitte study [100].

Moreover interoperability really poses a challenge also for businesses if they require an infrastructure that is intolerable in order to process things such as payments. We could take the visa corporation for instance which can be operated on a global basis across merchants and ATMs worldwide but with the lack of interoperability in this case it wouldn’t have the same usability across the world if it stayed isolated, no matter how scalable it might be in contrast.

In conclusion even though a number of projects have been working on blockchain interoperability solutions such as Polkadot, Cosmos and Chainlink, networks remain largely isolated. However, it is important to keep in mind that the blockchain space is still relatively new, and most of the aforementioned startups are at early stages of their road-map [99].

8.5 Conclusion

The Web, initially built with a great vision in mind, should connect people all over the world, enable them to share ideas and collaborate without the obstacles of the physical world. However, it has evolved over time to a space, which is dominated by very few powerful agents. Those entities collect large parts of the welfare, and optimize their operations to maximize profits. The main users of the Web, the general public, is exposed to that, without much protection. Regulation is one of the few tools that we have available, however, it is not efficient enough in such a dynamic world. Web 3.0 is trying to address these problems with a novel approach. Through mechanism design, it is trying to enforce rules in an absolute manner, and as a consequence reduce the amount of trust we have to put into application developers and service providers.

In this paper, we explored the nature of capitalism and how the current technological implementation of the web serves to attain an efficient economic system within capitalism. Blockchains empower new types of organization which directly contradict some of the fundamental principles of our current economic order. That is to say centralization, the private ownership of business knowledge as well as the degree to which information is scarce.

While there is adequate literature discussing blockchains as an ICT, there needs to be a healthy discourse about this immutable ledger from an institutional technology perspective. Blockchains allow for a new type of decentralized governing structure, which is categorized as a trust machine. The usability of blockchain technology in solving the issues of the Fourth Industrial Revolution are enormous. Someday in the far future, this new governing structure might replace coordination mechanisms which today feel untouched.

Nonetheless, there still are a multitude of obstacles ranging from scalability, privacy, security, government regulation and interoperability amongst others which have to be dealt with before blockchains can truly replace current coordination mechanisms.

Bibliography

- [1] Melanie Swan: *Blockchain: Blueprint for a New Economy*. O'Reilly Media, Inc. Sebastopol, California, 2015.
- [2] Max Mersch: *Which New Business Models Will Be Unleashed By Web 3.0?* Medium, April 2019. <https://medium.com/fabric-ventures/which-new-business-models-will-be-unleashed-by-web-3-0-4e67c17dbd10>, last visit December 15, 2019.
- [3] Jutta Steiner: *What The Heck Is Web 3.0 Anyway?* Forbes, October 2018. <https://www.forbes.com/sites/juttasteiner/2018/10/26/what-the-heck-is-web-3-0-anyway>, last visit December 15, 2019.
- [4] Marcella Atzori: *Blockchain Technology and Decentralized Governance: Is the State Still Necessary?* 2015, SSRN: <https://ssrn.com/abstract=2709713>.
- [5] Chris Berg, Sinclair Davidson, Jason Potts: *Blockchains Industrialise Trust*. November 2017, SSRN: <https://ssrn.com/abstract=3074070>, DOI: <http://dx.doi.org/10.2139/ssrn.3074070>.
- [6] Will Steffen, Wendy Broadgate, Lisa Deutsch, Owen Gaffney, Cornelia Ludwig: *The Trajectory of the Anthropocene: The Great Acceleration*. The Anthropocene Review 2(1), pp.81-98, January 2015, DOI: <https://doi.org/10.1177/2053019614564785>.
- [7] Fridolin Krausmann, Simone Gingrich, Nina Eisenmenger, Karl-Heinz Erb, Helmut Haberl, Marina Fischer-Kowalski: *Growth in Global Materials Use, GDP and Population During the 20th Century*. May 2009, DOI: <http://dx.doi.org/10.1016/j.ecolecon.2009.05.007>.
- [8] Timothy F. Bresnahan, Manuel Trajtenberg: *General Purpose Technologies “Engines of Growth?”* Journal of Econometrics, vol. 65, no. 1, pp. 83-108, 1995.
- [9] Christian Catalini, and Joshua S. Gans: *Some Simple Economics of the Blockchain*. 2016, SSRN: https://ssrn.com/abstract_id=2874598.
- [10] Christian Catalini and Catherine Tucker: *Seeding the S-Curve: The Role of Early Adopters in Diffusion*. 2016, SSRN https://ssrn.com/abstract_id=2835854.
- [11] Oliver Hart: *An Economists Perspective on the Theory of the Firm*. Columbia Law Review, 89: 1757-74, 1989.
- [12] Oliver Hart, John Moore: *Property Rights and the Nature of the Firm*. Journal of Political Economy 98: 1119-58, 1990.
- [13] Michael C. Jensen, William H. Meckling: *Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure*. Journal of Financial Economics, 3(4): 305-60, 1976.

- [14] Richard G. Lipsey, Kenneth I. Carlaw, Clifford T. Bekar: *Economic Transformations: General Purpose Technologies and Long Term Economic Growth*. Oxford University Press, 2005.
- [15] Ronald H. Coase: *The Nature of the Firm*. *Economica*, 4(16): 386-405 1937.
- [16] Ronald H. Coase: *The Problem of Social Cost*. 1960.
- [17] Geoffrey M. Hodgson: *Opportunism Is Not the Only Reason Why Firms Exist: Why an Explanatory Emphasis on Opportunism May Mislead Management Strategy*. *Industrial and Corporate Change*, 13(2): 401-18, 2004.
- [18] Geoffrey M. Hodgson: *The Approach of Institutional Economics*. *Journal of Economic Literature*, 36(1): 166-92, 1998.
- [19] Geoffrey M. Hodgson: *Conceptualizing Capitalism*. University of Chicago Press, Chicago, IL, 2015.
- [20] Geoffrey M. Hodgson, Thorbjørn Knudsen: *Darwin's Conjecture*. University of Chicago Press, Chicago, IL, 2010.
- [21] Frederick L. Nussbaum: *A History of the Economic Institutions of Modern Europe*. F.S. Crofts & Co., New York, 1933.
- [22] Douglass C. North: *Institutions, Institutional Change, and Economic Performance*. Cambridge University Press, Cambridge, MA, 1990.
- [23] Elinor Ostrom: *Governing the Commons*. Cambridge University Press, New York, 1990.
- [24] Elinor Ostrom: *Understanding Institutional Diversity*. Princeton University Press, Princeton, NJ, 2005.
- [25] Jason Potts: *Knowledge and Markets*. 2001, DOI: <https://doi.org/10.1007/PL00003865>.
- [26] Jason Potts: *The New Evolutionary Microeconomics*. Edward Elgar, 2000.
- [27] Andrew Schotter: *The Economic Theory of Social Institutions*. Cambridge University Press, Cambridge, 2008.
- [28] Jean Tirole: *Incomplete Contracts: Where Do We Stand?* *Econometrica*, 67(4): 741-81, 1999.
- [29] Sinclair Davidson, Primavera de Filippi, Jason Potts: *Blockchains and the Economic Institutions of Capitalism*. *Journal of Institutional Economics*, Cambridge University Press, 14 (4), pp.639-658, 10.1017/S1744137417000200, hal-01850927, July 2018.
- [30] Sinclair Davidson, Primavera de Filippi, Jason Potts: *Economics of Blockchain*. Public Choice Conference, Fort Lauderdale, United States. 10.2139/ssrn.2744751, hal-01382002, May 2016.
- [31] Niki Wiles: *The Radical Potential of Blockchain Technology*. 2015. <https://www.youtube.com/watch?v=JMT0xwmFKIY>.
- [32] Oliver E. Williamson: *Transaction Cost Economics: The Governance of Contractual Relations*. *Journal of Law and Economics*, 22(2): 233-61, 1979.

- [33] Oliver E. Williamson: *The Economic Institutions of Capitalism*. Free Press, New York, 1985.
- [34] Oliver E. Williamson: *Comparative Economic Organisation: The Analysis of Discrete Structural Alternatives*. *Administrative Science Quarterly*, 36(2): 269-96, 1991.
- [35] Oliver E. Williamson: *Opportunism and its Critics, Managerial and Decision Economics*. 14(2), 97-107 1993.
- [36] Basil S. Yamey: *Scientific Bookkeeping and the Rise of Capitalism*. *Economic History Review*, 1(2/3): 99-121, 1949.
- [37] Sir Isaac Newton: *The System of the World*. CreateSpace Independent Publishing Platform, 2015.
- [38] George Gilder: *Life after Google: The Fall of Big Data and the Rise of the Blockchain Economy*. Gateway Editions, 2017.
- [39] George Gilder: *Knowledge and Power: The Information Theory of Capitalism and How It Is Revolutionizing Our World*. Gateway Editions, 2013.
- [40] Michel Bauwens, Vasilis Kostakis, Alex Pazaitis: *Peer to Peer: The Commons Manifesto*. University of Westminster Press, 2019.
- [41] Lewis A. Kornhauser: *The Nexus of Contracts Approach to Corporations: A Comment on Easterbrook and Fischel*. Columbia Law Review Association, 2019.
- [42] Andrej Zwitter, Mathilde Boisse-Despiaux: *Blockchain for Humanitarian Action and Development Aid*. *Journal of International Humanitarian Action*, 2018.
- [43] Imran Bashir: *Mastering Blockchain: Deeper Insights Into Decentralization, Cryptography, Bitcoin, and Popular Blockchain Frameworks*. Published by Packt Publishing Ltd, March 2017.
- [44] Satoshi Nakamoto: *Bitcoin: A Peer-to-Peer Electronic Cash System*. 2008, <https://bitcoin.org/bitcoin.pdf>, last visit December 15, 2019.
- [45] Larimer, Daniel, Ned Scott, Valentine Zavgorodnev, Benjamin Johnson, James Calfee, Michael Vandenberg: *Steem: An Incentivised Blockchain-Based Social Media Platform*. 2016, <https://steem.io/SteemWhitePaper.pdf>, last visit December 15, 2019.
- [46] ICORating: *ICO Market Research Q1 2018*. August 2018, https://icorating.com/ico_market_research_q1_2018_icorating.pdf, last visit December 16, 2019.
- [47] World Economic Forum: *Building Blockchain for a better Planet*. Sep 2018, DOI: <https://doi.org/10.1093/biosci/biz088>.
- [48] Cathy O'Neil: *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, September 6, 2016.
- [49] *Drawn from the discussions at, and briefings prepared for, the International Dialogue on the Global Commons held in Washington DC (USA)*. Washington DC, October 2016.

- [50] McKinsey: *Resource Revolution: Meeting the World's Energy, Materials, Food, and Water Needs*. November 2011, <https://www.mckinsey.com/business-functions/sustainability/our-insights/resource-revolution>, last visit December 17, 2019.
- [51] Wood, G.: *Ethereum for Dummies*. Youtube, December 2015, https://www.youtube.com/watch?v=U_LK0t_qaPo, last visit December 15, 2019.
- [52] Web3.0 Foundation: *Homepage*. <https://web3.foundation>, last visit December 15, 2019.
- [53] EUR-Lex: *Directive (EU) 2015/2366 of the European Parliament and of the Council of 25 November 2015 on payment services in the internal market, amending Directives 2002/65/EC, 2009/110/EC and 2013/36/EU and Regulation (EU) No 1093/2010, and repealing Directive 2007/64/EC*. <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32015L2366>, last visit December 15, 2019.
- [54] World Wide Web Foundation: *History of the Web*. <https://webfoundation.org/about/vision/history-of-the-web>, last visit December 15, 2019.
- [55] The Economist: *The Great Chain of Being Sure About Things* <https://www.economist.com/briefing/2015/10/31/the-great-chain-of-being-sure-about-things>, last visit December 18, 2019.
- [56] W3.org: *Frequently Asked Questions*. <https://www.w3.org/People/Berners-Lee/FAQ.html>, last visit December 15, 2019.
- [57] IBM: *developerWorks Interviews: Tim Berners-Lee*. <https://www.ibm.com/developerworks/podcast/dwi/cm-int082206txt.html>, last visit December 15, 2019.
- [58] Brave Browser: *Homepage*. <https://brave.com>, last visit December 16, 2019.
- [59] Bloomberg: *Google and Facebook Too Can Be Disrupted*. <https://www.bloomberg.com/opinion/articles/2017-12-08/google-and-facebook-too-can-be-disrupted>, last visit December 16, 2019.
- [60] arstechnica: *Mozilla Co-Founder Unveils Brave, a Browser That Blocks Ads by Default*. <https://arstechnica.com/information-technology/2016/01/mozilla-co-founder-unveils-brave-a-web-browser-that-blocks-ads-by-default>, last visit December 16, 2019.
- [61] Interplanetary File System (IPFS): *Homepage*. <https://ipfs.io>, last visit December 16, 2019.
- [62] Tardigrade: *Homepage*. <https://tardigrade.io>, last visit December 16, 2019.
- [63] status: *Homepage*. <https://status.im>, last visit December 16, 2019.
- [64] steemit *Homepage*. <https://steemit.com>, last visit December 16, 2019.
- [65] The Guardian: *The End of Capitalism Has Begun*. <https://www.theguardian.com/books/2015/jul/17/postcapitalism-end-of-capitalism-begun>, last visit December 16, 2019.

- [66] Acton: *'Knowledge and Power': The Information Theory of Capitalism and How It Is Revolutionizing Our World*. <https://acton.org/knowledge-and-power-information-theory-capitalism-and-how-it-revolutionizing-our-world>, last visit December 16, 2019.
- [67] Harvard Business Review: *Limits of the Learning Curve*. <https://hbr.org/1974/09/limits-of-the-learning-curve>, last visit December 16, 2019.
- [68] O'Reilly: *What Is Web 2.0*. <https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html>, last visit December 16, 2019.
- [69] Webopedia: *Web as a Platform*. https://www.webopedia.com/TERM/W/web_as_a_platform.html, last visit December 16, 2019.
- [70] The Economist: *The World's Most Valuable Resource Is No Longer Oil, but Data*. <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>, last visit December 16, 2019.
- [71] Computerworld: *Data Storage Goes From \$1M to 2 Cents per Gigabyte*. <https://www.computerworld.com/article/3182207/cw50-data-storage-goes-from-1m-to-2-cents-per-gigabyte.html>, last visit December 16, 2019.
- [72] The Register: *Great Time to Shift Bytes: International Bandwidth Prices Are in Free Fall*. https://www.theregister.co.uk/2018/06/06/international_10_gbps_circuit_prices_plummeting, last visit December 16, 2019.
- [73] The Economist: *Drastic Falls in Cost Are Powering Another Computer Revolution*. <https://www.economist.com/technology-quarterly/2019/09/12/drastic-falls-in-cost-are-powering-another-computer-revolution>, last visit December 16, 2019.
- [74] Google: *About Google*. <https://about.google>, last visit December 16, 2019.
- [75] Channel4: *If Google Is Free, How Does It Make So Much Money?* <https://www.channel4.com/news/if-google-is-free-how-does-it-make-so-much-money>, last visit December 16, 2019.
- [76] The Guardian: *Google, Not GCHQ, Is the Truly Chilling Spy Network*. <https://www.theguardian.com/commentisfree/2017/jun/18/google-not-gchq--truly-chilling-spy-network>, last visit December 16, 2019.
- [77] TechTarget: *Information Arbitrage*. <https://whatis.techtarget.com/definition/information-arbitrage>, last visit December 16, 2019.
- [78] Hethoughts: *Information Scarcity*. <https://hethoughts.wordpress.com/2012/02/09/information-scarcity>, last visit December 16, 2019.
- [79] Stanford University: *Gödel's Incompleteness Theorems*. <https://plato.stanford.edu/entries/goedel-incompleteness>, last visit December 16, 2019.
- [80] Columbia University: *Theory of Capitalism* <https://capitalism.columbia.edu/theory-capitalism>, last visit December 16, 2019.

- [81] International Monetary Fund: *Winds of Change: The Case for New Digital Currency*. <https://www.imf.org/en/News/Articles/2018/11/13/sp111418-winds-of-change-the-case-for-new-digital-currency>, last visit December 16, 2019.
- [82] Reuters: *Father of Web Says Tech Giants May Have to Be Split Up*. <https://www.reuters.com/article/us-technology-wwf-father-of-web-says-tech-giants-may-have-to-be-split-up-idUSKCN1N63MV>, last visit December 16, 2019.
- [83] Blockstack: *Homepage*. <https://blockstack.org/about>, last visit December 16, 2019.
- [84] Solid: *Homepage*. <https://solid.inrupt.com>, last visit December 16, 2019.
- [85] Backfeed: *Homepage*. <http://backfeed.cc>, last visit December 16, 2019.
- [86] DAOstack *Homepage*. <https://daostack.io>, last visit December 16, 2019.
- [87] Polkadot: *Homepage*. <https://polkadot.network>, last visit December 16, 2019.
- [88] Fortune: *Hoax Over 'Dead' Ethereum Founder Spurs \$4 Billion Wipe Out*. <https://fortune.com/2017/06/26/vitalik-death>, last visit December 16, 2019.
- [89] Dataversity: *A Brief History of Machine Learning*. <https://www.dataversity.net/a-brief-history-of-machine-learning>, last visit December 16, 2019.
- [90] IBM: *Big Data Analytics*. <https://www.ibm.com/analytics/hadoop/big-data-analytics>, last visit December 16, 2019.
- [91] Cryptoeconomics: *The Blockchain Economy: What Should the Government Do?* <https://medium.com/cryptoeconomics-australia/the-blockchain-economy-what-should-the-government-do-c69cbdab7c3c>, last visit December 16, 2019.
- [92] Preethi Kasireddy: *Blockchains Don't Scale Not Today at Least but There's Hope*. Hackernoon, August 2017. <https://hackernoon.com/blockchains-dont-scale-not-today-at-least-but-there-s-hope-2cb43946551a>, last visit December 16, 2019.
- [93] Kenny Li: *The Blockchain Scalability Problem the Race for Visa Like Transaction Speed*. Hackernoon, January 2019. <https://hackernoon.com/the-blockchain-scalability-problem-the-race-for-visa-like-transaction-speed-5cce48f9d44>, last visit December 16, 2019.
- [94] Matteo Cagnazzo, Chris Wojzechowski: *Security and Privacy in Blockchain Environments*. Dotmagazine, June 2017. <https://www.dotmagazine.online/issues/innovation-in-digital-commerce/what-can-blockchain-do/security-and-privacy-in-blockchain-environments>, last visit December 16, 2019.
- [95] Ajay Chandhok: *Top Five Blockchain Security Issues*. Ledgerops, <https://ledgerops.com/blog/2019/03/28/top-five-blockchain-security-issues-in-2019>, last visit December 16, 2019.

- [96] Rick Martin: *Blockchain Security Vulnerabilities Risks*. Ignite, November 2018. <https://igniteoutsourcing.com/blockchain/blockchain-security-vulnerabilities-risks/>, last visit December 16, 2019.
- [97] Kevin Werbach: *People Don't Trust Blockchain Systems, Is Regulation a Way to Help?* The Conversation, February 2019. <https://theconversation.com/people-dont-trust-blockchain-systems-is-regulation-a-way-to-help-110007>, last visit December 16, 2019.
- [98] Chris Berg, Sinclair Davidson, Jason Potts: *The Blockchain Economy What Should the Government Do?* Medium, November 2017. <https://medium.com/cryptoeconomics-australia/the-blockchain-economy-what-should-the-government-do-c69cbdab7c3c>, last visit December 16, 2019.
- [99] Stephen O'Neal: *Blockchain Interoperability Explained*. Cointelegraph, September 2019. <https://cointelegraph.com/explained/blockchain-interoperability-explained>, last visited December 16, 2019.
- [100] Ryan Browne: *Five Crucial Challenges for Blockchain to Overcome*. CNBC, October 2018. <https://www.cnbc.com/2018/10/01/five-crucial-challenges-for-blockchain-to-overcome-deloitte.html>, last visited December 16, 2019.
- [101] Joe McKendrick: *Useful Metrics for Measuring Blockchain Results*. ZDNet, November 2019. <https://www.zdnet.com/article/19-useful-metrics-for-measuring-blockchain-results>, last visited December 16, 2019.

Chapter 9

Economics of Fifth Generation Cellular Networks

Rabiya Abdullah, Annesha Bhoumik, Dominik Jurilj, Manpreet Singh Sohal

The need for better, faster and more seamless connectivity has propelled telecommunications from wired technology to broadband cellular network technology (4G), that is currently available in more than 80 countries worldwide. However, as modern societies keep expanding the horizons of internet usage - from exchanging messages to driving autonomous vehicles - the need and ability to handle larger volumes of data at faster speeds is a key driving force behind the development of the next generation of telecommunication technology or 5G.

While the development of 5G is inevitable and is bound to create new ripples in the technological world, it is also not without its controversies; the Wall Street journal in a recent article remarked that the "5G Race Could Leave Personal Privacy in the Dust" [1]. On the economic front too, there are increasing concerns that the cost of installing infrastructure necessary for 5G will be too much for many of the world's developing economies and therefore only contribute to the increasing wealth gap in this world. This report is an attempt to summarize the economics of 5G technology and tries to weigh in on both the pros as well as the cons before arriving at a conclusion.

Contents

9.1	Introduction	35
9.1.1	The drivers behind 5G	36
9.1.2	The underlying technologies that will be at the heart of 5G	39
9.1.3	Status of the 5G Market	40
9.2	Economics and Implementation of 5G	41
9.2.1	Economic Stimulation by 5G Technology	41
9.2.2	Economic Impact of 5G in Major Economies	42
9.2.3	Economic Linkages with 5G	42
9.2.4	Costs of 5G Infrastructure	43
9.2.5	Challenges in Implementation of 5G	48
9.3	Security Aspects of 5G	48
9.3.1	Similarities and Differences in 5G Security compared to older 4G Technologies	48
9.3.2	Importance of Trust and other security standards in 5G	49
9.3.3	Life and 5G: The Security of Organic Beings	50
9.3.4	Economical Aspects of Security	51
9.4	Case studies	51
9.4.1	Industrial 5G	51
9.4.2	Case studies on 5G	54
9.4.3	Case 1: Trade war between the US and China	54
9.4.4	Case 2: Huawei ban in UK	56
9.4.5	Case 3: 5G in South-East Asia	57
9.4.6	Case 4: Germany deploying 5g with Huawei and situation in some other European countries:	57
9.4.7	Case 5: Huawei introducing 5G in Switzerland:	58
9.4.8	Forecasted Dominators in 5G technology	59
9.5	Conclusion	60

9.1 Introduction

The conclusion of a discussion paper published by the Centre for European Economic Research in 2016, titled "The economic impacts of telecommunications networks and broadband internet" [2], strongly suggests that there is a strong correlation and causation between adoption of higher speed wireless internet and economic growth, particularly in developing countries. Societies and governments, too, are keen to build on this correlation and it is therefore no surprise that 4G internet is now available in more than 86 countries for more than 50% of the time [3].

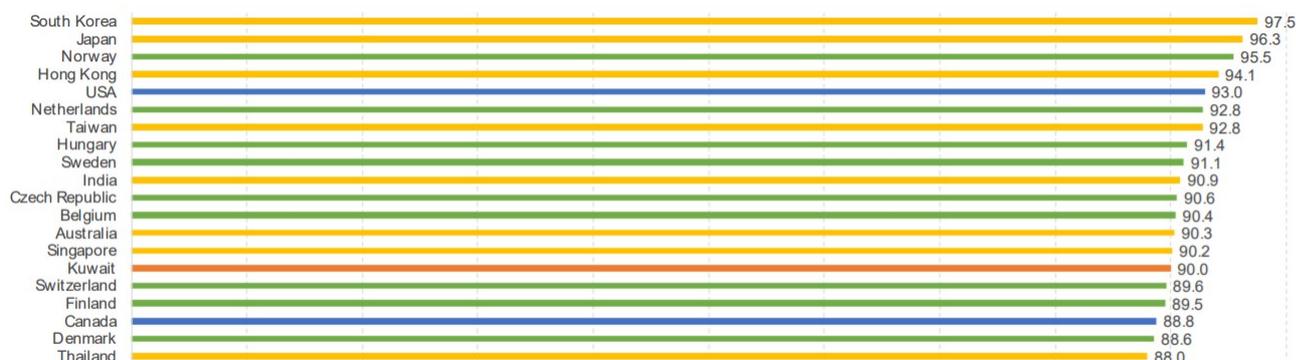


Figure 9.1: 4G Availability

The journey from the early days of telephone to the introduction of 4G has not been a short one. The cost of laying cables and other infrastructure for wired telecommunications was a key reason for poor wired telecommunication coverage in many developing countries, and it continues to be a problem even today. The advent and further development of wireless telecommunication has transformed both the nature of utilization as well as the utilization of phones.

Early wireless technology (0G) permitted the possibility to transfer audio signals without having to be in a fixed position. Technologies used included the Push to Talk (PTT or manual), Mobile Telephone Service (MTS), Improved Mobile Telephone Service (IMTS), and Advanced Mobile Telephone System (AMTS) systems. Where commercially available, these systems were an extension of the wired public telephone network to prevent it from interfering with other wireless networks such as police radio or a taxi dispatch system.

The first generation of wireless cellular technology was introduced in the early 1980s. It was an analog network that allowed the transfer of voice signals (between 20 kHz and 20 kHz) by first modulating them to higher frequencies (e.g. 150 MHz), and then transferring them using radio towers.

The advent of digital technology in the late 1980s and early 1990s heralded the next generation of wireless cellular technology (2G). The most significant benefit of digital vs. analog technologies was that unlike analog signals, digital signals can be encrypted, allowing the transfer of data in such a way that only the intended receiver can receive it. Further, compared to 1G or G, 2G allowed significantly more efficient use of the radio frequency spectrum enabling more users per frequency band. Finally, sending written messages (SMS), in addition to audio, was also made possible with the advent of 2G.



Figure 9.2: Increasing human connectivity, one "G" at a time

The early 2000s saw the introduction of 3G mobile services which created the possibility of browsing the internet at higher speeds on mobile phones. The arrival of 3G supported the introduction of smartphones, for example Blackberry, iPhone, and others running on Microsoft and Android operating systems (OS). The combination of multiple features such as email, multimedia messaging, mp3 music and voice calling led to widespread demand for mobile internet connectivity. In many ways, the introduction of 3G technology created the ground for internet-based economies. Entrepreneurial focus shifted towards providing an increased number of services using mobile applications.

The need for better, faster and more seamless connectivity resulted in the next generation of wireless technology or 4G. It made possible mobile web access at speeds greater than 50 Mbps in some countries (e.g. South Korea), IP telephony, gaming services, high-definition mobile TV, video conferencing, and 3D television.

Each generation shift in telecommunications, has had a direct impact on the nature and volume of economic activity. The last decade for example has seen the rise of mobile application giants like Snapchat (valued at \$19 billion [4]) as well as the development of new payment methods like WePay (China), Apple Pay (Global) and PayTM (India). The next generation of telecommunications, christened 5G, is expected to have a similar disruptive impact on economic and social life, the details of which are laid out in subsequent sections of this report.

9.1.1 The drivers behind 5G

1. **More data at faster speeds:** Data generated by smartphones has increased more than 10 times since introduction of 4G in 2009 and it is expected to increase a further ten times between 2016 and 2022 [5]. It is expected that the technologies behind 5G will make this possible at about 100 times the speed of current 4G networks, at around 10 Gbps [6]. In fact, at a test 5G network hosted by Korea Telecom during the Winter Olympics in 2018, speeds of up to 3.8 Gbps or 26 times the 4G network in Korea were observed [7].

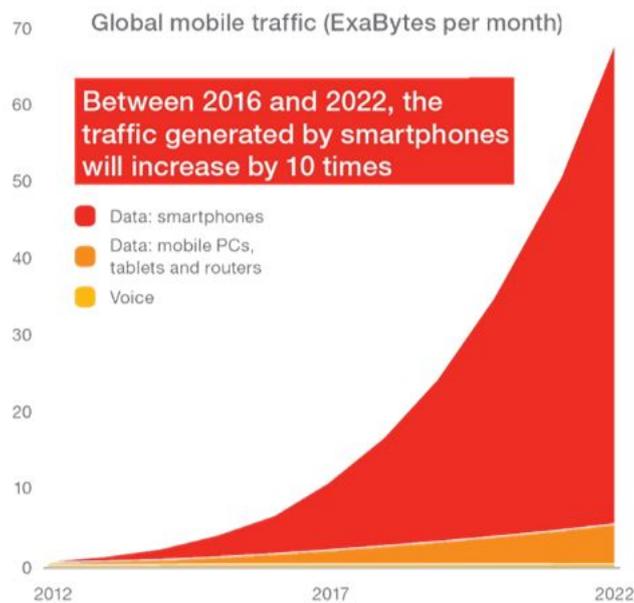


Figure 9.3: Global mobile traffic in ExaBytes per month [5]

2. **Lower Latency:** Latency refers to the delay that occurs before a connection is established. Current networks have up to 50 ms latency on an average connection, which means that data is not transmitted in real time. 5G could have latency as

low as 5 ms [8]. This means that Ultra-Reliable Low-Latency Communications or in brief mission critical applications that requires uninterrupted and robust data exchange will be possible with lesser chances of failure.

3. **Device Density and Energy Efficiency:** In addition to higher speeds and lower latency, 5G promises to greatly increase device density by allowing up to 900,000 more devices to be connected per square kilometer than 4G, which supports the connection of at most 100,000 devices per square kilometer [9]. Further, 5G networks will be approximately 100 times more energy efficient than 4G as the energy required to power each wireless device will decline [9].
4. **Tomorrow is today:** The new data-sharing network that is at the core of many technical applications would be almost impossible without 5G. Because it transmits data more efficiently, 5G has the potential for faster speeds and suffers from shorter lag times than the current 4G standard. That speed for example is critical for autonomous vehicles and vehicle-to-vehicle (V2V) communication, where timely decisions reduce the chances of collision, improve road safety, and save lives [10].

The transmission capacity and bandwidth offered by 5G will be critical in promoting the Internet of Things. Sensors in IoT devices, robots, machines, cars, and drones will be able to share large volumes of data in real time and therefore enable manufacturers, for example, to deliver faster over-the-air software updates and conduct quicker diagnosis of performance issues [10].

Further, unlike with 4G based real-time video communication, which suffers because data packets take too long to get to the destination, 5G will make virtual and augmented vision an everyday reality, thereby supporting new fields like remote-medicine or tele-maintenance, thereby increasing the quality, safety and ease of service [11].



Figure 9.4: Real-time data is used to control self-driving cars and change their behavior based on observations [10]



Figure 9.5: 5G will be at the center of the Internet of Things - making possible a society built around data-sharing [11]

9.1.2 The underlying technologies that will be at the heart of 5G

While the concepts behind 5G are rather interesting, it is the technology backbone that will ultimately make it a reality.

The underlying principle behind mobile communications is the transmission of voice, text and data via radio waves. An individual radio wave is defined by its frequency and wavelength, while a series of radio waves also requires to be defined within a certain bandwidth. All mobile communications so far have been in a frequency range below 6 GHz [12] [13]. For 5G technology however, it is intended that both sub-6 GHz as well as frequency bands in the mmWave range (24 GHz - 100 GHz) [14] are used.

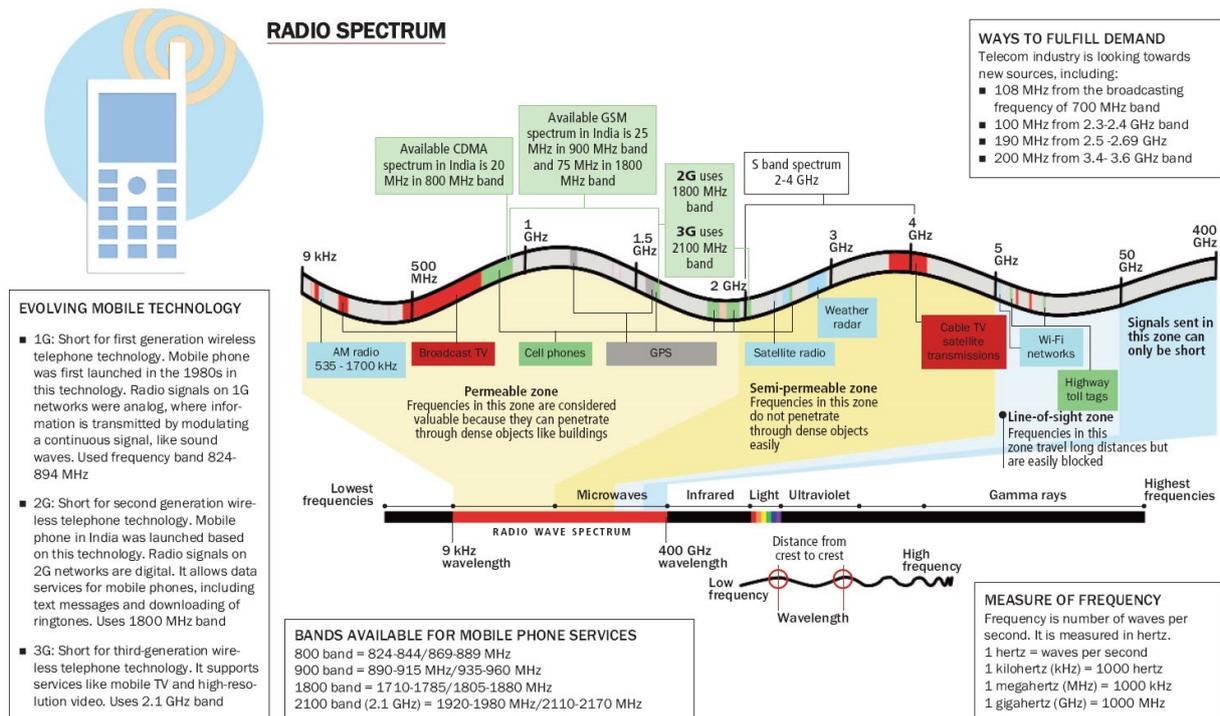


Figure 9.6: Radio spectrum - India [15]

As is mentioned in the picture above, radio waves in the frequencies above 5 GHz are easily blocked by physical obstructions like buildings or trees, even though they have the capacity to travel long distances otherwise. Therefore, the challenge of ensuring seamless communication while having to overcome the possibility of physical barriers is critical to the technologies at the heart of 5G [16], namely small cells, massive MIMO, beamforming and full duplex.

- **Small Cells:** It is likely that 5G networks will supplement traditional cellular towers with portable miniature base stations called small cells. These require minimal power to operate and can be placed every 250 meters or so throughout cities. To prevent signals from being dropped, carriers could install thousands of these stations in a city to form a dense network. The cost of setting up thousands of cells might make it financially untenable to set up in rural areas.
- **Massive MIMO:** Multiple-input multiple-output or MIMO describes wireless systems that use two or more transmitters and receivers to send and receive more data at once. 4G base stations have twelve ports for antennas that handle all cellular traffic: eight for transmitters and four for receivers. 5G base stations in turn support a hundred ports, meaning that they can send and receive signals from many times the number of users as compared to 4G. This technology is called Massive

MIMO. So far, this concept has only been tested in laboratories and is yet to be fully utilized in real time networks. The most important limitation and therefore the barrier which needs to be overcome is that installing multiple antennas also causes more interference if those signals cross.

- **Beamforming:** In order to address the concerns of interference arising from Massive MIMO and millimeter waves, cellular signals need to be transmitted such that it arrives at the user's end as a concentrated beam. This process is called Beamforming and it is akin to a traffic-signaling system for cellular base stations that identifies the most efficient data-delivery route to a user. One approach for example is using signal-processing algorithms to send individual data packets in many different directions, bouncing them off buildings and other objects in a precisely coordinated pattern before they arrive in the form of singular beam at the user's end.
- **Full Duplex:** 4G network transceivers take turns while transmitting and receiving information over the same frequency or operate on different frequencies when information is transmitted and received at the same time. 5G transceivers will be able to transmit and receive data at the same time, on the same frequency, using for example silicon transistors as high-speed switches, in a technology known as full duplex. This technology is expected to double the capacity of mobile networks as we know them today.

9.1.3 Status of the 5G Market

1. **Development and adoption of a global standard:** Like any other mode of communication, in the past, air interface was unique for different geographies. In 1G, the commonly used standards were NMT (Nordics, Switzerland, Netherlands and Russia), AMPS (North America and Australia), TACS (UK), C-450 (West Germany, Portugal, South Africa), Radiocom 2000 (France), TMA (Spain), RTMI (Italy) and TZ-801/802/803 (Japan) [17]. With time, standardization has become the norm worldwide with only two International Mobile Telecommunications (IMT) compliant standards in use throughout the world for 4G today - Long Term Evolution (LTE) Advanced and IEEE 802.16m or WirelessMAN-Advanced [18]. Continuing with the trend of standardization, 5G will be based on the New Radio access technology which will be adopted globally [19].



Figure 9.7: (Widely) Adopted Standards for Mobile Communications

5G NR defines two standard frequency ranges (FR) for mobile communication [14]:

- FR1: Sub-6 GHz frequency bands
- FR2: Frequency bands in the mmWave range (24 GHz - 100 GHz)

2. **Global Deployment of 5G [20]:** Pilot 5G networks have been rolled in more than 40 countries worldwide by 74 operators. So far, all 5G networks operate in the MHz range and therefore the full potential of the 5G market is yet to be realized.

Commercial Service		Pre-Commercial Service	
Australia	Philippines	Brazil	Russia
Austria	Qatar	Czechia	Sweden
Bahrain	Romania	Estonia	UAE
China	Saudi Arabia	Indonesia	
Finland	South Africa	Japan	
Germany	South Korea	Latvia	
Hungary	Spain	New Zealand	
Ireland	Switzerland	Norway	
Italy	United Kingdom	Poland	
Kuwait	United States	Portugal	
Maldives	Uruguay		
Monaco			

Figure 9.8: Global deployment of 5G

3. **The Swiss Connection - 5G in Switzerland:** There are two network operators who provide commercial 5G services - Swisscom and Sunrise. Both launched their 5G networks in April 2019. Sunrise uses 3500 MHz spectrum at a frequency of 100 MHz and 1800 MHz spectrum at a frequency of 20 MHz [21]. Swisscom on the other hand, currently uses the 3500 MHz spectrum at a frequency of 120 MHz to provide 5G services, but also has further options to utilize frequencies in the 700 MHz and 1400 MHz bands in the future [22].

9.2 Economics and Implementation of 5G

9.2.1 Economic Stimulation by 5G Technology

5G networks are expected to have major economic impacts. According to IHS Markit, a renowned global intelligence company, 5G is expected to stimulate the development of new business models as well as the world economy [29]. There are significant factors that allow 5G to boost the economic landscape. These factors have been discussed below:

1. In contrast to previous generation networks which were primarily connecting mobile phones, 5G can be used to connect a diverse range of devices such as appliances used at home, industrial electronic devices and machinery, cars and televisions [30].
2. 5G has the ability and capacity to connect a greater number of devices per cell in comparison to lower generation networks [30]. As per estimates, 5G is estimated to allow connection of 1 million devices per 0.38 square miles in contrast to connection of approximately 2000 devices per 0.38 square miles in the case of 4G [31].
3. 5G enables network slicing where a single network can provide distinct or differentiated services depending upon the use cases. Hence, the network can be thought of getting divided and operating in 'slices'. 5G offers the possibility to enhance each slice to offer the needed resources and service quality depending upon the latency, data flow rate, capacity and coverage requirements. This stands in contrast to all users or use cases receiving the same type of services regardless of their needs when using 4G. Hence, instead of being more about faster speed, 5G networks have the

leverage over lower generation networks by fulfilling needs when it comes to capacity and specific, differentiated services for different use cases [30].

Network slicing further strengthens the case of utility of 5G networks. Network slicing paves the way towards achieving greater efficiency and revenues for mobile network operators [33]. Management of each slice separately from the other slices is possible for the mobile network operators which means that the revenue and operating expenditure of each slice can be calculated and evaluated to see that how and where profits can be made [30].

4. The utility of big data gained enormous popularity considering the opportunities and possibilities associated with it. However, studies indicate that a large quantity of data is not being utilized as according to some studies, only 5% of the generated data is being used. A major hurdle in achieving the full benefits of big data is the inadequacy in transmission of data from the point where it is generated to the point where it can be processed and analyzed. The issue of data transmission can be dealt by deployment of 5G networks as they can increase the rate of transferring data. Utilization of 35% of the digitally generated data can be realized through 5G networks, as per estimates [30].

9.2.2 Economic Impact of 5G in Major Economies

Value created by 5G has significant implications for various industries which is why efforts are being made to develop the 5G ecosystem. 5G technology has the potential to greatly boost the e-commerce ecosystem especially in combination with other major technologies such as machine learning, Internet of Things (IoT), blockchain and artificial intelligence [30],[32]. Studies have been conducted to evaluate potential economic impacts of 5G networks in some major economies. The results have been tabulated below:

Table 9.1: Estimated Economic Impact of 5G in Major Economies and on the World [30]

Country	Year	Economic Impact	Study done by
China	2025	5G market worth U.S.\$ 167 billion (3.2% of China's GDP)	China Academy of Information and Communications Technology
India	2035	Economic stimulation worth U.S.\$ 1 trillion	A committee on India's 5G road map
UK	2030	Economic impact of U.S.\$ 226 billion	King's College London
USA	2024	Increase in GDP by U.S.\$ 500 billion and generation of 3 million jobs	Accenture
Global economy	2035	Generation of 22 million jobs and revenue worth USD 3.5 trillion with overall economic impact worth U.S.\$ 12 trillion	Qualcomm

9.2.3 Economic Linkages with 5G

Success of 5G is not possible in isolation. Development of certain industries is important for the success of 5G. The connection between these certain industries and 5G can be

conceptualized in terms of economic linkages. Success and use of 5G networks is dependent upon the industries that create the demand of 5G (forward linkage), industries that provide supply of 5G (backward linkage) as well as those industries that complement use of 5G (horizontal linkage) [30]. Focus on these linkages plays a significant role in utilizing 5G to its full potential and the benefits it has to offer.

1. Forward Linkage

Technologically advanced use cases require networks that offer fast speed, reliability and low level of latency. Economies that place great importance on such use cases create a higher demand of 5G networks [30]. An example of industry that produces the demand of 5G is autonomous vehicle industry [34]. This was witnessed when Hyundai and South Korea developed 5G equipped autonomous buses for 2018 Winter Olympics [30].

2. Backward Linkage

Development of industries and sectors that provide the needed inputs for 5G is crucial. Low labor costs and other inputs are important. Considering this, if we compare China and the US, equipment required to build wireless network is 35% cheaper in China which puts China in a better position when it comes to the cost. Moreover, research and development activities concerning 5G are also important here. 32% of the contribution towards the development of 5G globally was made by China alone, since March 2018 [30]. Furthermore, China has spent US\$ 24 billion more in comparison to US when it comes to wireless communication development and is working towards the realization of its economic plan because of which US \$ 400 billion have been reserved for investments regarding 5G [35].

Simultaneously, device manufacturers offer backward linkage as well. 5G tablets have already been released by Samsung. These tablets were used during the 2018 Winter Olympics by the viewers to locate players on a 3D map of the field [30].

3. Horizontal Linkage

Horizontal linkage exists when activity in one sector provides boost to activity in another sector. In the case of 5G, horizontal linkage exists with other technologies like artificial intelligence and big data as greater benefits can be realized by combining these technologies with 5G. This has encouraged mobile operators to gain advantage by focusing on such horizontal linkages. For example, China Mobile plans to increase 5G's operational efficiency by making use of artificial intelligence [30].

Similarly, China Unicom, a telecommunication company, and Tencent, a Chinese multinational conglomerate company, are running a laboratory together for carrying out research and development on major technologies. Major technologies include network slicing, edge computing and positioning services with greater accuracy. At the same time, Baidu, Chinese search provider and China Unicom plan to combine artificial intelligence with 5G while focusing greatly on big data and internet of vehicles [30].

9.2.4 Costs of 5G Infrastructure

Where 5G presents a hopeful picture in terms of economic impacts, it is also important to look at the costs associated with 5G. 5G technology and devices require more investment as 5G networks are able to perform better by reducing the distance to the user in terms of antennas and computing power. According to technology consultancy company, Xona Partners, 5G would need six times more investment than 4G networks to provide a similar coverage [30].

Where mobile operators look forward to gain benefits from 5G use cases and the adoption of Internet of Things (IoT), they also have to consider the increase in infrastructure investment. The use cases of 5G primarily determine the requirements of the infrastructure so that those use cases can be supported by 5G. Use cases can be broadly divided into three groups: advanced mobile broadband, Internet of Things and mission-critical applications. An increase in network performance by 10 times over existing levels in terms of scale, reliability, latency and throughput is needed to support the 5G use cases. Hence, this creates the need for the mobile operators to invest in all network domains including radio access network (RAN) infrastructure, spectrum, core networks and transmission [36].

9.2.4.1 5G Infrastructure Investment Approach

Mobile operators can evolve their existing 4G infrastructure when it comes to infrastructure development for 5G as many elements of 5G are built upon or are an extension of the elements of 4G. As long as the incremental revenue potential of 5G remains uncertain, this approach will be a favorable choice for many operators as they aim to minimize the infrastructure investment [36].

However, the increase in traffic will lead to a point where upgradation of the existing network will not be able to fulfill the demand. This will create the need for developing new macro sites or small cells which will be a significant factor behind increase in the cost of network. The timeline below shows the point in time when some selected countries will run out of capacity [36].



Figure 9.9: Timeline when at least 50% of the sites in the country will face capacity constraints [36]

9.2.4.2 Evolution of Infrastructure for 5G

Mobile operators may vary in their approach towards infrastructure development and corresponding investment for 5G but there will be some actions and trends prevalent among operators, regardless of the approach taken. Some of these are:

- Competition for spectrum:** Spectrum from 3.5 gigahertz to 80 gigahertz is being tested for 5G. However, to acquire 3.5 gigahertz bands over the short-to-medium term, followed by 26 gigahertz and 28 gigahertz bands is the main target of most mobile operators. Greater bandwidth followed by an increase in air capacity will hence be achievable because of the new spectrum. Mobile operators will have to increase investment in infrastructure to a great extent even if new spectrum is made available in order to tackle certain shortcomings and issues. For example, limitations in propagation can come up with the use of high frequency spectrum. Operators may also have the possibility to increase capacity by refarming spectrum to 4G and 5G as demand and use of 2G and 3G goes down [36].

Moreover, according to a policy report by GSMA (a trade body that represents the interests of mobile operators globally), 5G requires spectrum within three frequency

ranges in order to support the 5G use cases and fulfill its coverage requirements. These frequency ranges are Sub-1 gigahertz, 1-6 gigahertz and above 6 gigahertz. Large scale coverage across urban and rural areas in addition to provision of Internet of Things services can be supported by Sub-1 GHz. Coverage and capacity capabilities associated with 5G can be achieved through frequency range of 1-6 GHz. Lastly, above 6 GHz is for 5G ultra high broadband speed [37].

- **Inclination towards small-cells:** Densification of current networks with macro sites can satisfy the increase in demand or traffic in rural and semi urban areas. However, this will not be enough in urban areas with high population. Mobile operators will have to turn to deploying small-cells as concentration of traffic increases and as higher spectrum bands are used [36]. Hence, large scale deployment of small cells will lie at the heart of infrastructure development for 5G [38].

Furthermore, it was interesting to find out from a study that sites with traffic density above 0.5 petabyte per square kilometer had a cell radius of less than 200 meters which leads to the need for small-cells, in a European city. Similar density pattern has been observed in other cities and many others will also experience this level around 2020. The following figure shows the levels of traffic densities in some major cities around the world [36].

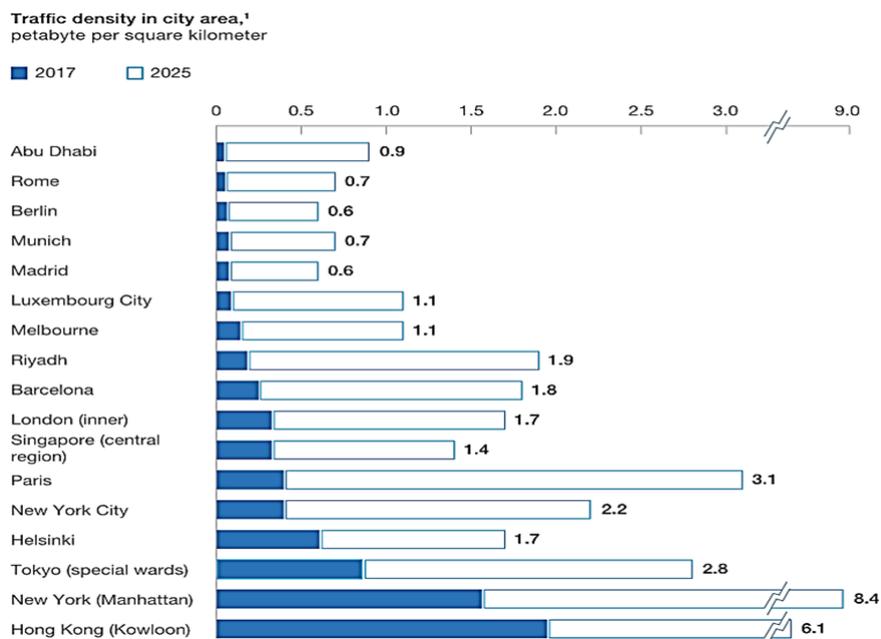


Figure 9.10: Traffic Density in Urban Areas [36]

- **Need for Fiber transmission:** Fiber only transmission will become important for improving the transmission rates. Fiberization is crucial for facilitating small cell deployment in urban settings in addition to ensuring that networks are able to satisfy latency and capacity needs for 5G [36].

9.2.4.3 Increase in Infrastructure Cost in Future

Mobile operators can actively make investments in 5G infrastructure or they can delay investing in 5G as long as it seems manageable while network/infrastructure is upgraded. Regardless of the approach taken by mobile operators when it comes to 5G infrastructure investment, an increase in infrastructure costs in the future is estimated as increase in traffic is expected and is evident [36].

GSMA analyzed situation in 4 urban cities. The study estimated that the network capacity will fail to meet the demand of data based on the existing assumptions such as regarding financial limitations, network arrangements and level of technological development. It was noted that highly dense urban areas will face lack of satisfaction of around 50% of the network/data demand. The estimated demand will require major infrastructure investment [39].

Furthermore, an analysis was carried out in a country in Europe where cases of three mobile operators were considered who decided to choose the latter approach discussed above, that is, they chose to delay active 5G infrastructure investments. As part of the analysis, it was estimated that there will be a major increase in the total cost of ownership (TCO) for RAN between 2020 and 2025 in comparison to the level in 2018. The figure below exhibits this prediction. For example, TCO would increase by 60% in the scenario where it is assumed that there will be a 25% annual data growth [36].

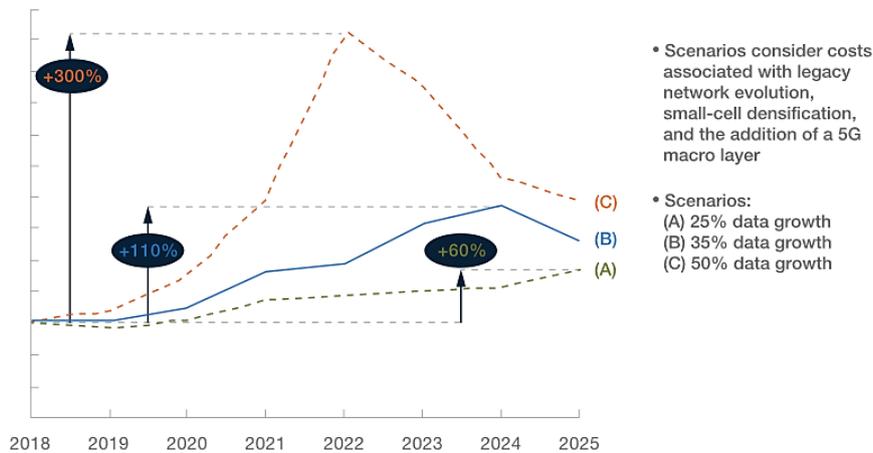


Figure 9.11: Total Cost for Mobile Access Networks [36]

The infrastructure cost for the existing network can be divided into 4 groups: development of new macro sites, deployment of small cells, development of 5G layer and upgradation of the existing network. As mobile operators will go for new macro sites, small cells, and the 5G layer between 2020 and 2025, these areas will form a greater chunk of the total cost of ownership as can be seen in the figure below (assuming 35% annual data growth) [36].

9.2.4.4 Cost Control in 5G Deployments

Where increase in cost is expected for mobile operators when moving towards 5G, certain strategies can be used to control the cost. One strategy is based on network sharing. In general, a typical mobile network operator in Europe can go for network sharing and consequently retrieve around 20% of operational expenditure and reduce the cost of passive Radio Access Network (RAN) components by 50% which typically accounts for half of the total network expense [40].

Operators can decide on the depth or extent of network sharing which refers to small cells or 5G IoT macro layer sharing or coming up with varying sharing models depending upon whether urban or rural areas are being covered [41]. Network sharing in urban areas has the potential of prevention of getting into complicated and lengthy processes for acquiring sites because of regulations. Simultaneously, network sharing in rural areas can decrease the period for getting return on network investment [40].

At the same time, the idea of fixed networks is also worth consideration. This means that a single 5G network can be developed which can be used by all the operators through wholesale access. Ownership of spectrum will still be the deciding factor when it comes to

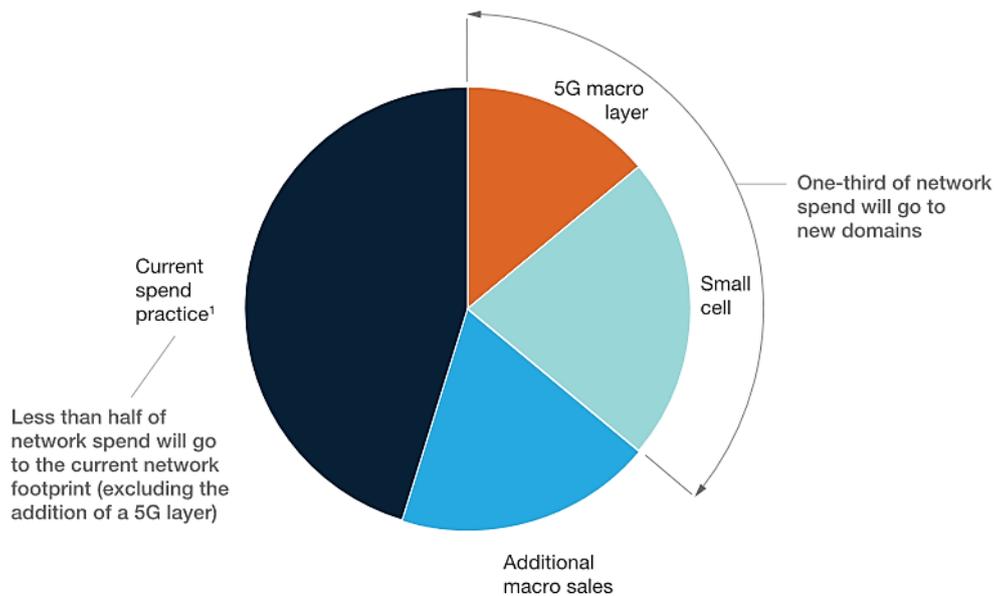


Figure 9.12: Network Expenditure between 2020 and 2025 [36]

entering the market and mobile operators will still compete when it comes to the services that they offer. 5G technology can develop on the current network sharing models that are in place for the lower generation networks such as Multi-Operator Radio Access Network (MORAN) and Multi-Operator Core Network (MOCN) but with the addition of new features like network slicing that enables allocation of resources depending upon traffic or use cases among mobile operators [41].

When deciding to operate individually and separately from other operators, mobile operators can not only diminish their profits but also experience physical limitations when going for network densification in urban areas. This is being said considering the installation of required equipment and underground fiber transmission network for the densification of the network in urban areas that are already crowded and will have to experience major physical disturbances. Moreover, cost reduction accompanied with better network quality are key factors that encourage network sharing among operators. For example, estimates suggest that 50% of the cost can be decreased in the case of each mobile operator if three mobile operators share the network. The figure below shows that 5G investments can be lowered by more than 40% through sharing of 5G small cell deployment and development of a common 5G IoT macro layer on a national level, as per simulations from a case [41].

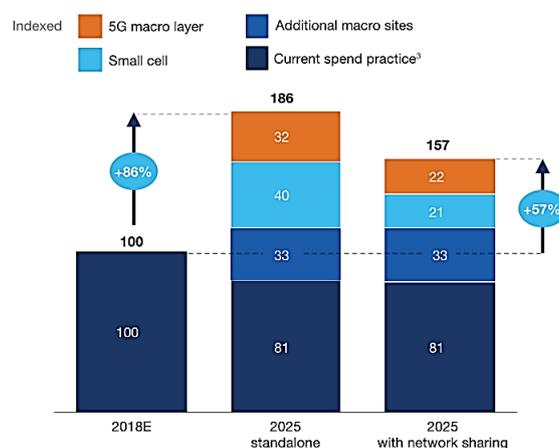


Figure 9.13: Change in cost with network sharing [41]

Moreover, the strategy of network sharing is also important in terms of visual pollution and urban disruption which can increase with greater deployment of fiber and installation of infrastructure equipment in urban settings especially when developing the infrastructure for 5G [41].

9.2.5 Challenges in Implementation of 5G

In addition to the cost considerations discussed above, there are other challenges when it comes to the implementation of 5G. Some of the major challenges are as follows:

1. Small Cell Deployment

Numerous administrative and financial requirements and obligations on operators as part of a country's regulations and policy may slow down the process of small cell deployment. Long permitting processes can take upto 24 months in the case of some countries as local authorities take time to approve applications for small cell deployment. Operators may also experience long procurement procedures as local authorities take time to grant them the right to install the small cell equipment on street furniture. Moreover, high fees to install equipment on street furniture and outdated regulations provide further hindrances [42].

At the same time, effort needs to be made on identification of numerous locations where small cells are to be placed. This can also be a cumbersome process [43].

2. Fiber Backhaul Deployment

Lack of adequate availability of fiber networks will be a significant obstacle as deployment of fibre backhaul networks is needed for small cells [42],[38]. For example, UK has 2% fiber penetration rate, one of the lowest fiber penetration rates in Europe while the European average is 9%. Considering the feasibility in terms of cost and revenue when it comes to implementation of fiber backhaul, operators may think about wireless backhaul technologies [42].

3. Spectrum Allocation

Harmonized allocation of spectrum is important. The advantages of harmonized allocation include minimization of radio interference along border and facilitation of international roaming. Coordination and understanding between telecommunication organizations, national regulatory authorities and global community is required for identification and allocation of spectrum globally. Hence, ensuring this coordination and understanding may emerge as a major challenge [42].

4. 5G Compatible Devices

Widespread availability of devices that are compatible with 5G standards and spectrum is fundamental for the implementation and success of 5G [42]. Availability of 5G compatible devices is heavily dependent upon how expensive it can be to make them [34]. Hence, shortage in devices that are compatible with 5G will hinder the widespread use and implementation of 5G.

9.3 Security Aspects of 5G

9.3.1 Similarities and Differences in 5G Security compared to older 4G Technologies

There are multiple quite similar technological design principles in 5G as there are in 4G, which also leads to 5G taking over some limitations and risk factors in the area of security.

In general, the connection goals are almost the same, trying to ensure following points [44]:

- Authentication
- Integrity
- Privacy (Data, Location, Identity)
- Availability

There are of course many types of attacks that are achieved in the same way as in 4G, since the 5G technologies' backwards compatible radio technologies such as 2G, 3G, and 4G let 5G inherit the underlying security issues [45] [46]:

- **Device Threats**
Bots, DDoS, Man-in-the-Middle, Malware, Tampering, etc.
- **Air Interface Threats**
Jamming, Eavesdropping, MitM, etc.
- **Network Threats**
Access Control, Rogue Nodes, API Vulnerabilities, App Server Vulnerabilities, etc.

The main differences lie in how the newer 5G technologies try to achieve these goals and defenses, as well as what they have to do differently to older systems. Where the 3G and 4G end-to-end authorization was simple, efficient, and mostly secure enough, the era of the IoT complicates this approach. While 4G connections are used almost exclusively by mobile devices such as smartphones, the IoT aspect introduces a huge influx of new devices to the network. The 4G authentication (network-based hop-by-hop) and connection becomes far too inefficient of a design principle [47].

The higher density of devices, that is handled by a similar increase in density of Base Stations in the form of small cells, calls for different approaches in handling the security issues. Firstly, the infrastructure has to be robust and secure [47]. Home eNode B (HeNB) femtocells can be physically altered/attacked, which could affect both end-user and mobile operators. Cells need to be unapproachable by unauthorized third parties, which would mean they have to be either in hard to reach places (e.g. such as on the side of walls, a few Meters above the ground) or physically protected places (e.g. behind locked doors). This would also protect against physical relocation of the femtocell. Physical protection might further give the opportunity to store credentials on a device in proximity of the cells. Another option is to store the credentials in a separate, protected domain.

Other security issues are the configuration and protocol attacks, against which it gets harder to protect than in the 4G technologies because of the principle of dividing different services into different security options [45]. While opening opportunities of splitting and separately investing into research and security when these services are clearly split, it is hard to tell in which ways it is beneficial for the mobile communication aspects.

9.3.2 Importance of Trust and other security standards in 5G

The robustness of the 5G network is a very critical point in the success of it. The basic ideas for network infrastructure include virtualization and the more traditional approach of physical alteration to the network access point. Virtualization is used to keep the networks efficient and the investment costs low. If 5G was, similar to 4G, structured so that every node is physically isolated, a large network would become impossible to manage in terms of scalability, antenna correlations, and coupling [49].

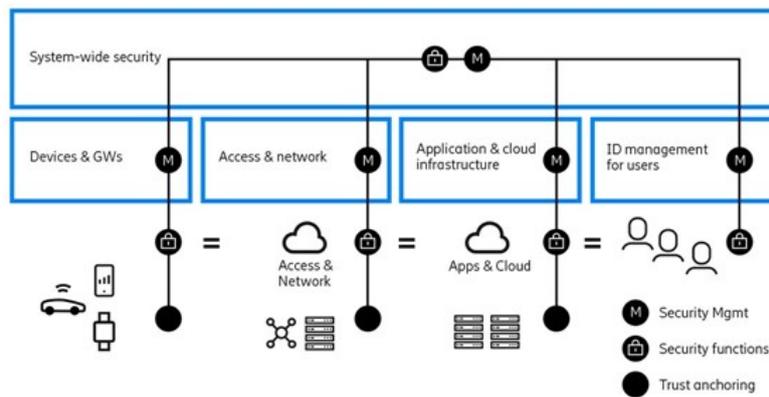


Figure 9.14: Ericsson's System-wide Security [48]

Virtualization would mostly take place through network slicing, which allows for flexibility when choosing how to implement security for different services. There would be no additional costs for infrastructure since the same nodes can be used by all of the desired services split into different virtual networks, which would lead to possibilities of dividing services into e.g. IoT, Mobile Broadband, and low-latency applications (e.g. vehicular communications) [50].

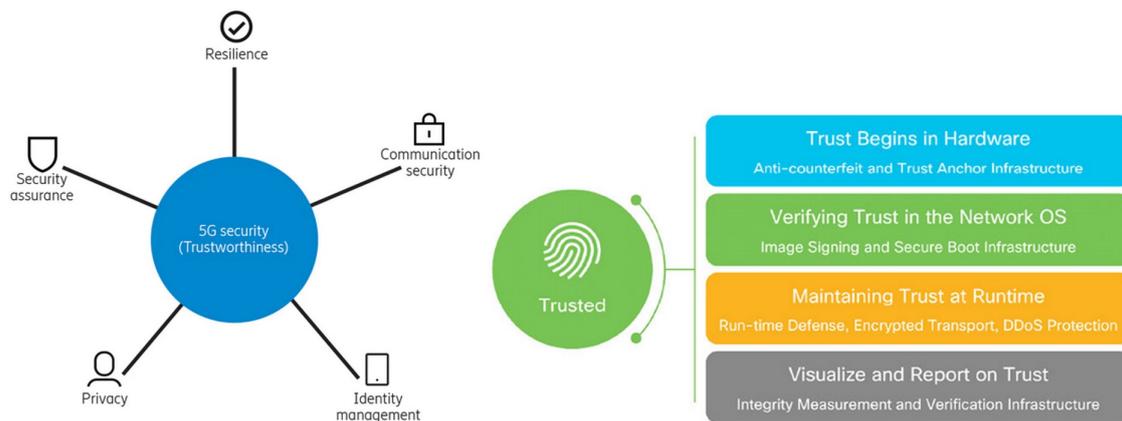


Figure 9.15: Ericsson's Trust Model on the left, Cisco's Trust Model on the right [51][52]

Apart from infrastructural robustness, there are new Trust Models, which define key points in Authentication Management. While having multiple varieties of Trust Models, usually different for each provider of these services, the main points are rather similar. Main features contain both physical robustness as well as the resilience of the software aspect, which contains security management regarding communication, identity, privacy (location, data, etc.). Depending on the provider, additional solutions are offered, one of which is a fast response to actual attacks or malfunctions such as Ericsson's PSIRT (Product Security Incident Response Team). Usually the provider takes care of both implementation and security, more in the economical chapter.

9.3.3 Life and 5G: The Security of Organic Beings

Many news sites have reports about the danger of having so many stations radiating with a high power in densely populated areas. Doctors are warning about cancer and various other organic problems, which could be caused or accelerated by this radiation.

While it is currently unclear whether this could become a problem in the near future, companies investing in these technologies and their stakeholders need to be careful about

a possible backlash of the population. Riots and vandalism could quickly shake up the industry of 5G as a whole, should the radiation prove to be a significant health risk.

Current investigations assume that the effect is negligible if there is nothing organic in the immediate and close vicinity of a cell. However, it remains to be seen if the cells in a densely populated city can be arranged in a meaningful and harmless way. If compared to 4G technologies, this does not surprise, since similar effects have been observed with older technologies as well.

Multiple studies investigating health risks from 5G are underway. Although they could shed some light onto the whole debate of 5G being a risk or not, it has to be noted that it will still be only a short-term study. Long-term effects would have to be studied over a longer time period for a conclusive result about its effects on organic beings. At the time of writing this paper, there are no recognized papers on this subject released, the topic itself is mostly found on news report websites mentioning the marches against 5G [53][54][55].

9.3.4 Economical Aspects of Security

”Security as a service (SECaaS) is an outsourced service wherein an outside company handles and manages your security. At its most basic, the simplest example of security as a service is using an anti-virus software over the Internet [56].”

With the vertical separation of technology, while still being in the same network, comes the opportunity of offering SECaaS in different areas. While some companies might offer both the infrastructure as well as the security measures (e.g. Ericsson), others might only offer security measures (e.g. Anti-Virus Software against virtual threats).

In both scenarios, the 5G network opens many business opportunities and structures. Companies can offer a multitude of either specialized (e.g. a service of patient-doctor interaction with biometric authentication [47]) or broad (e.g. Ericsson’s IoT platform [57]) security services. As mentioned when discussing the Trust Models, the 5G network opportunities include both hardware and software specific aspects of products. It remains to be seen what further possibilities may arise in combination with nationwide networks, where a provider offers network access points across the country with their own security measures which are not as extensive as what private solutions may provide. Security software will be one of the most likely contenders, another one being hardware (such as specialized routers) that has their own security encoding for enhanced safety.

In conclusion, because of the IoT aspect in 5G, early investments in this area might be profitable, since this trend is likely to break through as the future of technology and communication. It remains unclear however, in which cases the change to 5G technologies is worth it from a security point of view. The opportunities for SECaaS to penetrate the market is surely there, but with the little experience we currently have considering all the aspects around infrastructure, virtual networks, verticality, etc. in 5G specifically, it remains to be seen if the change (from a mobile communication standpoint) makes sense.

9.4 Case studies

Prior to discussion about the case studies, let us take a brief overview of Industrial 5G and its implications.

9.4.1 Industrial 5G

Industry 4.0 with intelligent factories and the Industrial Internet of Things (IIoT) - this is the future of industrial manufacturing. Making production facilities and intralogistics

more flexible, autonomous, and efficient requires appropriate communications framework conditions and comprehensive connectivity. The new 5G communication standard is making new progress in this area. One of the benefits of 5G is the significantly higher bandwidth that allows much more data to be sent simultaneously than previously. Estimates start at ten gigabits per second which is 10 times more than with 4G. Moreover, 5G will have a considerable lower latency and greater reliability than existing mobile technologies. Also, the number of subscribers that would be able to be connected in a defined area will also be higher [58].



Figure 9.16: Industrial 5G- wireless network of the future

On the path to a new era: How 5G will change industry: A factory site in 2025: Goods, spare parts, and finished products are being transported between delivery bays, production facilities, and warehouses by a fleet of autonomous vehicles which is precisely coordinated with the manufacturing schedule. Infinite devices are connected with each other in production and transfer data from the entire production line in a matter of milliseconds. Cameras on a conveyor belt can, for example, recognize a foreign body and bring a robot to rest instantly. The field engineer is able to carry out remote maintenance and service tasks easily and effectively using augmented reality without leaving the site. According to the Mobile Economy 2019 Report (GSMA), 15 percent of the world's mobile communications will be operating on 5G as early as 2025. Huge talks are about a wireless network which can include many things because of its bandwidth: From automated racking systems, and production robots, to air conditioning systems, and control panels. An all-encompassing network which allows an industrial plant to be controlled wirelessly - reliable, super-quick, or equipped with very high bandwidth. Therefore, Siemens has engaged itself to this new communication standard from the outset in order to support standardization and industrial implementation.

Added value for users and industry: Looking back on the development of mobile networks over the last 40 years shows that they have always generated added value for users and industry alike. Even the first commercial mobile network, in other terms, the first-generation network (1G), allowed us to talk to each other while, in other words, mobile telecommunication. 2G networks boosted the arrival of text messages, 3G provided the Internet into people's hands, and 4G upgraded with music and video streaming. However, for industry, 1G applications might be very expensive because of analog voice transmission, and limited network coverage. The next generation, 2G, brought text messages into display including simple data transmission for industrial telecontrol applications. 3G

allowed long-distance activity and remote access, for example, in teleservice where users could interact with remotely installed applications. 4G provided high-performance mobile remote access and that is not the end of the networking. 5G wireless mobile communications would possibly bring further improvements, focusing on increased bandwidth, greater reliability, lower latency, and many more connected devices.

A vision with three key scenarios: The "3rd Generation Partnership Project (3GPP)" is responsible for the global standardization of mobile networks, including the 5th generation. The 5G vision was established in an initial phase of the development of the most recent standard. This consists of three key scenarios or aspects which are anticipated for 5G mobile networks. The first key scenario, Enhanced Mobile Broadband (eMBB), includes improvements with respect to 4G. The main objective is the realization of data driven applications which need high data rates with large-scale network coverage. A common example is the growing need for HD high-quality streaming of music and videos on mobile device, for example, smartphones.

It is also possible to foresee augmented-reality (AR) applications for industry which could be supportive to field engineers. The second scenario, Ultra-Reliable Low-Latency Communication (URLLC), entails demands for high reliability and low latency industrial applications which prove to be challenging. For instance, mobile robots, autonomous logistics, automated guided vehicles (AGVs), or safety applications. The third scenario, massive Machine-Type Communication (mMTC), focuses on connecting a large number of devices in a small space. In practice, this means applications for the Industrial Internet of Things (IIoT) where a unit area has a high device density. The devices send or receive data continuously at longer intervals to make sure that the smallest bandwidth is being utilized. Another example includes processing industry where many sensors are installed (e.g. to measure temperature, pressure, flow) such as to support process monitoring in a plant.

Actual implementation and its limits: Despite of the many potential advantages of 5G, we should understand that not all of the functionalities will be available right from the start and could not be grouped together in most of the cases. A series of capabilities were defined which 5G had to have in order to satisfy the stated objectives from the three key scenarios. This comprises downlinks with peak data rates of 20 Gbits/s and a maximum latency of 1 millisecond, as well as specifications with respect to mobility, density, energy efficiency, spectrum efficiency, and area network capacity. 5G has a phased release schedule to allow for these commitments and the prescribed timeline for the new standard.

Spectrum as a key factor: The provision of 5G has a higher spectrum requirement than previous generations of mobile communications. The spectrum is owned by governments. Few of them are royalty-free such as ISM bands, Industrial Scientific and Medical but for mobile networks, governments places the frequencies on auction to mobile network operators in order to build national public networks. However, industrial 5G permits the network to be customized to every application. For example, URLLC and mMTC could be more beneficial than eMBB for different industries. In private networks, the end user can determine which parameters are set and operate the network in a way which is best for the specific application. However, sufficient spectrum must be available to industry for these types of private networks. In Germany, the Bundesnetzagentur (Federal Network Agency - BNetzA) has decided to reserve 100 MHz from 3,7 GHz to 3,8 GHz for local industrial use. This provides German companies with the opportunity to lease a spectrum for an annual payment for exclusive use in their own operating facilities and hence to ensure data protection [59].

9.4.2 Case studies on 5G

9.4.3 Case 1: Trade war between the US and China

- One of the key reasons behind the U.S.'s trade war with China is the desire to get ahead in 5G.
- The technology is foreseen as a backbone from everything from driverless cars to future cities.
- America and China are in a race to become the leader in 5G and set the standards that will define the next generation of mobile internet.

The story behind:

There is a key technology could actually be one of the main reasons behind why the world's largest economies have started a trade war. That technology is 5G – the next generation of mobile internet. It could make easier for people to download movies in few seconds, And it could lead to a good mobile internet experience. However, 5G is much more than high-speed mobile internet. It is being observed as a technology that could support the next generation of infrastructure, from the billions of internet-connected devices which are expected to come online in the next few years, to smart cities and driverless cars.

What the countries think is at stake:

Mobile internet demands 'standards' that can be agreed upon globally so that companies that make telecom equipment, as well as mobile carriers, can deploy the technology around the world.

Now the real race is on as tech firms like ZTE and Huawei and European companies such as Nokia and Ericsson move ahead to take a lead in 5G. U.S. chipmakers like Qualcomm and Intel are also pulling up their socks to move ahead in 5G. According to the experts, it is about who is going to define and control the model, the architecture, and the agenda of 5G. At stake is trillions of dollars of economic value is at stake eventually. In 2035, 5G is expected to enable 12.3 trillion dollars of global economic output, according to IHS Markit [60].

How 5G is related to trade war between the US and China:

5G will serve as the foundation to support the next generation of infrastructure, comprising billions of internet-connected devices powering smart cities, wondering new VR and AR applications and driverless cars. Because of these reasons, President Donald Trump wants the US to lead in 5G.

What's at stake is likely to determine whether the US will continue to maintain its technological edge and shape geopolitics for the next couple of decades or if it'll cede that control to China, which is already working hard in technological dominance in order to become a world superpower.

Huawei is right in the middle of this trade war. A year ago, most Americans had likely never heard of the tech giant. Now it is in the news nearly every day as a highlight of the US-China trade war. Huawei is the one of the main supplier in the 5G market. However, national security experts claim that the company is closely tied to the Chinese government that could prove dangerous to the US and its partners. The reason being Huawei's gear is suspected to shut down critical communications giving rise to potential future conflicts.

Huawei is also representative of a bigger issue the US is facing. As China continuously working to transform from a country known for making toys to one that leads in advanced technologies like artificial intelligence, autonomous vehicles, robotics and 5G, it's accused of a state-led industrial policy that US intelligence officials say relies on intellectual property theft, forced technology transfers, cyber crime and discriminatory treatment of foreign investment. Therefore, this concern has led to unsupported trade practices

resulting in increased tariff on the Chinese goods by Trump along with blocking Huawei from access to US markets.

Reasons for the race to 5G:

1. The country which will lead in the development and deployment of 5G technology in the near future will see more economic growth and will have more power. The leader of 5G would be expected to gain hundreds of billions of dollars in revenue over the next decade, with multiple job opportunities across the wireless technology sector. For the US, it is to maintain the technological and economic lead it gained with its 4G wireless technology. But for China, it's an opportunity to surpass the US and the West to become the economic and geopolitical superpower as it wanted to be for long.
2. Winning 5G is not so much a 'race' as it is a process. Characterizing 5G as a contest specifies its great technological progress and the policy challenges that progress presents. 5G should be more than a political talks; the new network should represent the need for a meaningful policy strategy.

How tech giant Huawei is closely related to this trade war:

Huawei is one of the biggest developer of 5G equipment, and its technology is also considered to be the most advanced. And it is the second largest smartphone maker behind Samsung, having surpassed Apple last year.

The company, founded in 1987 by a former officer of the Chinese People's Liberation Army, is suspected to have close ties to the Chinese government according to CIA, Directors of FBI and Team of National Security Agency(NSA). However, Huawei on their end has denied these claims. The company has also been accused by the US Justice Department, in indictments that included 23 counts of alleged theft of intellectual property, obstruction of justice and fraud. Therefore, Trump blacklisted Huawei, adding it to the US "entity list," which ceases it from buying US products and services. (Including Google mobile services as well.) [61].

Possible discussions between the US and China to settle on trade deal:

The image shows a screenshot of a CNBC news article. The top navigation bar includes 'MARKETS', 'BUSINESS', 'INVESTING', 'TECH', 'POLITICS', and 'CNBC TV'. The main headline is 'China and US had 'constructive discussions' about phase-one trade deal'. Below the headline, it says 'PUBLISHED SUN, NOV 17 2019-12:23 AM EST | UPDATED SUN, NOV 17 2019-10:37 PM EST'. The author is Spencer Kimball (@SPENCERKIMBALL). There are social media share icons for Facebook, Twitter, LinkedIn, and Email. A 'KEY POINTS' section lists three bullet points: 1. The two sides had "constructive discussions" about "each other's core concerns" and agreed to remain in close contact, according to the Chinese Ministry of Commerce. 2. The call came at the request of Treasury Secretary Steven Mnuchin and U.S. Trade Representative Robert Lighthizer, the ministry said. 3. The Dow Jones Industrial Average closed at record highs Friday on renewed optimism about trade talks.

Figure 9.17: discussions for "phase-one" trade deal

The two sides had settled on terms of "constructive discussions" about 'each other's core concerns' and agreed to remain in close contact, according to the Chinese ministry. There have been conflicting reports about the state of trade negotiations in recent days. The discussions hit a statement this week as the U.S. pushes Beijing for greater concessions on intellectual property rights and forced technology transfers in exchange for a rollback of tariffs. If both sides reach a phase one agreement, the level of tariff could be taken back by the US [62].

Trump threatens higher tariffs on Chinese goods if the latter does not agree on a trade deal:

US President Donald Trump threatened higher tariffs on Chinese goods if China does not agree on a trade deal. Financial markets, which have proven reactive to developments in the ongoing trade war, largely shrugged off Trump's latest warning. The U.S. and China were about to agree to the aspects of 'phase one' trade deal in October 2019, but the officials from both the countries sent mixed signals about the deal leaving it in a absurd position [63].



Figure 9.18: Trump threatening China to settle on phase-one deal

China's reaction to Trump's warning:

Chinese officials and media remained silent even after US President Donald Trump threatened for the second time to "raise the tariffs even higher" on Chinese imports if a trade deal could not be accomplished. China's no reaction to Trump's warning has boosted the 17-month trade war eventually slapping a tariff of billions of dollars on each other's goods [64].

China remains silent as Trump repeats threat to hike tariffs if trade war deal not reached soon

- Chinese officials and state media have been silent on threats from US president to raise tariffs on Chinese goods if 'phase one' trade deal not settled soon
- Senior White House officials still optimistic a deal can be reached to end 17-month tariff war

Figure 9.19: China's reaction to Trump's warning

9.4.4 Case 2: Huawei ban in UK

UK government has imposed a ban on Chinese telecom supplier Huawei from supplying core parts of the future 5G mobile phone network, after decision being taken at meeting of ministers of the National Security Council (NSC). Britain's intelligence agencies are also getting cautious with Huawei but have not called off a complete ban. UK's spy agency is trying to figure out the "opportunities and threats" caused by the Chinese technology. Ministers have raised worried concerns posed by Huawei at the NSC meeting. The decision came into effect after the meeting of the Britain's spy agency as they were getting indications of a possible cyber attack from China or Russia where the useful information could be revealed in a matter of seconds. The UK is also concerned about how the spy agency will affect the British business and consumers. The UK government is

not expected to disclose the name of any countries as a specific threat, but hackers from Russia, China, North Korea and Iran are generally considered by the intelligence agencies the most likely sources of danger to British cybersecurity [65].

9.4.5 Case 3: 5G in South-East Asia

1. Chinese phone operator Huawei made a statement that it is ready to emphasize on 5G infrastructure across South-east Asia, apart from US warnings that its tech could be used to misuse data. The firm has emerged out as key factor in the US-China trade war that has seen tit-for-tat tariffs imposed on hundreds of billions of dollars worth of goods.

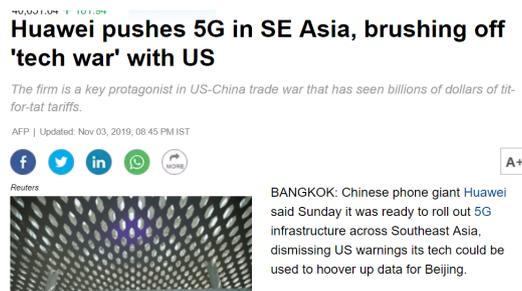


Figure 9.20: 5g in South-East Asia

President Donald Trump's administration has warned that Huawei's equipment could allow China to spy on other countries and has effectively ceased American companies from selling US technology to the firm. However, Huawei has continuously denied the accusations with the fact that there is some other reason such as tech envy apart from trade war. Thailand and the Philippines have likely ignored the cybersecurity warnings in order to deploy the ultra-fast 5G network provided by the China's biggest smartphone maker, while Vietnam stayed away from Huawei. Huawei president said that they are facing some unproved conflicts and they will support the countries to grow a level ahead in the field of communication rather it's south-asia or some other countries.

2. Host country Thailand in south-asian summit has welcomed Huawei with open arms, permitting it to set up a test bed at a major university near the Thai capital. Huawei had already invested US 5 billion dollars in the trials and has been invited to conduct similar tests in other Southeast Asian markets.
3. The Philippines' Globe Telecom is also planning to launch South-east Asia's 5G broadband service using Huawei technology.
4. However, not all countries have been eager to sign up. Vietnam has quietly taken side of the US on the burning issue avoiding the Chinese tech firm and taking decisions favoring some alternative technologies like Ericsson and Nokia [66].

9.4.6 Case 4: Germany deploying 5g with Huawei and situation in some other European countries:

- Germany will not ban Chinese telecom giant Huawei from helping to build its national 5G networks, ignoring the actions from the US to ban the company over serious security concerns. The move from Germany is a big blow to the U.S., which has been pressuring its partners to exclude Huawei from 5G infrastructure, claiming that it is on the front for Chinese espionage. Huawei has already tested few 5G

trials with Deutsche Telekom and also known to have supplied current equipment to all of Germany's telecoms operators. Germany's decision to allow Huawei to deploy 5G using its network could also have consequences for the rest of Europe.

Germany set to allow Huawei into 5G networks, defying pressure from the US

PUBLISHED WED, OCT 16 2019-7:23 AM EDT | UPDATED WED, OCT 16 2019-10:02 AM EDT

Chloe Taylor
@CHLOETAYLOR141

SHARE [f](#) [t](#) [in](#) [✉](#)

KEY POINTS

- Germany will not ban Chinese telecoms giant Huawei from its national 5G networks, snubbing calls from the U.S. to exclude the company over national security concerns.

Ad

Figure 9.21: 5g in European countries

- Other EU member states, including France and the U.K., are still thinking on new agendas to employ Huawei. Meanwhile Britain is also planning to re-consider its decisions on the Huawei. France is also seeking collaboration with Huawei but with restrictions on the table in order to preserve its national security [67].

9.4.7 Case 5: Huawei introducing 5G in Switzerland:

Sunrise has taken the lead to build 5G network in Europe by shaking hands with its strategic partner Huawei. Sunrise and Huawei will work together under the 5G innovation center, which is first in Europe, to research and develop 5G applications for both the private and business sectors.

Sunrise and Huawei Open the First European 5G Joint Innovation Center

Oct 14, 2019

in [f](#) [t](#) [t](#)

[Opfikon/Dübendorf, Switzerland, October 14, 2019] Working together with its strategic partner Huawei, Sunrise has built the leading 5G network in Europe. Now with the first 5G Joint Innovation Center in Europe, Sunrise and Huawei will work together to research and develop 5G applications for both the private and business sectors. The joint innovation center will additionally help to build a Swiss 5G ecosystem, using Sunrise's headquarters in Opfikon to introduce live 5G application scenarios that have already been launched or are about to be commercialized.

Figure 9.22: Sunrise for 5G in Switzerland

- Sunrise 5G use cases:

1. Smart farming: Environmentally friendly technology for increased production:

Researchers are using the new 5G network to help farmers optimize their cows' milk production and track their feeding behavior. The aim of this implementation is to acquire highest possible production.

Moreover, 5G transmission provides high-resolution data transmission in real-time. This ultimately allows farmers to directly monitor the calving process of their cows, by using a high-resolution camera. Therefore, it is also working on identifying the appropriate time for insemination. This technology is still under some testing at the moment and is being carried on to real-life testing.

On the other hand, on the field, drones could be deployed that can send data through the antennas to cloud for processing thus allowing farmers to see the results in real time. As an example, the drones could be used to detect the nitrogen content

of plants, so that fertilizer can be applied in a more precise and resource-efficient manner.

2. Smart manufacturing: 5G as a key factor for industrial digitization:

With Sunrise's 5G network, GF Machining Solutions in Biel laid the new foundations for the factory of the future. 5G gateways were installed and rolled out across the entire factory floor. It could be beneficial in the way that machines could be placed anywhere in the factory. Download speeds and transfer rates could drastically rise with 1.1 Gbps on the factory floor; which is about ten times faster than the speeds that were being recorded before the gateways were installed. All of this means that GF Machining Solutions could achieve to wirelessly and securely connect their machines to their cloud services and infrastructure while taking benefit from minimal latency.

Thus, the 5G network allows predictive maintenance in almost real time. This concept enables the particular system to predict when a machine needs maintenance and intervention, and unplanned downtime can be reduced or avoided altogether, thus improving machine uptime and reduce costs [68].

9.4.8 Forecasted Dominators in 5G technology

1. China, United States, Japan and Korea would be growing in numbers to account for more than half of the world's subscribers to 5G mobile networks by 2025, leaving Europe lagging behind. Europe is lagging because of consumer take-up which have impact on its slow progress over deployment of 5G. Yet the picture will be different in businesses, where 5G will be capable of running 'smart' factories using connected robots, devices and sensors.

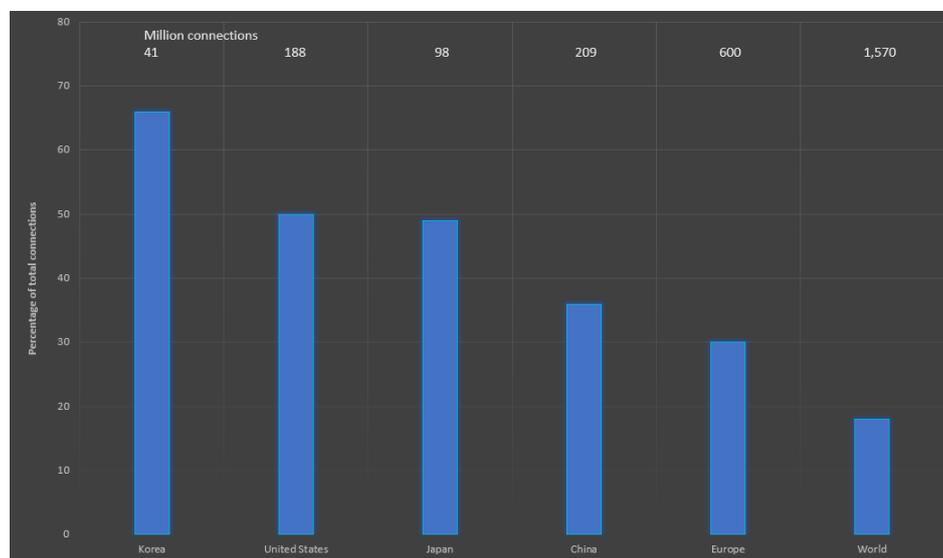


Figure 9.23: Forecasted dominators in 5G technology

2. In Korea, 66 percent of mobile connections will be on 5G network by mid-decade, stated in GSMA Intelligence forecast in a 100-page study, followed by the United States on 50 percent and Japan on 49 percent.
3. 1.57 billion people worldwide are expected to adopt 5G by 2025 - or 18 percent of total mobile users. Past experiences shows that mobile carriers can hike 5G prices by 15-20 percent, offering 'more for more' unlimited data plans. But, if the past is anything to go by or being concerned then those data plans for more gains will definitely be competed by other operators. [69].

9.5 Conclusion

The former British Prime Minister Benjamin Disraeli had once remarked that "Change is Inevitable. Change is Constant." Nowhere is it as true as in the case of the telecommunication industry, and in particular, wireless cellular networks. However, as with any new technology, the pros and cons need to be weighed and considered before concluding on the economics of 5G networks.

The concerns with 5G are primarily related to high investment costs, health risks and the security architecture of 5G networks. Given the high investment costs of installing new small cells and antennas, the argument that this might lead to increasing disparity between developed and developing countries, cannot entirely be negated. Further, as investing in rural areas might be prohibitively high in some countries (e.g. India /USA), 5G networks can also lead to further disparity in service available within the same country - e.g. healthcare, industrial and even educational.

The health impacts of 5G are not yet fully clear [23]. A recent study by the National Toxicology Program in the US found that high exposure to 3G radio frequency radiation (RFR) led to some cases of cancerous heart tumors, brain tumors, and tumors in the adrenal glands of male rats, and hence the suggestion that similar exposure to RFR is also harmful for humans, cannot completely be ruled out [24]. As 5G networks will require a larger number and higher density of small cell towers and antennas, the health risks, and its corresponding economic implications cannot be eliminated. Further, the security architecture of 5G networks are far more complicated than that for previously used 4G or 3G networks. This, therefore, raises the risk that 5G equipment can be compromised by government agencies and used to spy on people.

Further, as of 2018, only 5 countries had a network coverage area greater than 90% that supported 4G. Among large countries, the US was the only country to feature among the top 10 countries [26]. Moreover, when it comes to speeds, no country has managed to achieve the maximum theoretical capacity of 50 Mbps, and instead maximum speeds have been limited to 90% or 45 Mbps. It would therefore not be unfair to say that the fullest possible utilisation of 4G is yet to be seen. In this context, therefore, it is but natural that the 5G networks that have been launched are more of an extension of existing 4G networks, operating in the MHz bandwidth at a fraction of the maximum obtainable speeds. However, recent test results in South Korea [27], and more recently by T-Mobile in the US [28], gives hope that the promised benefits of higher speeds and lower latency are not far away. The 5G wave is coming, but if economies are to fully ride it, they need to ensure complete utilisation of existing 4G bandwidth while investing in technologies that will enable GHz range bandwidths for 5G.

Despite the concerns and limitations mentioned above, the fact that 5G is inevitable is beyond debate. The economic and social benefits that are likely to materialize as a result of high data speeds and low latency itself are sufficient to push for its implementation - Autonomous vehicles, interconnected industries, smart homes, and augmented and virtual reality, have the possibility of transforming societies in a significant way and 5G is at the very heart of this transformation [25]. What is therefore required is that governments, international telecommunication agencies and companies work together to take necessary steps to alleviate the concerns and limitations related to 5G and ensure that it becomes an enabler for all of mankind.

Bibliography

- [1] 5G Race Could Leave Personal Privacy in the Dust; <https://www.wsj.com/articles/5g-race-could-leave-personal-privacy-in-the-dust-11573527600>, Last visit: December, 2019.
- [2] Bertschek Irene, Briglauer Wolfgang, Huschelrath Kai, Kauf Benedikt, Niebel Thomas: *The Economic Impacts of Telecommunications Networks and Broadband Internet: A Survey*, ZEW - Leibniz Centre for European Economic Research, August, 2016. <https://www.econstor.eu/bitstream/10419/145104/1/866030050.pdf>.
- [3] Peter Boyland: *The State of Mobile Network Experience*, Opensignal, May, 2019. https://www.opensignal.com/sites/opensignal-com/files/data/reports/global/data-2019-05/the_state_of_mobile_experience_may_2019_0.pdf.
- [4] Snap Market Cap; https://ycharts.com/companies/SNAP/market_cap, Last visit: December, 2019.
- [5] Future Mobile Data Usage and Traffic Growth; <https://www.ericsson.com/en/mobility-report/future-mobile-data-usage-and-traffic-growth>, Last visit: December, 2019.
- [6] Huawei Plans \$600m Investment in 10 Gbps 5G Network; <https://www.independent.co.uk/life-style/gadgets-and-tech/huawei-plans-600m-investment-in-10gbps-5g-network-8924124.html>, Last visit: December, 2019.
- [7] KT's Millimeter Wave 5G Network Transmitted 3800 TB of Data During Winter Olympics; <http://bit.ly/36GiNXY>, Last visit: December, 2019.
- [8] 5G Latency - Reality Checks; <https://www.senki.org/5g-latency-reality-checks/>, Last visit: December, 2019.
- [9] Eight Reasons Why 5G Is Better Than 4G; <https://connect.altran.com/2018/03/eight-reasons-why-5g-is-better-than-4g/>, Last visit: December, 2019.
- [10] Why We Need 5G Cellular Service; <https://www.asme.org/topics-resources/content/need-5g-cellular-service-part-2>, Last visit: December, 2019.
- [11] 1 Million IoT Devices Per Square Km- Are We Ready for the 5G Transformation?; <http://bit.ly/2Z3btmD>, Last visit: December, 2019.
- [12] Simon Johansen, *1G, 2G, 3G, 4G, 5G*, https://its-wiki.no/images/c/c8/From_1G_to_5G_Simon.pdf.
- [13] LTE Frequency Bands; https://en.wikipedia.org/wiki/LTE_frequency_bands, Last visit: December, 2019.

- [14] 5G; http://sharetechnote.com/html/5G/5G_FR_Bandwidth.html, Last visit: December, 2019.
- [15] Spectrum Allocation in India: Journey So Far; <https://www.downtoearth.org.in/coverage/science-and-technology/all-about-mobile-spectrum-33106>, Last visit: December, 2019.
- [16] Everything You Need to Know About 5G; <https://spectrum.ieee.org/video/telecom/wireless/everything-you-need-to-know-about-5g>, Last visit: December, 2019.
- [17] 1G; <https://en.wikipedia.org/wiki/1G>, Last visit: December, 2019.
- [18] Committed to Connecting the World; http://www.itu.int/net/pressoffice/press_releases/2010/40.aspx#.Xc_dS1xKhPY, Last visit: December, 2019.
- [19] What is 5G New Radio; <https://5g.co.uk/guides/what-is-5g-new-radio/>, Last visit: December, 2019.
- [20] List of 5G NR Networks; https://en.wikipedia.org/wiki/5G_NR_frequency_bands, Last visit: December, 2019.
- [21] Sunrise Expands 5G Footprint to Over 300 Markets”, ; <https://www.telegeography.com/products/commsupdate/articles/2019/11/04/sunrise-expands-5g-footprint-to-over-300-markets/>, Last visit: December, 2019.
- [22] Swisscom Aims for Nationwide 5G Coverage by Year-end, Devices to Hit Market Next Month; <http://bit.ly/2PxsKkL>, Last visit: December, 2019.
- [23] How Worried Should You Be About the Health Risks of 5G?; <https://www.howtogeek.com/423720/how-worried-should-you-be-about-the-health-risks-of-5g/>, Last visit: December, 2019.
- [24] Cell Phone Radio Frequency Radiation; <https://ntp.niehs.nih.gov/whatwestudy/topics/cellphones/index.html>, Last visit: December, 2019.
- [25] The Pros and Cons of Going 5G Wireless; <https://xyzies.com/pros-cons-going-5g-wireless/>, Last visit: December, 2019.
- [26] The State of LTE(February 2018); <https://www.opensignal.com/reports/2018/02/state-of-lte>, Last visit: December, 2019.
- [27] After Seven Months, Here’s What South Korea Can Teach Us About 5G; <http://bit.ly/2raRq9q>, Last visit: December, 2019.
- [28] 5G Speed is Data Transmission in Real Time; <https://www.telekom.com/en/company/details/5g-speed-is-data-transmission-in-real-time-544498>, Last visit: December, 2019.
- [29] IHS Markit: The 5G Economy: How 5G will contribute to the global economy; Report, November, 2019, <https://www.qualcomm.com/media/documents/files/ihs-5g-economic-impact-study-2019.pdf>
- [30] N. Kshetri: The Economics of the Fifth Generation Cellular Network; IT Professional, 21(2), March, 2019, pp.77-81, <http://bit.ly/2Qn7gY5>

- [31] 5G - Connection Density - Massive IoT and So Much More; <http://bit.ly/2FgV58Z>, Last visit: January, 2020.
- [32] N. Kshetri: 5G in E-Commerce Activities; IEEE IT Professional, 20(4), 2018, pp.73-77, https://libres.uncg.edu/ir/uncg/f/N_Kshetri_5G_2018.pdf
- [33] Network Slicing; <https://www.ericsson.com/en/digital-services/trending/network-slicing>, Last visit: January, 2020.
- [34] Five of the biggest challenges facing 5G; <https://www.futurithmic.com/2019/02/26/five-biggest-challenges-facing-5g/>, Last visit: January, 2020.
- [35] China is vastly outspending the US on 5G infrastructure, expert says; <https://cnb.cx/2uceHZv> Last visit: January, 2020.
- [36] The road to 5G: The inevitable growth of infrastructure cost; <https://mck.co/35yLtSx>, Last visit: January, 2020.
- [37] 5G Spectrum: GSMA Public Policy Position, Report, July, 2019, <http://bit.ly/36rqHV1>
- [38] MdM. Ahamed and S. Faruque: 5G Backhaul: Requirements, Challenges, and Emerging Technologies, in *Broadband Communications Networks: Recent Advances and Lessons from Practice*; chapter from book, September, 2018, <http://bit.ly/2Q1xH0j>
- [39] Will mobile infrastructure keep up with rising demand?; <http://bit.ly/2MT9cFU>, Last visit: January, 2020.
- [40] K. Samdanis, X. Costa-Perez and V. Sciancalepore: From network sharing to multi-tenancy: The 5G network slice broker. IEEE Communications Magazine, 54(7), July, 2016, pp.32-39, <http://bit.ly/2SQ1MH9>
- [41] Network sharing and 5G: A turning point for lone riders; <https://mck.co/38Y9QLI>, Last visit: January, 2020.
- [42] Itu.int.: Setting the Scene for 5G: Opportunities & Challenges; Report, 2018, <http://bit.ly/2tz6heA>
- [43] Why 5G networks can't succeed without a small cell revolution; <https://pwc.to/2ZN6RRU>, Last visit: January, 2020.
- [44] Ferrag, M. A., Maglaras, L., Argyriou, A., Kosmanos, D. & Janicke, H.: Security for 4G and 5G cellular networks: A survey of existing authentication and privacy-preserving schemes; Journal of Network and Computer Applications, Volume 101, January, 2018, pp. 55-82, <http://bit.ly/36kx78B>
- [45] Mantas, G., Komminos, N., Rodriuez, J., Logota, E. & Marques, H.: Security for 5G Communications; Fundamentals of 5G Mobile Networks, May, 2015, pp. 207-220, <http://bit.ly/39xrGFs>
- [46] Khan, K., Kumar, P., Jayakody, D. N. K. & Liyanage, M.: A Survey on Security and Privacy of 5G Technologies: Potential Solutions, Recent Advancements and Future Directions; Article in IEEE Communications Surveys & Tutorials, July, 2018, <http://bit.ly/35my0wd>

- [47] 5G Security: Forward ThinkingHuawei White Paper; https://www.huawei.com/minisite/5g/img/5G_Security_Whitepaper_en.pdf, Last visit: November, 2019.
- [48] System-wide Security; <https://www.ericsson.com/en/security/a-guide-to-5g-network-security>, Last visit: November, 2019.
- [49] Siddiqi, M. A., Khoso, M. & Aziz, A.: Security Issues in 5G; Conference Paper (ICCMS 2017), February, 2017, https://www.researchgate.net/publication/314176863_Security_Issues_in_5G_Network
- [50] What is the difference between network slicing and Quality of Service?; <http://bit.ly/37yIwSE>, Last visit: November, 2019.
- [51] 5G security - enabling a trustworthy 5G system; <http://bit.ly/2Z0Pb8k>, Last visit: November, 2019.
- [52] What is Trust?; <https://www.cisco.com/c/en/us/solutions/service-provider/5g-what-is-trust.html>, Last visit: November, 2019.
- [53] Is 5G technology bad for our health?; <https://www.medicalnewstoday.com/articles/326141.php#1>, Last visit: November, 2019.
- [54] Risiken und Gefahren von 5G; <http://bit.ly/39DAZUc>, Last visit: November, 2019.
- [55] The 5G Health Hazard That Isn't; <https://www.nytimes.com/2019/07/16/science/5g-cellphones-wireless-cancer.html>, Last visit: November, 2019.
- [56] What is Security as a Service? A Definition of SECaaS, Benefits, Examples, and More; <https://digitalguardian.com/blog/what-security-service-definition-secaas-benefits-examples-and-more>, Last visit: November, 2019.
- [57] IoT Platform; <https://www.ericsson.com/en/internet-of-things/iot-platform>, Last visit: November, 2019.
- [58] About Industrial 5g; <https://assets.new.siemens.com/siemens/assets/api/uuid:622a2d3a-0885-472f-a7ba-df2af2440748/background-what-is-industrial-5g.pdf>, Last visit: January, 2020.
- [59] How 5G will change industry; <https://assets.new.siemens.com/siemens/assets/api/uuid:e48c77f2-2d4d-4425-80c8-4a8a588280f6/background-how-5g-will-change-industry.pdf>, Last visit: January, 2020.
- [60] Major factor behind US-China Trade war; <https://cnb.cx/39CreWs>, Last visit: January, 2020.
- [61] How 5G tied to trade war; <https://www.cnet.com/news/how-5g-got-tied-up-in-a-trade-war-between-trump-and-china/>, Last visit: January, 2020.
- [62] US-China discussion about trade deal; <https://www.cnb.com/2019/11/17/china-and-us-had-constructive-discussions-about-phase-one-trade-deal.html>, Last visit: January, 2020.
- [63] Trump threatens China over trade deal; <https://cnb.cx/2STXL4C>, Last visit: January 2020.

- [64] China reaction to Trump's warning; <https://www.scmp.com/economy/global-economy/article/3038547/china-remains-silent-trump-repeats-threat-hike-tariffs-if>, Last visit: January, 2020.
- [65] Huawei ban in UK; <https://www.theguardian.com/technology/2019/apr/24/may-to-ban-huawei-from-supplying-core-parts-of-uk-5g-network>, Last visit: January, 2020.
- [66] Huawei deploying 5G in South-east Asia; <http://bit.ly/2Mwd88K>, Last visit: January, 2020.
- [67] Germany introducing 5G with Huawei amid US pressure; <https://cnb.cx/2ukNDrh>, Last visit: January, 2020.
- [68] Situation of 5G in Switzerland; <http://bit.ly/36o0Egb>, Last visit: January, 2020.
- [69] Survey (study) on the future 5G dominators; <https://www.reuters.com/article/us-telecoms-5g/usa-china-japan-and-korea-to-dominate-5g-study-idUSKBN1XH0RC>, Last visit: January, 2020.

Chapter 10

The Economics of Multi-Access Edge Computing

Haishan Fei, Andreas Knecht, Dmytro Polyansky

Multi-Access Edge Computing (MEC) is a recent buzzword in the technology industry. Our paper endeavours to explain the economic prospects as well as challenges of this network infrastructure with a particular emphasis on stakerholders, the regulatory environment as well as the security and technological challenges that surround it. Through examples and case study analysis we show why we think of MEC as more of a long term challenge rather than a short term reality in the technology industry today.

Contents

10.1 Introduction	69
10.2 Definitions	69
10.3 Enablers	70
10.3.1 Virtualization	70
10.3.2 Network Function Virtualization	70
10.3.3 Software Defined Networking	70
10.3.4 Radio Communication Standards	70
10.3.5 Virtual Machine Migration Algorithms	71
10.4 Use Cases	71
10.4.1 Vehicular Networking	72
10.4.2 Infotainment	73
10.4.3 Smart City	74
10.4.4 Augmented Reality	74
10.4.5 Tactile Internet	75
10.4.6 Internet of Things	75
10.4.7 Content Delivery Caching	75
10.4.8 Hyper-Targeted Advertising	75
10.5 Stakeholders	75
10.5.1 Stakeholders of MEC in General	76
10.5.2 Stakeholders of Vehicular Networking	78
10.6 Concerns of Policymakers, Users and Stakeholders (Economics and Policies)	78
10.7 Security and Regulations	79
10.7.1 Physical Security	80
10.7.2 Privacy	80
10.7.3 Reliance on Legacy Systems	80
10.7.4 Cost Optimization	81
10.8 Challenges and the Market Situation	81
10.8.1 Competition	83
10.8.2 The Market Behind V2X Technology	83
10.8.3 The Market for Infotainment	83
10.8.4 Challenges in Vehicular Networking	84
10.8.5 Challenges in Autonomous Driving	84
10.9 Summary and Conclusions	85

10.1 Introduction

The past years saw the advent of cloud computing as well as an ever growing number of devices connected to the Internet [6]. The ubiquitous deployment of computing devices and smart sensors at the very edge of the network results in vast amounts of data that need to be aggregated, processed and analyzed in order to benefit from those devices. Following the current cloud computing paradigm, this work is predominantly performed in centralized datacenters, where the computing power for the tasks is provisioned. There are applications, however, where the latency from sending data to a centralized cloud and back to the consumer at the edge of the network is not acceptable. For example, one might expect to see autonomous driving cars. If two such cars get into an endangering situation, communication via a centralized cloud is not feasible. The two vehicles need to communicate in a Vehicle-to-Vehicle (V2V) situation (10.4.1), ensuring the lowest possible latency to avoid consequences. Multi-Access Edge Computing (MEC), also previously known as Mobile Edge Computing, enables this kind of communication by provisioning a cloud computing instance at the edge of the network, for example at the nearest cellular base station.

In this paper, the current development of MEC is analyzed. In Section 10.2, MEC is defined and differentiated from other paradigms such as Cloud Computing, Internet of Things and lastly Fog Computing. Section 10.3 presents the underlying technological enablers that are required for MEC. In Section 10.4, various use cases are illustrated to show the vast array of applications, where MEC can be deployed. The Stakeholders from these use cases are further investigated in Section 10.5. Section 10.6 discusses concerns regarding MEC and Section 10.7 analyzes the regulatory perspective. The challenges and market situation of MEC are discussed in Section 10.8. Finally, the paper is concluded in Section 10.9.

10.2 Definitions

MEC can in the simplest terms be described as “...ability to run IT based servers at the network edge, applying the concepts of cloud computing” [5]. The location of such servers is not included in the definition, as not to restrict the possibilities of MEC. One typical application might be directly located at the radio access network elements [5] that are part of the telecommunication infrastructure. The founders of the MEC industry initiative [5] characterize MEC by five keywords:

- **On-Premises:** The edge is local, and can be decoupled and run in an isolated network. This is relevant for secure systems.
- **Proximity:** The local proximity between the edge and end devices allows rapid processing of data generated by the end device. The edge might even be able to have direct access to the end device.
- **Lower latency:** As consequence of proximity, latency is considerably reduced.
- **Location awareness:** With the edge being part of a local network, it is able to gather location data for each connected device and thus enable location-aware applications, such as indoor-pathfinding.
- **Network context information:** As the edge is tightly connected to the local network station, network statistics can be used to offer a richer mobile broadband experience to end-users.

Other terms are sometimes used in literature. The term *Fog Computing* was mainly coined by Cisco and describes efforts to distribute cloud computing resources across the network, predominantly to the network edge. The term *Follow-Me-Cloud* describes a cloud computing platform that follows a mobile user to ensure consistent low-latency access (see also Section 10.3.5). A *Cloudlet* is a cloud datacenter that is located at the network edge and, therefore, an equivalent concept to a MEC host or MEC datacenter.

10.3 Enablers

A number of technologies, some of which have been around for a long time, some of which are recent developments and some of which are not yet fully deployed, are required for MEC and its use cases to achieve their full potential. We present those that are frequently mentioned in literature about MEC.

10.3.1 Virtualization

Virtualization is often used to run an entire operating system in a container that is sandboxed from the underlying hypervisor platform. These days, virtualization is also used to sandbox distinct programs or software components without running a full operating system in the container. Virtualization is an enabling technology of MEC, because it helps abstract away heterogeneity of hardware of MEC platforms and performs an important security function by sandboxing the applications running on the MEC platform.

10.3.2 Network Function Virtualization

Network Function Virtualization (NFV) is the practice of enabling network functions, which are traditionally implemented in hardware, to run in software. This allows for a much greater amount of flexibility in defining network functions for a given use case or adapting and optimizing network functions after deployment of the network function device.

10.3.3 Software Defined Networking

NFV also enables Software Defined Networking (SDN), which aims to separate network forwarding functions and network control functions. Thus, the network is split into a data plane, a control plane and a management plane [21]. These technologies also give rise to network slicing, which is the division of a single physical network into multiple virtual slices, each optimized for a specific function. These technologies could enable MEC for many new use cases on cellular hardware without compromising throughput and latency for existing use cases that make use of the hardware.

10.3.4 Radio Communication Standards

In literature about MEC, radio communication standards are frequently mentioned as enabling technologies. Most prominent is the 5G cellular network. The Long Term Evolution (LTE) and LTE-Advanced (LTE-A) cellular network standards are also often mentioned. Additionally, there exists a standard specifically for non-cellular close-range communication between vehicles and road side units (RSUs): IEEE 802.11p. The general concept of adding wireless connectivity to vehicular environments is known as wireless access in vehicular environments (WAVE). 802.11p is a standard similar to those known as WiFi (i.e. 802.11a/b/g/ac). Because wide coverage of 802.11p would entail large investments into

802.11p infrastructure at the side of all roads, it only offers short and intermittent connectivity when available [3]. Most automotive connectivity solutions involving 802.11p are likely hybrid solutions that also involve other radio communications standards and that forward data via 802.11p links between vehicles until a vehicle with cellular connectivity forwards the data to the Internet [8].

3G is the 3rd generation of cellular networks. It is rarely mentioned in MEC literature because the data rates are not high enough and latency not low enough to enable many use cases that MEC promises. LTE is the 4th generation cellular network standard and offers higher data rates than 3G and is sometimes mentioned in conjunction with MEC. However with the upcoming advent of 5G, the 5th generation cellular network standard, which promises even higher data rates and much lower latency than LTE, 5G is most frequently cited as an enabler of MEC. Concretely, the reduced latency of 5G in conjunction with the lower latency of MEC compared to cloud computing is frequently cited as an important requirement for some MEC use cases.

Even though V2V communication via 3G and LTE is certainly possible, these standards have the drawback that direct connectivity between vehicles is not possible without involving, for example, an eNodeB. LTE-A is an increment of LTE and provides the function *LTE Direct* [15], which enables vehicles to communicate with each other directly. This enables a much-needed reduction of network load for upcoming smart vehicle applications. We didn't find any definite evidence that 5G will provide a similar function like *LTE Direct* in its first release, but it seems certain that such an important feature would be implemented in 5G at some point.

Other radio communication standards such as 802.11a/b/g/ac (WiFi), ZigBee and Bluetooth are used for IoT applications, such as building automation and sensor networks, as well as for mobile phone based applications. These standards are not used in automotive connectivity, however.

10.3.5 Virtual Machine Migration Algorithms

Since most use cases of MEC deal with end users that are mobile (i.e. that are moving around), such as pedestrians connected via their smartphone or cars connected via a vehicle-to-anything (V2X) connection, it is not enough to provide a cloud application on a host located at the network edge. To ensure the lowest latency and highest throughput, the cloud application must follow the users geographically or topologically while they are moving around. A variety of algorithms are proposed that optimize the migration of virtual machines (VMs) running MEC applications (e.g. [2, 8, 12, 14]). Optimization criteria are minimizing the number of migrations, which can put a heavy load on the backhaul network, while minimizing access latency for the MEC application and maximizing throughput.

10.4 Use Cases

This section presents a non-exhaustive collection of use cases that can profit from MEC. The power of the use cases to excite end users is a critical requirement for the adoption of MEC. If MEC cannot enable use cases that end users really need, large investments into MEC might be premature. Since we see vehicular networking as the most promising of the use cases of MEC, we discuss that use case in further detail. Additionally, we introduce the infotainment, smart city, augmented reality (AR), tactile internet, internet of things, content delivery caching and hyper-targeted advertising use cases.

10.4.1 Vehicular Networking

Vehicular networking is considered as key technologies required a series of application related to traffic, individual cars, passengers and pedestrians [9].

10.4.1.1 Introduction

As we can see, this network contains the connection among vehicles and infrastructure. In vehicle side, we have V2V, vehicle-to-pedestrian(V2P), vehicle-to-infrastructure(V2I) and vehicle-to-network(V2N). In infrastructure side, the edge computing (EC) focuses on the overall safety, the road condition and comprehensive management under traffic circumstances.

Every EC device serves as an individual data station to collect and process the data. With the communication among vehicles, and infrastructure, a highly informed vehicular network is formed to enhance the safety level of vehicular network and improve traffic efficiency and management applications. It acts as an ecosystem instead of a system with a centralized monitor because every device can give out their own decisions from its own situation.

The aimed capacities of EC devices can be summarized as: radio communication capabilities, network communication capabilities, vehicle absolute positioning capabilities, vehicle communication security capabilities and other vehicle capabilities [9].

Compared to the traditional cloud-based data collecting and processing method, it would be better to collaborate both cloud computing and edging computing together to serve the functions of vehicular networking. The cloud computing means to process all the raw data in the Cloud, which is a highly centralized data center. The responsibility of data sources is to send all the data to the cloud and download the processed data from the cloud, as required the cloud center to have a powerful data processing ability.

Cloud computing has numerous benefits, while there exist some problems if cloud computing is the only contributor to the vehicular networking. EC can be the solution to the following problems because of its different characteristics. It features individual data processing, which means that sources help to process the data at the edge without directly sending it the centralized cloud. The problems and solutions are as followed:

The first problem is the large data quantity. Every day the quantity of the data produced in a smart city is incredibly huge. It is nearly impossible for several data centers to collect all the data and process by themselves. EC enables all the vehicles to process the raw data near the edge of the data source instead of directly sending them to the cloud and downloading all the processed data from the center.

The second problem is the data latency. Data latency comes with the problem of complicated data transmission way. Low data latency is crucial to the vehicular networking and EC applies a simplified network structure. Autonomous driving vehicles controlled by EC technology can have very short response time.

Another strength of EC is Location Awareness. EC collects and processes data on the location of the edge, which adds to its location awareness advantage. This characteristic facilitates the following decision-making and responding process. Due to all the mentioned, it is highly advantageous to apply EC into the development of vehicular networking and smart city traffic system.

10.4.1.2 V2X Technology and Autonomous Driving

A smart city requires the connection among autonomous vehicles, the environment and all the facilities on the road. Here are some of the advanced technologies that are paving the way to the future of vehicular network: V2X, V2N, V2I, V2V. The design aim of autonomous driving with V2X technology is to ensure the safety level of road traffic and

autonomous driving. What is most challenging is the autonomous vehicles have to collect and process massive data in real time (like 2 GB/s) with extremely low latency. For example, when an autonomous car is going to enter an intersection in one minute, it has to predict potential dangers up to a few seconds to ensure the safety of the vehicles. In other words, the processing speed of EC could partly determine the safety of autonomous driving.

In general, EC device of each car collects and processes its own data instead of directly sending it to the cloud. With EC technology, these devices can communicate with other cars immediately with the processed real-time data. While the system of autonomous driving is really complex. The system tightly integrates many technologies, including cooperative sensing, location awareness, decision making, data storage and so on. [11].

V2V implies the direct connection between two vehicles with the help of EC. Current automation systems in traffic mostly rely on ultrasonic, radar and camera. These technologies act as sensors to see and analyze the surroundings for the driving cars. While the problems are the limited view and sensory ability of these sensors. As a result, sometimes they cannot spot other cars immediately or even ignore some hidden objects, which posts a threat in road safety. Vehicles equipped with V2V technology can connect with each other and communicate in real time. This technology enhances the level of road condition predictability and safety. For example, every year a significant percentage of accidents in all countries happened at intersection because of the ignorance of other cars or traffic rules. With the technology of V2V, the EC device of each cars can communicate with other vehicles, provide information and assistance to the driver immediately to avoid such accidents.

With current technical level of data collecting and processing, V2V can be extended to V2X, vehicles to everything. This idea implies that vehicles with built-in electronics can communicate with all the things in the surrounding places, including V2V, V2P, V2I and V2N. All the vehicles, pedestrian, infrastructure and even network are put into a huge zone. Besides the advantage of better ensuring road safety, active applications of V2V, V2P, V2I and V2N can also enhance the traffic efficiency.

With the introduction of vehicular network and V2V, V2P, V2I and V2N technology, Autonomous driving technology has already been the innovative tendency on the road. Previous autonomous vehicles apply a variety of sensors to perceive their surrounding with fog computing, but no EC. With the application of V2V, V2P, V2I and V2N technology, it becomes possible to create a new kind of autonomous driving, which can be called Cooperative Autonomous Driving. Cooperative autonomous driving is divided into two categories: one is cooperative sensing and the other is cooperative decision [11]. Cooperative sensing is about sharing sensing information between V2V and V2I. Compared to the traditional individual sensors, this cooperative sensing can extend the limited range of traditional sensors and provide more information to both of the connected cars. The cooperative decision allows cars in a range with connect ties to make decisions together, which decreases the probabilities of accidents and collision among the nearby cars.

10.4.2 Infotainment

In general, infotainment is a type of media which can provide information and entertainment. With the development of city life, researchers focus more and more on the combination of information and entertainment. As a result, a variety of infotainment with different functions and application come out. One of the most popular application of infotainment is in-vehicle infotainment.

In current society, entertainment has become an essential part of everyone's life and automobiles have become the places people stay for several hours every day. Therefore, it becomes a tendency to develop a practical and new media as in-vehicle infotainment.

In-vehicle Infotainment is a collective system for entertainment and information, which can be enjoyed in a car. In-vehicle infotainment devices or technologies can provide entertainment services such as movies, games, TV, navigation and various services that are linked with mobile devices.

The automobile industry develops in a high pace and more and more communications-related requirements are added to the production of automobiles. Besides the basic requirement of traffic safety and traffic efficiency, infotainment has become a new focus of many automobile manufacturers and software companies. Automobile manufacturers like Mercedes-Benz and BMW and also some software companies such as Google, Microsoft are providing in-vehicle infotainment systems for developing these services.

Developing the infotainment system on the basis of EC technology could bring several business benefits. In the terms of the production of in-vehicle infotainment, a new value chain is expected. Automobile manufacturers cooperate with software companies to develop much more user-friendly infotainment system, which may form an ecosystem with both innovation and business value. Another benefit is about business transformation: Since new industry has been added to the market, cases are that some software companies or automotive parts companies might transform into Infotainment companies. Enterprise merger can also be applied between automobile manufacturers and software companies.

10.4.3 Smart City

The smart city use case is concerned with increasing efficiency of running a city by using IoT devices distributed across it. Examples include garbage bins that autonomously report when they need emptying and street lights that report when the bulb is defective. In this way, human and other resources can be deployed more efficiently and degradation of infrastructure can be monitored. An often-mentioned use case example where MEC can prove to be a valuable enabler is a mobile phone-based search for a lost child [19]. When a lost child is reported, the recently captured images and videos of all cell phones within a certain radius of the lost child are scanned for a visual match of the lost child in the image frame. But distributing the images and videos to a central server would entail uploading large amounts of data to a single server, risking network congestion due to limited bandwidth, as well as privacy concerns. With MEC, the image of the lost child is distributed instead to all cell phones in a given area, which autonomously search their recently captured photos and videos for a visual match of the lost child in the frames of the captured images and report only matches back to the backend server.

10.4.4 Augmented Reality

AR is a growing market that aims to augment the users' perception of the world around them with useful and/or interesting additional digital content that is overlaid over the users' field of vision (e.g. using AR glasses or by holding their phone camera in front of them). Examples include providing the user with timetables for public transport within their field of vision of the public transport station. Since the bandwidth requirements for AR applications are high, MEC could provide important offloading opportunities so that the AR data doesn't have to be fetched all the way from the cloud [17, 16]. Additionally, since the image processing places a high demand on the device's CPU, expensive calculations could be offloaded to a MEC server, as long as the latency is low enough and the bandwidth high enough. Some examples of AR use cases intersect with assisted driving, such as displaying additional information on a car's head-up display (HUD). A simple example would be to display a warning when a vehicle ahead (this is not limited to the vehicle directly in front, there might be multiple vehicles between the reporting and the receiving vehicle) reports that it's braking hard. A more complex use case could

be making vehicles “transparent”. In that use case a vehicle would stream a live feed of recordings from cameras attached to the vehicle’s front to the vehicles behind it. The following vehicles could then overlay the vehicle in front with its video feed, making it appear transparent and allowing the users to see obstacles “through” the vehicle.

10.4.5 Tactile Internet

The concept of tactile internet is similar to AR, but focuses more on the goal of minimizing latency when a user uses the Internet on their mobile device. The ultra-low latency promised by 5G technology can help achieve the required quick response times on the end users’ devices [18] and MEC can further help bring latency times down.

10.4.6 Internet of Things

IoT applications cover many use cases, including smart city, smart home and healthcare. Examples of smart home applications include a fridge that reports the stock of foods or may even automatically generate a shopping list and control of home appliances on the go, such as starting the robot vacuum cleaner before getting home. An example of a healthcare IoT application is the monitoring and autonomous reporting of environmental conditions when medicinal products or organs are transported. IoT devices that are not attached to the power grid usually have stringent energy consumption constraints. Offloading computations into the cloud to save energy on the IoT device [19, 1, 4, 22]. Depending on the amount and the nature of the data, offloading into the cloud may present concerns about the data volume, privacy and latency. Offloading only to the nearest device that is connected to the power grid could be a MEC-based solution to this problem.

10.4.7 Content Delivery Caching

With the growing market of streaming content such as video and music, transmission of that data from a few central cloud servers to many becomes a bandwidth bottleneck. Content delivery caching [21, 4] is a technique employed to cache the content to be delivered at a geographically and topologically closer location to the end user. While large content delivery networks might already employ this technique using their own servers in many geographically distinct locations, MEC can offer standardized platforms to host content delivery caching applications virtually everywhere at the network edge and help minimize bandwidth bottlenecks with content streaming.

10.4.8 Hyper-Targeted Advertising

Hyper-targeted advertising could provide relevant advertising for local businesses to all users in close vicinity [16]. Instead of relying on the Global Positioning System (GPS) location of users, a MEC-based approach could run the advertising application on a MEC device and target all users connected to that edge device. The added value for the end users may be small, but the promise of hyper-targeted advertising could incentivize more companies to invest into the deployment of MEC with the hope of getting a return on their investment through hyper-targeted advertising.

10.5 Stakeholders

In this section stakeholders of MEC are discussed. These include drivers behind the technology, such as researchers, standardization organizations and technology companies, as

well as, for example, governments putting constraints on MEC, such as privacy regulations. First, stakeholders of MEC in general are discussed and then stakeholders for the specific use case of vehicular networking.

10.5.1 Stakeholders of MEC in General

We identified the following stakeholders:

Telecommunications companies These are the companies that provide telecommunications connectivity to end users, including voice calls, short message service (SMS) and cellular Internet access. They are also called mobile network operators (MNOs).

The telecommunication infrastructure has gone through many stages since its inception, including telegraph and circuit switched telephone landlines, analog cellular phones and multiple generations of digital cellular phones. While 2G, 2.5G, 3G and 3.5G technologies still used circuit switching for voice calls, 4G uses packet switching for all data. Additionally, starting with 2G, telecommunications companies started offering Internet access to their users, taking a role that was previously reserved to Internet service providers. With MEC, the telecommunication companies would become platform providers for application developers, offering the service of running third-party applications on their edge infrastructure.

Internet service providers can offer similar MEC infrastructure to those users that connect to the Internet via their cable link instead of via the cellular network.

Automobile industry The automobile industry is making large investments into research and development of assisted and autonomous driving technologies. These technologies benefit greatly from the decreased latency, higher connectivity and increased V2V bandwidth offered by MEC technologies. Thus, car manufacturers have a large interest in speedy deployment of MEC and 5G.

Standardization bodies and working groups Multiple organizations have been formed with the goal of facilitating and promoting the adoption of MEC or some of its use cases. These include:

- The Third Generation Partnership Project (3GPP) works on specifying the evolving cellular standards with the goal of bringing better connectivity, higher throughput and support for new use cases to the people. Thus, their focus is much broader than MEC specifically, but since all potential MEC technologies run on top of cellular standards specified by 3GPP, coordination with them is an integral part of developing MEC.
- The European Telecommunication Standards Institute (ETSI) is concerned with work on cellular standards inside Europe. As such, it is a member of 3GPP.
- The MEC Industry Specification Group (ISG) inside ETSI is tasked with proposing standards that enable MEC technology on the existing and upcoming cellular standards.
- The Network Function Virtualization (NFV) ISG inside ETSI is tasked with proposing standards concerning NFV. Because NFV is an enabling technology for MEC the MEC ISG cooperates with the NFV ISG to align their work.
- The 5G Automotive Association is a cross-industry association to align the work of the automobile industry and telecommunications industry with the goal of achieving better connected vehicles in the future.

- The European Automotive Telecom Alliance (EATA) is a European cross-industry association with similar goals to the 5GAA. The 5GAA and EATA have signed a partnership agreement to support each other in their similar aims.
- The Open Edge Computing Initiative is an alliance of companies from multiple industries with the goal of driving MEC technology forward. Goals they name include providing attractive example edge applications for live demonstrations and running a real-world EC test center. Further, they work with all other stakeholders to drive adoption. Members include Microsoft, IBM, Intel, Nokia, T-Mobile and Vodafone. The fact that so many large companies and drivers in the technology and telecommunications industry have joined the initiative is testament to the interest in MEC by large companies from multiple industries.
- The Next Generation Mobile Networks (NGMN) Alliance is another alliance with members from the telecommunications industry, including operators, manufacturers and researchers. Hence, proposals for the direction of development of MEC can also come from this organization.
- The Institute of Electrical and Electronics Engineers (IEEE) is an association on the topic of electrical engineering. IEEE publishes multiple radio communication standards, including 802.11p.

Governments The drive behind MEC and 5G comes from private companies and organizations. Users of MEC technology, however, include public establishments, such as hospitals and schools, and cities. Furthermore, governments have a stake in these technologies, however, when it comes to the legal side. Privacy laws restrict the distribution of data on these networks (see Section 10.7). MEC, however, can help address legal restrictions regarding privacy because data is distributed less (i.e. only to the MEC node and not to a central cloud). If a city employs smart city systems they also become a stakeholder of the underlying technologies in that way.

Cellular hardware manufacturers include companies that manufacture hardware for cellular network infrastructure, such as base stations, as well as companies that manufacture cellular end devices such as mobile phones. Huawei is hugely involved in developing the 5G hardware that is deployed around the world. Further, they are involved in supplying pre-commercial or commercial hardware that are used in test beds, such as the Open Edge Computing Initiative's real-world test center. Nokia as a cell phone manufacturer is a part of the Open Edge Computing Initiative.

Cloud companies offer cloud services, such as platform-as-a-service (PaaS), infrastructure-as-a-service (IaaS) and generally anything-as-a-service (XaaS). Examples of companies offering cloud services include Google, Microsoft and Amazon. With the drawbacks of traditional cloud services for some use cases (as discussed in Section 10.1), these companies could invest in infrastructure at the network edge or extend their application portfolio to run on infrastructure at the network edge provided by telecommunication companies. The authors are not aware of any such efforts by cloud companies, except that Microsoft joined the Open Edge Computing Initiative. Cloud companies might employ their own geographically distributed infrastructure, but they seem to be awaiting for MEC to mature before they adopt this technology.

End users are private individuals who use technology products that might make use of MEC. On one hand end users are the consumers or beneficiaries of all the aforementioned technologies, including cellular internet, IoT device services, smart city applications, intelligent automobile safety systems, etc. Thus, the advancement of

MEC technology benefits them indirectly (e.g., through improved smart city systems in their home city) and directly (e.g., from higher mobile Internet bandwidth). On the other hand, end users could also rent out some of their unused, available computation resources to mobile and IoT devices and, thus, join the collection of nodes running part of the MEC platform [22].

IoT Solution Providers are hardware and application developers that develop IoT solutions, such as those deployed in smart cities. They are interested in wireless communication and computation offloading technologies that help them save power on their devices. During the development stage they might be frequent users of test beds such as the Open Edge Computing Initiative's real-world test center. They are not directly concerned with development of MEC technology, however, because their focus lies on their own IoT devices.

Technology companies such as IBM, Microsoft and Intel seem to see potential in MEC technology and by joining the Open Edge Computing Initiative they joined a driving organization behind the technology. For Intel the goal is to supply hardware that is deployed in the hardware running these new platforms. For Microsoft and IBM it is more about having a stake in all sorts of technology that might have a high relevance in the future.

Application developers are third-party application developers who might want to develop applications running on a MEC platform. Thus, without any specific knowledge of the technical details of the cellular infrastructure hardware and software, a good amount of abstraction must be built into the MEC application programming interfaces (APIs) that application developers use to build software for MEC platforms.

10.5.2 Stakeholders of Vehicular Networking

Currently there is almost no vehicular network in a wide practical range, while it is necessary to get an insight to the stakeholder and market behind it. The market can be complex. To be clear, we can split the market into three parts: Vehicles, Road Administrators and Service Providers. First of all, in order to get everything connected, the hardware of EC devices should be produced and applied to all the vehicles and infrastructure in the vehicular network. As to the road administrators, they act more like a middleware to organize a range of vehicles with the technology of V2V and V2I. Besides of the technical problems, the privacy problem of individual vehicles should be protected. A vehicular network is rather complex with a great number of cloud subsystem and edge subsystem. The market of service providers behind each subsystem need to be well-designed and organized.

10.6 Concerns of Policymakers, Users and Stakeholders (Economics and Policies)

We have already seen that by bringing storage related capabilities and the brunt of computational power from the core network to the edge network, providers are able to reduce latency and back-haul bandwidth requirements.

However, there are many challenges related to this relocation. Apart from the regulatory challenges (see Section 10.7), the amount of responsibility that each edge has increases dramatically. This is also compounded if the edge is part of a service which relies on essentially real-time information being provided at a high standard of service where even

low fault tolerances are not acceptable. Furthermore these edges may also have to provide several services at once. As noted before, while EC has the potential to decrease bandwidth requirements while also simultaneously increasing the speed at which tasks are performed, it also has a double edged sword when it comes to responsibility, regulation and management.

On the technical side, there are many challenges ranging from security concerns, to viability in an environment where the user is constantly mobile. For example, if a user in a given MEC environment is mobile, how many times will they face disconnections due to being out of range? With so many types of devices out there, all these edges need to have an optimal way to perform handoff. Furthermore, a mobile user presents challenges for a network. Will it be viable to track how many times a certain user switches to a different edge? Will it be acceptable to charge someone based on their mobility within a network? These are all important economical questions that need answering.

MEC creates a plethora of opportunities for not only mobile service providers to provide there services, but also for third party organizations to take advantage of them. For example, the use case of cars being able to communicate with many road-side sensor devices increases the safety further of the vehicle operation. No longer would the car need to account for round-trip time of information from a central server to the actual car – instead the car can now communicate in what is essentially real time with the sensors in its proximity. Furthermore, the sensors are also able to propagate this information to other nearby sensors (without the use of a central server) and cars which need to know the information as vehicles in a different area of the city do not need to know about an isolated event. Not only does this increase safety on the roads, but it massively alleviates both the bandwidth requirements as well as the resource requirements required for such a system. Vehicular networks can also be extended to increase not only V2I communication, but also V2V. Taking advantage of parked cars as an information source while in motion for certain information is something that can be leveraged in a MEC environment. However, it is important to distinguish the two types of infrastructure from as V2V should not be used for mission-critical and safety tasks and should mostly be used for information for the vehicles occupants that is not time sensitive.

10.7 Security and Regulations

Security is one of the most important areas of discussion in the MEC scene at the moment. Various stakeholders, regulators and developers all need to consider the security behind an application which utilizes MEC. Applications which were discussed earlier in the paper (such as Vehicular systems and content caching) is not immune to various threats by a relatively sophisticated adversary.

Defending against:

1. Man in the Middle Attacks
2. Falsified Information and Logs
3. Loss of Policy Enforcement
4. Loss of Data

MEC networks are quite prone to certain attacks such as Man in the Middle Attacks. Taking example a parked car who may be transmitting information about a road-accident ahead which causes the unsuspecting driver to take an alternative detour route. Although this may not be a big deal from the safety perspective, things like this have an effective to cause a vehicular network to be compromised and deemed untrustworthy. It may be

tempting to initially create many Value added Services for a vehicle network, but unless they are properly secured, they are unlikely to be approved by regulators and even more unlikely to be trusted by the majority of the population as they hear about possible malicious use of such technology against them; falsified Information (especially if certain data has the potential to be changed down the road) is a major problem. Even more-so, if intra-vehicle communication is enabled in some of these cases, vehicles should be preventing from transmitting false information (for example, information that a car is braking ahead when it is not should not happen and the vice-versa of this account is even more dangerous).

It is always important to consider security at the forefront of EC development as there are many layers to secure. Without adequate security, verification and privacy mechanisms, any benefits that developments in EC bring to the world are quickly voided if there is no confidence in the security of the technology from the user and as well as the regulator side.

10.7.1 Physical Security

An apparent but otherwise often so often addressed issue is that the security of MEC networks is also concerned with the physical aspect in a sense. This is because many there are many nodes in the network and guarding it the same way say a central server is guarded is not always feasible. However, on the other hand, these networks also offer certain advantages in securing data because of the shorter distance between a user and such a node. Furthermore, because of the physical layout of the network (and its accessibility from various geographic locations), successful attacks can happen on any one of the nodes, which compromise the network in that particular geographic region – or even worse – propagate to another section of the network.

10.7.2 Privacy

When certain data is said to be processed by other edge nodes (say in various geographic regions), privacy regulations become a major concern. This happens when data is processed by entities that are not controlled by a user. Furthermore, these considerations are even more complicated when different regulations exist for such regions (e.g. different country or even different state). Requirements for fault tolerances, forensic evidence storage all vary and all require different levels of service even though a network may only be capable of a certain level of operation.

10.7.3 Reliance on Legacy Systems

Perhaps one of the most important challenges that many MEC networks currently or in the near future will face is the heavy reliance on legacy systems. For example, the most popular localization methods today (Satellite GPS) are not always so compatible with the real-time needs and low fault tolerances that many MEC networks will be designed to have. For example, GPS signal is not always available (e.g. tunnels). Secondly, GPS signals can easily be spoofed by an adversary. Furthermore, attenuation in GPS is also a problem in many circumstances. Innocent things such as atmospheric conditions, solar activities, natural impediments and presence of electromagnetic fields can have dire consequences for certain MEC networks that make them not only unfeasible to develop, but also dangerous for humans. On the malicious side of things, spoofing a GPS signal is actually not difficult for even a less-sophisticated adversary these days – especially if some of the receivers are meant to work in low-power/low overhead environments where many of the popular anti-spoofing methods are not physically possible to integrate.

10.7.4 Cost Optimization

The efforts required to deploy such a network are extraordinarily large. The initial up-front investments need to be carefully managed and more importantly, with large areas to cover, need to be optimized from both a technical and financial side. Furthermore, the mounting security and regulatory challenges that edge networks will face and being under constant scrutiny for privacy and security risks pose a massive barrier to entry for many innovators due to being cost prohibitive to accomplish this and satisfy all the requirements.

10.8 Challenges and the Market Situation

MEC is faced with several challenges, including technical challenges and the magnitude of required initial investment. Additionally, the added value for end users needs to be large enough to warrant large investments into the required infrastructure. If the only advantages over traditional cloud computing are somewhat reduced latency and improved throughput, the incentive might not be big enough. Furthermore, we observe that MEC has seen very little real-world adaptation since the concept was introduced. However, there is evidence of a large interest by some companies, including the members of the Open Edge Computing Initiative and Amazon who is buying edge-based startups.

Swisscom reports 11184 LTE cell towers in Switzerland [20]. Even when assuming that 100% of these cell towers would have to be fitted with extra hardware to enable MEC, the cost of the pure hardware deployment should be below 500.000.000 USD (this would amount to 44.707 USD per MEC server, which seems reasonable, assuming the MEC servers don't consist of very special hardware). The cloud computing market is estimated at roughly 200.000.000.000 USD in 2020 [10]. Assuming a fair share of the worldwide cloud market and approximating the Swiss population fraction to $\frac{1}{1000}$ of the world population, even with a vastly unrealistic 100% profit margin, the hardware investment would take 2,5 years to pay for itself. And if MEC offered by Swisscom doesn't fully replace the cloud market, the income will be much lower. Furthermore, the large uncertainties around the concrete architecture of MEC, the lack of knowledge of who would pay for MEC-enabled services and the fact that deployment of the 5G cellular network – which is also a very large investment with uncertain promises of returns [7] – is just starting, serves as a barrier for investment into MEC.

We think that the largest drive to advance MEC will come from the automotive industry and from company alliances that include large technology companies and telecommunication providers. In the automotive industry, automobile manufacturers push for assisted driving systems that make automobile use safer and more convenient for their users, public media frequently report on advances in automated driving systems and many high-end cars already have many advanced safety systems included. Therefore, there is a strong drive to implement cellular systems and Internet infrastructure that best enable those systems. While car manufacturers do not work on MEC technologies, associations, such as the 5GAA, EATA work on driving the automobile industry interests within the telecommunication and Internet industries.

We believe that some augmented and automated driving applications can already be implemented based on hybrid systems of 3G, LTE and 802.11p in conjunction with MEC technologies and smart VM migration algorithms (see Section 10.3.5) or even using traditional cloud applications. Thus, only technologies that are already deployed are used. One question that remains is who responsible for setting up RSUs? Potential parties include the owners of the roads, such as cities, counties and states. We don't see a complete coverage with RSUs as likely because the required investment into infrastructure large. But RSU-based traffic safety applications could be very valuable and cost-efficient

in specific locations, for example, around traffic lights and intersections for smart intersection management. With LTE-A and 5G, the power of these applications will greatly increase. Not only are the latency much lower and the throughput much higher, if the cellular standards provide device-to-device capabilities, such as *LTE Direct* (see Section 10.3.4), inter-vehicle communication becomes much more reliable and latency is greatly reduced. Together with improved location-awareness brought by these technologies, much more advanced automotive safety applications can be deployed.

In addition to automotive safety applications, V2X communication can also enable traffic efficiency applications, such as smart traffic guidance or smart intersection management. While these do not directly improve the safety of automobile use, they can benefit all road users and the environment by increasing road use efficiency and minimizing traffic jams. Even though MEC is a very novel realm in the field of computer science today, there are many investors that are already seeking to secure themselves a place as a new entrant in the field. Although individual investors may find it difficult to break into the field of EC, large investment firms are already partnering with certain companies to create a partnership once certain technologies progress in the coming months/years – namely 5G. Investment vehicles like Real-Estate Investment Trusts (REITS) that are focused on data center geared locations have become popular in the recent years. There are many startups that buy up space on urban rooftops and other inconspicuous places with efforts down the road to either place their own edge nodes or rent the space out to clients who can put their own edge nodes in place for their network. There is a push for so-called “micro-data centers” so that MEC networks can do real-time computing on-board the device.

Perhaps one of the most powerful impacts of MEC, is that sectors which have traditionally seen less use and reliance on technology stand to make drastic improvements in efficiency, scalability and overall productivity as the barrier to entry will become less as technology develops to fully support such systems.

Company alliances such as the Open Edge Computing Initiative invest a lot into proof-of-concept (PoC) systems and test beds that showcase the powers of MEC. To achieve this, the expertise of companies from multiple industries are required, including manufacturers of cellular hardware, technology companies and MNOs. The showcased use cases are often IoT and smart city use cases.

Another challenge is the abstraction of the complexity of the systems hosting MEC platforms. MEC employs so many technologies, including NFV and SDN and is hosted on a complex system that has to interact with all the nodes that make up the vicinity of the cellular network base station. Orchestration of MEC applications on a MEC platform is complex. Ideally, no special knowledge should be required to develop applications that run in a virtual container on a MEC platform. Multiple models of abstracting heterogeneity between MEC platforms have been proposed. Shi et al. [19] propose a model of a computing stream, where a stream of well-defined instructions is executed on the MEC platform. Most other authors, including the whitepaper by the MEC ISG [5] agree that virtualization should be the technology to abstract away details of the MEC platform from the application developers.

We believe that the standard proposed by the MEC ISG is very powerful because it specifies a very generic architecture that is based on virtualization and Service-Oriented Architecture and includes a communication service that allows the applications to communicate with the underlying platform and with each other in a well-defined manner. The level of maturity of the specification seems high. The generic nature of the specification ensures that the architecture can be used for a wide range of use cases.

10.8.1 Competition

There are also certain barriers to investment on the edge – namely security reasons. For example, many companies are reluctant to put their content on the edge due to fears of exposing certain trade secrets and operational procedures. As a result, in its infancy stages of development, MEC will not impact services like *Amazon AWS* and other cloud providers too much. At the beginning cloud providers are not necessarily in direct threat of everything moving to the edge. Rather, the cloud will still be an integral part of operations at first – it will just be complemented by certain edge nodes and not all the data will go back to the cloud. This is even more likely as at first regulators and companies will not be able to streamline everything to the edge for privacy and security reasons. However, as the industry develops and the backbone for such technological activities (such as 5G infrastructure) improves, we may see much less reliance on traditional cloud based computing. However, for the moment it seems that traditional cloud based providers will be able to expand their business models with this new EC paradigm parallel to their current service offers. This means they will be able to develop and offer EC services alongside their traditional cloud based service models as a value added service.

At the moment, Amazon is actually “poaching” edge-based startups as it reacts to the ever more common shift to EC development techniques. Furthermore, Amazon actually considers EC as an extension of the cloud and posits that there would be no edge without the cloud. Although at present and in the near-future they may be right, as technology develops which would enable more data processing to occur at the edge instead of the cloud, Amazon may lose its appeal if it does not innovate its services correctly.

10.8.2 The Market Behind V2X Technology

The V2X technology acts as a revolutionary technology which can open up an era for the automobile industry and vehicular networking. Several countries including US, Germany, UK and China apply the V2X technology for the traffic efficiency, traffic safety and other benefits. North America is reported to be the largest V2X market until 2016 and European market follows as the second largest. The increasing concerns on the safety of pedestrians and vehicles on the road have led to the adoption of V2X technology by the European Transport Safety Council and is expected to boost the market. [13].

The key driving factor of the market is software. Software can record and organize the data content and connect vehicles together. Some software with can be use in the field of traffic regulation and overall management. Therefore, the demand and cost of the software is crucial to the future development of V2X technology application.

10.8.3 The Market for Infotainment

Since nowadays, the in-vehicle infotainment technology is mostly applied to the premium and luxury cars. Therefore, the sales of premium and luxury cars can be a key factor for the in-vehicle infotainment market. As the development of in-vehicle infotainment, the adoption in low and mid segment cars can also drive the market. The adoption decides the demand of the system since infotainment system market is an on-demand market. Security concerns can also boost the market because one of the most important functions of infotainment systems is to share the information, which may require the enabling technology of V2V, V2I, V2P and V2N technology. Another crucial factor is the public regulation from the government and road administrators. Since the construction of vehicular network has already been on the way, the infotainment technology would one day be asked to apply to all the vehicles on the road to keep the communication network complete and ensure that all the vehicles are connected and informed with the immediate

messages. Besides, as the electric automobile market are growing rapidly, its growth may prompt the growth of the infotainment industry because the infotainment systems can help the drivers locate the nearby charging stations. As the requirement of using cars to travelling long-way is growing, it may also increase the motivation for people to buy cars with infotainment systems for entertainment in cars and car navigation functions.

The in-vehicle infotainment market has several service divisions including entertainment, navigation, E-call and other services. Among all the division, the navigation division ranks top in the market size these days. Therefore, the growth demand in automobile navigation can be the key boosting factor. Also, the market for E-call also grows fast. It is obvious that in current stage the most important function of infotainment system belongs to the information and practical sides. Therefore, the quality and availability should be seriously concerned and the market of the entertainment side of infotainment systems is still blank. Innovative entertainment functions are expected.

10.8.4 Challenges in Vehicular Networking

A complete vehicular network combines both vehicular edge subsystem and Cloud subsystem. Cloud subsystem relies on the technology of cloudification, which will not be discussed here. As the vehicular edge subsystem, they have to meet several requirements such as low-latency and reliability. There are three main technical challenges to overcome: The first challenge is the processing speed, which is one of the important decisive factors for the safety of vehicular network. With the development of computer and internet, there are several ways out for the problem. One of the solutions is 5G technology. In vehicular network, there are two typical ties in V2V connection: Weak tie and Strong tie. Generally, one car always has strong ties with cars of family members or friends while most of the ties are weak ties with strangers. 5G can strengthen the weak ties. Also, better hardware architecture is essential. Several reports review and compare different kinds of hardware but currently no conclusion on which one is the best for vehicular network.

For the second challenge, we need to focus on the recovery and adjustment abilities of the vehicular network. It is not only important for the individual vehicle but also the cooperative ability to recover and deal with the emergence. The cooperative emergency response ability needs to a real-time information system and strong middleware. Different from the individual hardware, a middleware is an operation platform to bind different infrastructure and vehicles together. There are also some solutions like the application of robot operating system to facilitates the communication between V2V and V2I. Since every car is bind in the system, it is of great importance to focus on the second problem or the whole network may fail because of a single error.

The third problem is the energy problem. How to keep operating the great number of servers and devices for the whole vehicular network is always one of the main concerns of the construction.

10.8.5 Challenges in Autonomous Driving

Since the autonomous cars with EC technology need to be put into the vehicular network, to better guarantee the operation of whole system, there are several challenges. Autonomous vehicles equipped with numerous sensors and EC system are rather expensive, so from economics sides, it is of great importance to reduce the cost of the whole construction and turn it into normal prices which can be available for most of the car drivers. From technical sides, we can also discuss the challenge of cooperative decision and cooperative sensing. The challenge of cooperative decision is to handle the dynamic changing environment with a temporary coverage of V2X communications. 5G technology can strengthen the weak ties among vehicles and establish a regional stable vehicular

networking. As to the cooperative sensing part, one problem is the investment in the EC systems on public infrastructure, the other is the privacy problem. We need to identify what information can be share with vehicles and what should remain as private data.

10.9 Summary and Conclusions

MEC is a proposed concept of bringing cloud computing resources closer to the network edge, minimizing latency and improving throughput. MEC promises improved functionality for a number of use cases, including vehicular networking and autonomous driving, smart city, IoT, AR, tactile internet, improved content delivery caching and hyper-targeted advertising. A variety of technologies enable the concept of MEC, such as virtualization, NFV, SDN, new radio communication standards and advanced VM migration algorithms. The 5G cellular network is a large driver behind MEC because it also offers low-latency, high-throughput access to the cellular network. Assisted and automated driving are high-profile applications that generate a lot of media interest. Many organizations and companies, including ETSI, Microsoft and Amazon drive MEC.

The Internet in its history has gone through alternating phases of centralization and decentralization. First, the Internet was a network of computers at research institutions, computers were big and access to the Internet was limited to researchers. With the advent of the PC and the World Wide Web, everybody gained access to the Internet, thereby taking part in a phase of decentralization of the Internet. With cloud computing, computing resources have been moved to central servers, offering software as a service or offloading computation from the end devices. MEC would constitute another instance of a decentralizing development in the history of the Internet and, thus, may seem almost like a logical step. However, there are also large challenges for MEC, most prominently the large required investment. Like the 5G network, the potential for large returns on the investment is not guaranteed and it seems that companies are not making large investments into MEC.

At the same time, we are seeing some companies picking up MEC related endeavours. Economically and technologically, we see the industry still using cloud based services in the near future as the main type of service. Moreover, the large amount of regulatory oversight and security improvements that need to be sorted out before MEC can be deployed on a large scale also forces us to see MEC as more of a long term challenge rather than a short term reality.

Bibliography

- [1] AHMED, A., AND AHMED, E. A survey on mobile edge computing. In *2016 10th International Conference on Intelligent Systems and Control (ISCO)* (2016), IEEE, pp. 1–8.
- [2] AISSIOUI, A., KSENTINI, A., GUEROUI, A. M., AND TALEB, T. On enabling 5g automotive systems using follow me edge-cloud concept. *IEEE Transactions on Vehicular Technology* 67, 6 (2018), 5302–5316.
- [3] ARANITI, G., CAMPOLO, C., CONDOLUCI, M., IERA, A., AND MOLINARO, A. Lte for vehicular networking: a survey. *IEEE communications magazine* 51, 5 (2013), 148–157.
- [4] BECK, M. T., WERNER, M., FELD, S., AND SCHIMPER, S. Mobile edge computing: A taxonomy. In *Proc. of the Sixth International Conference on Advances in Future Internet* (2014), Citeseer, pp. 48–55.
- [5] ETSI, I. Mobile-edge computing–introductory white paper. *White Paper, Sept* (2014).
- [6] EVANS, D. The internet of things: How the next evolution of the internet is changing everything. *CISCO white paper 1*, 2011 (2011), 1–11.
- [7] GRIJPINK, F., MÉNARD, A., SIGURDSSON, H., AND VUCEVIC, N. The road to 5G: The inevitable growth of infrastructure cost. *McKinsey* (2018). <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-road-to-5g-the-inevitable-growth-of-infrastructure-cost>, Accessed: 2019-12-10.
- [8] HUANG, C.-M., CHIANG, M.-S., DAO, D.-T., SU, W.-L., XU, S., AND ZHOU, H. V2v data offloading for cellular network based on the software defined network (sdn) inside mobile edge computing (mec) architecture. *IEEE Access* 6 (2018), 17741–17755.
- [9] KARAGIANNIS, G., ALTINTAS, O., EKICI, E., HEIJENK, G., JARUPAN, B., LIN, K., AND WEIL, T. Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions. *IEEE communications surveys & tutorials* 13, 4 (2011), 584–616.
- [10] LIU, S. Total size of the public cloud computing market from 2008 to 2020 (in billion U.S. dollars). <https://www.statista.com/statistics/510350/worldwide-public-cloud-computing/>, 2019. Accessed: 2019-12-10.
- [11] LIU, S., LIU, L., TANG, J., YU, B., WANG, Y., AND SHI, W. Edge computing for autonomous driving: Opportunities and challenges. *Proceedings of the IEEE* 107, 8 (2019), 1697–1716.

- [12] LIU, X., ZHANG, J., ZHANG, X., AND WANG, W. Mobility-aware coded probabilistic caching scheme for mec-enabled small cell networks. *IEEE Access* 5 (2017), 17824–17833.
- [13] MARKETRESEARCH.COM. V2X Market for Automotive by Communication Type (V2C, V2D, V2G, V2P, V2V and V2I), Offering Type (Hardware and Software), Connectivity Type (DSRC and Cellular), Propulsion Type (ICE and EV), Technology Type, and Region - Global Forecast to 2025. <https://www.marketresearch.com/MarketsandMarkets-v3719/V2X-Automotive-Communication-Type-V2C-11415056/>, 2018. Accessed: 2019-11-13.
- [14] MUMTAZ, S., HUQ, K. M. S., ASHRAF, M. I., RODRIGUEZ, J., MONTEIRO, V., AND POLITIS, C. Cognitive vehicular communication for 5g. *IEEE Communications Magazine* 53, 7 (2015), 109–117.
- [15] MUMTAZ, S., HUQ, K. M. S., AND RODRIGUEZ, J. Direct mobile-to-mobile communication: Paradigm for 5G. *IEEE Wireless Communications* 21, 5 (2014), 14–23.
- [16] NOKIA AND INTEL. Increasing mobile operators value proposition with edge computing. *Technical Brief* (2013).
- [17] PATEL, M., NAUGHTON, B., CHAN, C., SPRECHER, N., ABETA, S., NEAL, A., ET AL. Mobile-edge computing introductory technical white paper. *White paper, mobile-edge computing (MEC) industry initiative* (2014), 1089–7801.
- [18] SATYANARAYANAN, M. The emergence of edge computing. *Computer* 50, 1 (2017), 30–39.
- [19] SHI, W., CAO, J., ZHANG, Q., LI, Y., AND XU, L. Edge computing: Vision and challenges. *IEEE Internet of Things Journal* 3, 5 (2016), 637–646.
- [20] SWISSCOM. Locations of LTE cell towers. <https://opendata.swisscom.com/explore/dataset/locations-of-lte-cell-towers/information/?disjunctive.powercode&sort=-id>, 2019. Accessed: 2019-12-10.
- [21] TALEB, T., SAMDANIS, K., MADA, B., FLINCK, H., DUTTA, S., AND SABELLA, D. On multi-access edge computing: A survey of the emerging 5g network edge cloud architecture and orchestration. *IEEE Communications Surveys & Tutorials* 19, 3 (2017), 1657–1681.
- [22] YI, S., LI, C., AND LI, Q. A survey of fog computing: concepts, applications and issues. In *Proceedings of the 2015 workshop on mobile big data* (2015), ACM, pp. 37–42.

Chapter 11

Botnet Economics and Business Models

Famos Tobias, Mannhart Thomas, Tham David, Waltert Gian

Botnets are large networks of infected machines used for malicious purposes and have caused tremendous economic damage since they came into existence. They have been covered exhaustively in media with attacks shutting down popular websites or even impacting the Internet of an entire country. By analyzing the Mirai botnet in our case study, we illustrate the dangers botnets pose using new technologies with missing incentives to secure them. In the first part of the paper, we give an overview of the several technical aspects, such as definitions of botnet related terms, the different types of botnet models, and possible solutions against DDoS attacks. Following this part is an economic analysis of the botnet underground market, and its economic impact on different stakeholders, which highlights the different revenue streams and incentives of botnet owners and attackers. In the last part of this paper, we take a closer look at Mirai, a recent botnet, which took the Internet by storm in 2016 with various attacks on high-profile targets. We provide a comprehensive overview of Mirai with a focus on the economic impact it caused and still causes to this day, especially in the case of the Dyn 2016 cyber attack. The botnet landscape developed into a sophisticated market for Internet criminals with a new and more prominent focus on generating revenue. In order to suppress the thriving of this underground market, new forms of countermeasures must be put in an action targeting the revenue generated using botnets. To achieve this goal, parties such as ISPs and IoT manufacturers have to contribute more to the fight against botnets.

Contents

11.1 Introduction	91
11.1.1 Motivation	91
11.1.2 Malware in numbers	91
11.1.3 Botnet Landscape	92
11.1.4 Organisation / Outline	93
11.2 Background	93
11.2.1 Botnet	93
11.2.2 DDoS Defense Systems	96
11.3 Economics	100
11.3.1 Stakeholders	100
11.3.2 Revenue Streams and Incentives for Botnet Owners	103
11.3.3 Modeling Costs	106
11.4 Case Study: Mirai	108
11.4.1 The Botnet	108
11.4.2 The Dyn attack	111
11.4.3 Other major incidents	112
11.4.4 Future	113
11.5 Summary	114

11.1 Introduction

11.1.1 Motivation

Online crime has shifted in the last 20 years. Where earlier attacks and malicious actions were performed by individual hackers, trying to break things mostly for fun, there are now organized groups built like companies that act as malicious actors in the cyber space. Cybercrime is a field of high revenue. It is estimated that the global revenue from cybercrime is at least \$1.5 trillion [47]. In context, that would be roughly equivalent to the GDP of Australia in the year 2018 [72]. The damage caused by cybercrime is estimated to have risen about 12% from 2017 to 2018 [1] and in total numbers estimated to be \$600 billion [44].

Currently, some of the most significant threats in the cybersecurity landscape are coupled to botnets, a network of computers controlled by attackers [60]. Botnets can be used for various malicious intents and actions such as Distributed Denial of Service (DDoS) attacks, generation of spam, information theft, and data exfiltration.

There is an increasing number of devices connected to the Internet every day. Furthermore, there are predictions that there will be 75 billion IoT devices in the world by the year 2025 [64]. Many of them are poorly secured, if at all. These facts lead to a severe incentive to botnet providers to target not only personal computers and smartphones but also IoT devices such as security cameras, washing machines, or doorbells. Devastating examples of the abuse of vulnerabilities in IoT devices, such as default or hardcoded credentials, can be found in mass. Large botnets like Linux.Aidra, Mirai, or Bashlite infect tens of thousands of devices and use them for malicious purposes.

The high number of devices connected to the internet, paired with the trend for more bandwidth and manufacturing devices with more CPU power and better network cards to output higher bandwidths, gives malicious actors more possibilities to abuse foreign infrastructure and devices. Additionally, there is a lack of incentives for device owners to secure their devices since there is no direct, perceptible impact from an insecure device onto them.

11.1.2 Malware in numbers

As the revenue from cybercrime is growing, the development of malware is too. Researchers estimate that there are 300'000 to 1'000'000 new viruses developed every day [44]. Currently, the fastest-growing malware thread is ransomware [44]. The FBI estimates that there are more than 4'000 ransomware attacks performed every day, with \$209 million in ransom paid in the first quarter of 2016. These facts represent a rapid growth compared to the \$24 million total in ransom payment throughout the year 2015. The deployment of malware onto personal computers is also changing and growing. The Symantec Threat Report states that the first delivery of malware has changed in recent years, from malicious URLs to malicious attachments. From the malicious attachments seen by Symantec, 48% are Office files with macros used to download malware. An astonishing rise in the cybercrime landscape is the mobile ransomware. This type of malware targets mobile devices and locks them down.[65]

Furthermore, the threats concerning IoT devices are equally rising. Symantec has seen an increase in IoT attacks in the year 2017, and the attack numbers are stabilizing in the year 2018 [65]. The most attacked devices in their IoT honeypots were routers and security cameras. Together they accounted for 90% of the IoT attacks, with routers having 75% and security cameras at 15% [65]. Telnet was the most used protocol, with 90% of attempted attacks using it in 2018, which represents up from 50% usage in the year 2017 [65].

In terms of threads, the top three accounted for over 75% of the attacks on the IoT honeypots. Namely, they were Linux.Lightaidra, Linux.Kaiten and Linux.Mirai. Thus the Mirai botnet still is an active thread with 15% of attacks originating from it.[65] Overall, one can state that phishing is still the most popular and easiest way to commit a cybercrime. Due to the low costs of phishing and the perceived low risk of getting caught phishing remains an attractive attack vector.[44]

11.1.3 Botnet Landscape

Due to the high publicity distributed denial-of-service attacks get, the public opinion is that most botnets are being used for only DDoS attacks. But there are many other usages of botnets such as web injection, URL spoofing, DNS spoofing, and data collection.[19] We will assess what criminals target with their botnets, followed by a short description of a few big botnets.

As observed by Kaspersky Lab, the most substantial part of the attacks performed was targeted at financial service providers such as banks, online stores, and other resources for financial transactions. This category made for 77.5% of the attacks observed by Kaspersky. The second-largest category is the Global Portals and Social Networks category. This includes email providers, search engines, and social network platforms [19]. The third-largest category, with 5.08%, are the providers of various products. Part of this category is hosting providers, telecom providers, and providers of other services. The full distribution of the targeted categories is shown in figure 11.1.

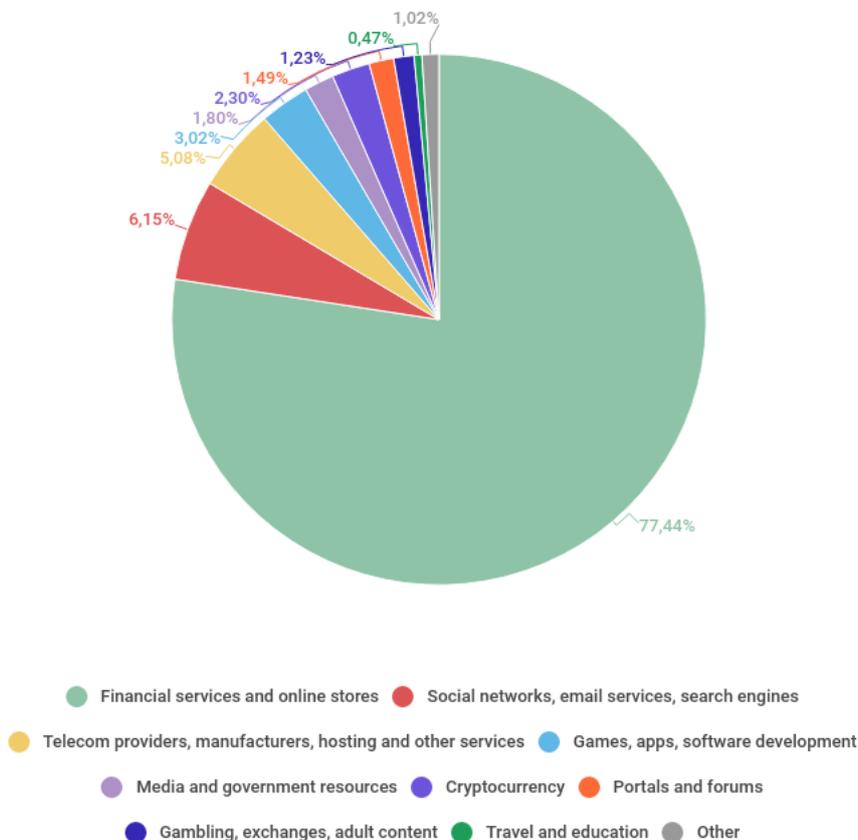


Figure 11.1: Categories of Attacks by Target [19]

In terms of the botnet landscape there are many big botnets [10]. We chose to describe three of them.

Emotet is a banking Trojan that has emerged into spam and malware as a service botnet that is globally distributed. It is distributed via corrupted attachments or links in emails sent from an already compromised device. According to Proofpoint, the Emotet botnet was responsible for half of the spam emails generated in the first quarter of 2019, all containing malicious attachments or links. Emotet's command and control server structure are very complex. The Emotet actors maintain over 100 command and control servers, which are updated frequently. This makes Emotet's C&C (command and control) structure highly resistant to single interruptions of a C&C server.[10]

Mirai is a botnet targeting IoT devices. Due to its simple usage for scaling the botnets up, it has become a popular platform for distributed denial-of-service attacks. The original Mirai code has been published and is evolving into different variations. In the first quarter of 2019, Netscout saw over 20'000 unique samples of the Mirai code [53]. To infect its bots, Mirai scans for IoT devices and tries a dictionary of default passwords and usernames. Mirai's command and control structure is quite simple. For a given botnet, there is only one command and control server.[10]

Necurs is a botnet initially used for spam and malware distribution, such as bank trojans and ransomware. Over time, it evolved into a tool capable of more malicious activities such as crypto mining, DDoS attacks, and more. Similar to Emotet, Necurs is distributed mainly through phishing emails. Known phishing attempts are Russian dating scams and pump-and-dump stock scams. Necurs targets mostly unpatched or pirated versions of windows. Thus its main targets are in developing nations.[10]

11.1.4 Organisation / Outline

In the following paper, we want to show the importance of further research about botnets and the importance of botnets itself. By providing an overview of botnets and their possible usage, we provide awareness about their dangers and highlight their influence on the economy. This is underlined by applying the stated analysis and facts onto a botnet in the case study.

11.2 Background

11.2.1 Botnet

11.2.1.1 Definition

A botnet is a network of infected end-hosts called bots (also referred to as zombies or drones) under the control of a botmaster. Botnets recruit vulnerable machines using primary methods of malware distribution. The botmaster is the manager of the botnet and controls his bots by using command and control (C&C) channels. These channels are used by the botmaster to distribute his commands to his bots. These channels can use multiple communication mechanisms like P2P or Internet Relay Chat (IRC). A vast majority are using IRC [2].

11.2.1.2 Malware

Malware stands for malicious software and describes any software that is designed to damage or exploit the infected machine. In our case, the malware is the code installed on

the vulnerable machine that turns it into a bot and allows the botmaster to take control. Malware spreads using multiple techniques, and one specific malware can use more than one of these.

A virus is a malicious piece of code that hides inside a host program. It replicates itself when the host program is executed and inserts its code into other programs.

Worms are standalone programs that replicate and spread themselves when introduced to a computer network.

The name Trojan horse or Trojan describes any malware that hides their real and malicious intent. This can be a seemingly useful program, an unsuspecting e-mail attachment, or an interesting advertisement.

11.2.1.3 Revenue Models

In today's botnet economy, the attacker using a botnet is rarely the creator and botmaster of the used botnet. Botnets are traded and rented, to generate revenue without performing attacks [39]. There already exist services for development, distribution, and hosting of botnets. In many cases, this even includes customer support. Bottazi and Me (2014) call this model Cybercrime-as-a-Service (CaaS) [9].

With Distributed Denial-of-Service (DDoS) attacks, revenue can be generated by executing attacks on competitors in the market and getting a part of their market share or executing attacks on behalf of a competitor in exchange for payment. Another possibility is cyber extortion by threatening a DDoS attack if a certain amount of money is not paid to the attacker.

Theft and fraud: Botnets are responsible for the majority of spam distributed through the internet (i.e., phishing mail). They are used to host fake websites (phishing websites) to steal personal and valuable information, and they enable click fraud [69]. Click fraud is the exploitation of pay-per-click online advertising by imitating a real user and clicking on the advertiser's link. The advertiser has to pay the publisher and the advertising network, therefore creating a potential conflict of interest.

11.2.1.4 Command and Control

Command and control channels (C&Cs) are used to communicate with the bots in a botnet. Those channels are based on basic internet communication protocols. The majority of known botnets are using C&Cs based on the IRC (Internet Relay Chat) protocol. The botmaster can send and receive messages to and from his bots through a centralized command and control mechanism. The communication is in real-time and is highly successful. There are known botnets using the HTTP protocol for their C&C. This approach is still centralized, but the botmaster cannot directly interact with his bots. The bots have to contact the C&C server to get their commands. [21]

11.2.1.5 Architecture

The architecture refers to the model of communication between the botmaster and his bots. We divide these models into three categories: centralized, decentralized, hybrid, and unstructured.

In the centralized model (Figure 11.2), there are one (or few) central points responsible for the exchange of commands and data between the botmaster and his bots. This central point is the C&C server and runs on a machine with a high bandwidth connection. This server runs a communication service (mostly IRC). This model provides a small latency, which makes it easy to control the botnet. A disadvantage of this model is the high vulnerability of botnet communication, because of the centralized point is responsible for the whole C&C communication. If this central point gets discovered and eliminated, it

renders the whole botnet useless. This threat can be reduced by using multiple redundant servers controlling the same botnet.[75]

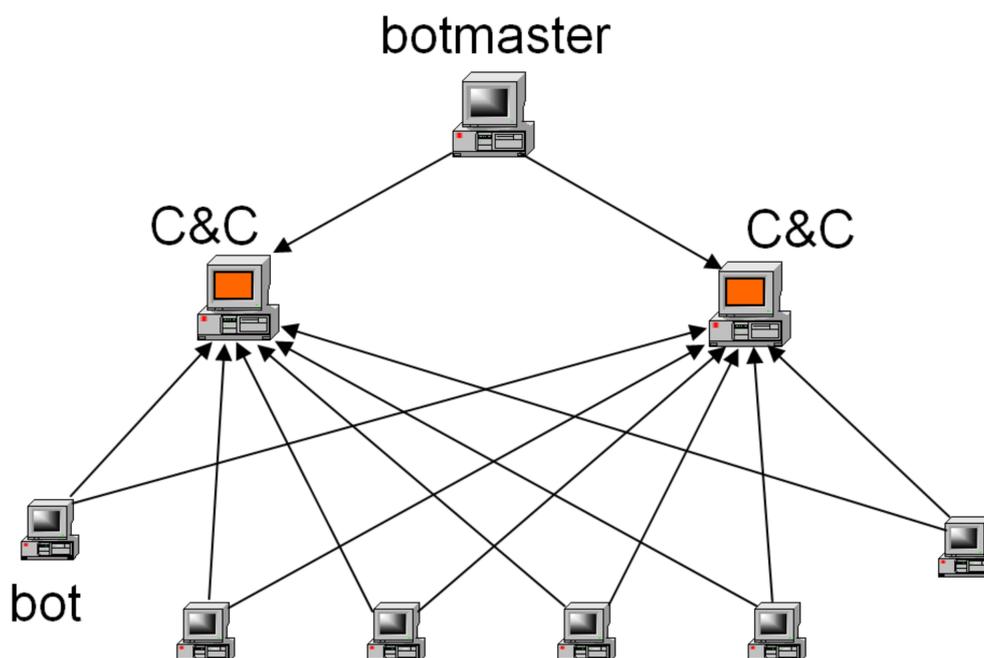


Figure 11.2: Command and control architecture of a Centralized model.[75]

The decentralized model (Figure 11.3) is a way to eliminate the vulnerabilities of the centralized approach. By using a P2P pattern for C&C, the system no longer depends on a few selected servers. The bots are interconnected and function as hosts and as servers simultaneously. Each bot knows a fraction of the botnet to send and receive the commands of the botmaster. So even if some bots are detected and taken down, the rest of the botnet continues to receive commands. Zeidanloo and Manaf (2009) stated that the use of P2P based communication in botnets would be used dramatically in the near future because botnets using this form of communication are much more challenging to detect and destroy.[75]

Wang et al. (2010) proposed a hybrid model (Figure 11.4) that uses the best of both worlds. Such a botnet would contain two types of bots. The Servant Bots, which serve as clients and hosts, have static and routable IP addresses. These Servant Bots are accessible from the whole internet, which means they are not behind a firewall that restricts incoming traffic. The Client Bots do not accept incoming connections and have a peer list containing only Servant Bots. The Client bots regularly connect to the Servant Bots in their peer list and forward any new commands to the whole list.[70]

An unstructured communication model is based on the principle that if a bot receives a command from the botmaster, it randomly scans the internet for other bots to propagate the message. In this case, the rest of the botnet will not be affected if a bot gets discovered. The drawbacks are incredibly high latency, the noticeable scanning and that it cannot be guaranteed, that every bot in the botnet receives the command.[15]

11.2.1.6 Underground Markets

As Thomas et al. (2006) state in their article, that there are entire IRC networks dedicated to cybercrime. Those IRC servers are not hidden but easy to find and easily accessible. The participants in these underground networks often use encryption to hide their identity. Most of these networks have channels for helping new members and reporting fraud. Members known to have committed fraud are recorded and shared on the network as

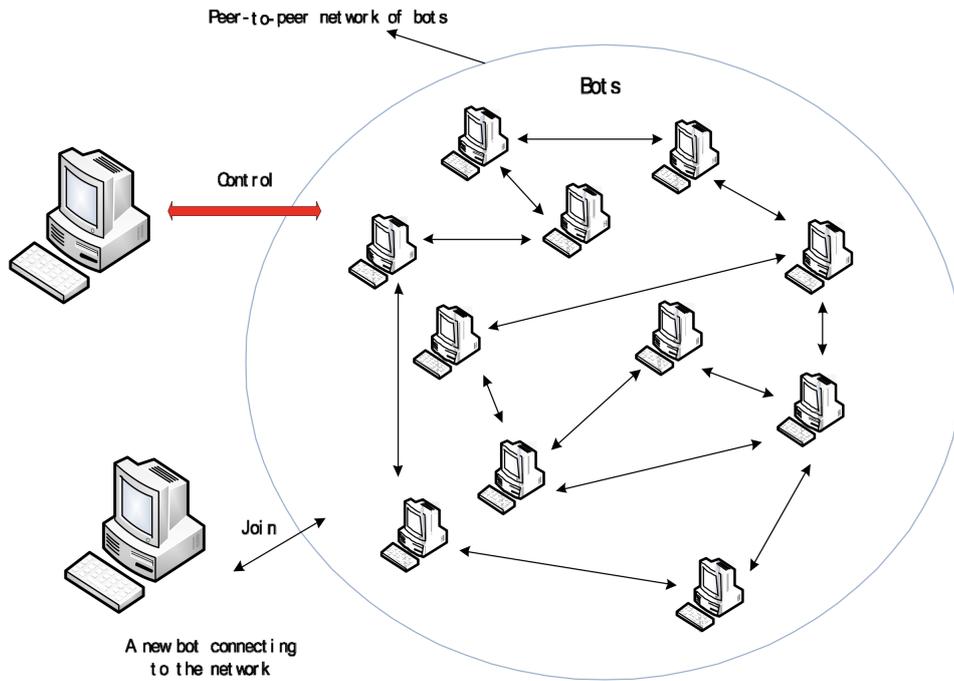


Figure 11.3: Example of Peer-to-peer Botnet Architecture.[75]

a form of self-policing. There are also other such underground networks using HTTP, Instant Messaging, Peer-to-Peer (P2P) and other forms of communication.[66]

11.2.1.7 Distributed Denial-of-Service Attacks (DDoS)

A Denial-of-Service (DoS) attack attempts to prevent the legitimate use of a service by either overwhelming the service with a huge amount of traffic or exploiting an application or protocol by sending malformed packages and causing the service to freeze or reboot. A Denial-of-Service (DoS) attack becomes a Distributed Denial-of-Service (DDoS) attack if the source of the attack is multiple distributed entities or rather bots in a botnet (Figure 11.5).[50]

An increasingly popular form of DDoS attack is DDoS amplification attacks, like the Memcached DDoS attack on GitHub on February 28, 2018. The attacker abused Memcached instances to amplify the attack by up to 51'000 times the sent data initially. The Memcached's response has been targeted to addresses used by GitHub.com using IP address spoofing. This amplification resulted in a peak of 1.35Tbps via 126.9 million packets per second sent to GitHub's services (Figure 11.5).[25]

11.2.2 DDoS Defense Systems

11.2.2.1 Technical Solutions

a. Aggregate-based congestion control (ACC)

Mahajan et al. (2002) introduced two aggregate-based congestion control (ACC) mechanisms. The job of the first, local ACC is to identify the aggregates responsible for the congestion of service and to throttle the throughput of these.[45]

The identification of the offending aggregates is challenging. The overload may be chronic due to an under-engineered network or unavoidable because the load shifted due to routing changes. The traffic might cluster in multiple dimensions (i.e., a particular server or network link) and the attacker may change their target to avoid detection.[45]

The next challenge is to determine how much an aggregate should get throttled. The goal is to keep the service running during an attack, which implies they cannot block those

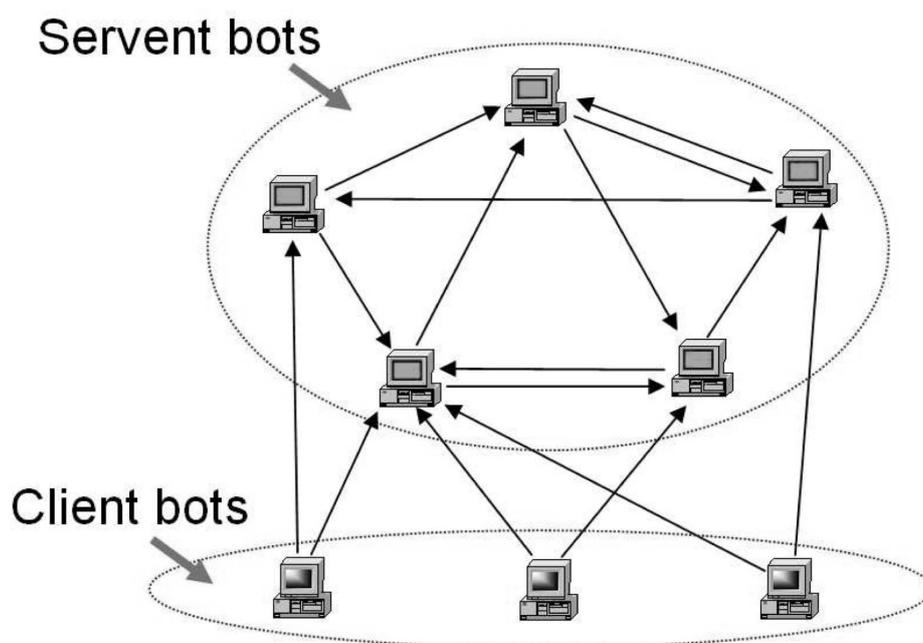


Figure 11.4: C&C architecture of the proposed hybrid P2P botnet.[70]

aggregates completely. The rate limit is chosen so that the remaining traffic can at least maintain some level of service.[45]

The second ACC mechanism is the pushback mechanism. The congested router pushes the aggregate throttling upstream to the adjacent routers sending him a significant fraction of the congesting traffic. A pushback strategy propagates upstream to save bandwidth by dropping packets early and to focus on the routes or routers responsible for the congestion and still letting legitimate traffic through. [45]

The pushback will not be effective against a DDoS attack if the traffic is distributed evenly across the inbound connections. This can be possible if the attacker uses an amplification (or reflector) attack with sufficiently distributed reflectors. Pushback can also overcompensate by throttling aggregates upstream that would not have been a problem downstream. Alternatively, if the pushback algorithm is not able to differentiate between malicious and legitimate traffic coming from the same edge network, that does not support the pushback mechanism. This fact could even be abused by blocking legitimate incoming traffic from a particular source by launching the attack from a host that is close to this source.[45]

b. Max-Min Fair Server-Centric Router Throttles

Yau et al. (2005) proposed and tested a router throttle mechanism that aims to increase the precision in blocking malicious and protecting legitimate traffic. This is a server-initiated proactive approach, where the server protects itself by installing a router throttle at an upstream router. The throttles depend on the current demand and have to be negotiated dynamically between server and network.[73]

If a server load crosses the load limits, the server installs the throttle at some of its upstream routers. There is a defined upper and lower limit, and if reached, the throttle rate is reduced further or increased again.[73]

The fair throttle algorithm tries to minimize the throttling on well-behaving routers, by not just decreasing the allowed traffic from every router, but by multicasting a uniform leaky bucket rate (i.e., the throttle rate) to all targeted routers. This way, busier routers drop much more packets than calm ones, which may not have to drop any.[73]

c. Probabilistic Packet Marking

Park and Lee (2001) discussed the effectiveness of IP traceback mechanisms based on probabilistic packet marking (PPM). Probabilistic packet marking aims to reduce the

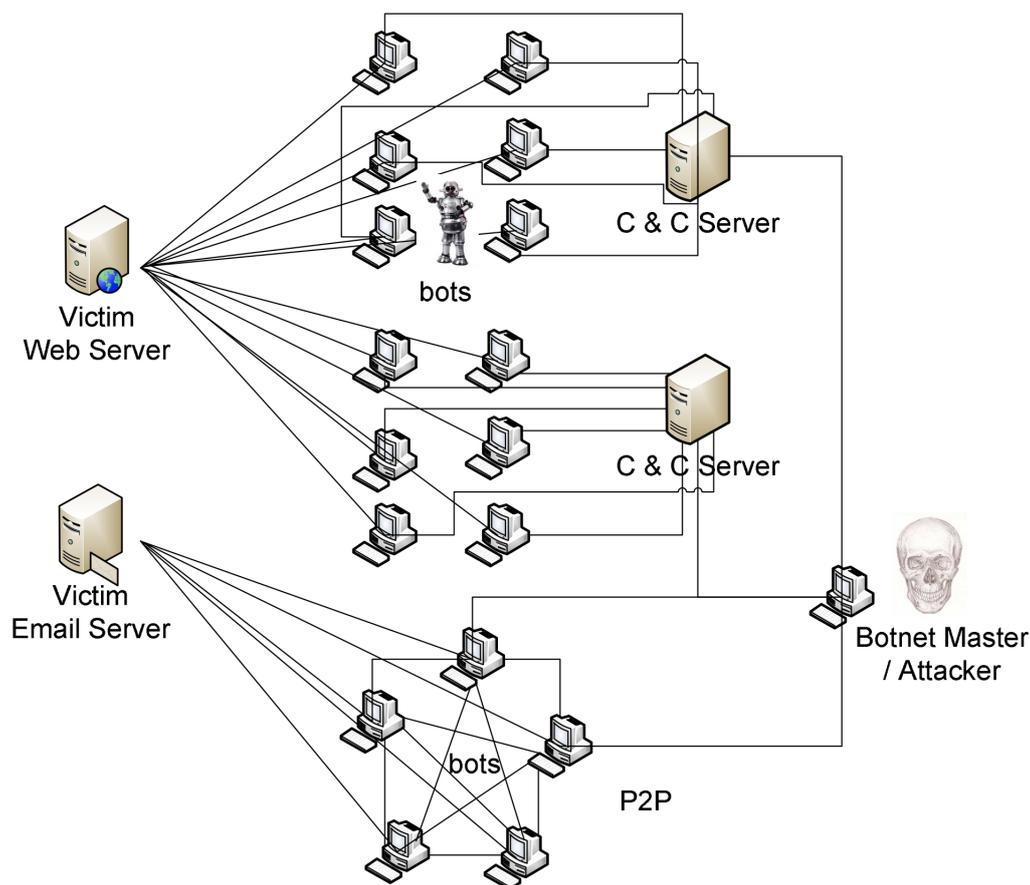


Figure 11.5: A scenario of botnet attacks launched by robot computers (bots) controlled by the botnet master and attacker.[39]

overhead produced by deterministic packet marking (DPM). In DPM, each router inscribes its local path information onto the passing packet. Those markings allow the destination node to trace back the traversed path of a packet but increases the packet header size linearly with every hop.[56]

In PPM, the packet header provides a constant space for traceback information, and after every hop, the router overwrites this information with a probability $p < 1$. This allows the destination node to trace back the whole path of the incoming packets, if the attack volume is high enough but also leaves the possibility open, that some packets still contain the original markings of the attacker, who can try to confuse the traceback. [56]

In the case of DDoS attacks with a high number of attacking sources, this technique gets rendered more and more useless. The uncertainty grows more and more while the probability of finding the sources becomes smaller and smaller.[56]

d. Hash-Based IP Traceback

Snoeren et al. (2001) present a hash-based IP traceback technique that allows us to trace back the origin of a single IP packet. They developed the Source Path Isolation Engine (SPIE). The engine enables us to trace back a packet with a copy of the packet, its destination, and approximate reception time. The main goals of this technique are to reduce memory requirements and not to increase the vulnerability to eavesdropping of a network.[62]

To achieve those goals, SPIE produces a 32-bit packet digest using a hash function. As hash input, it uses the invariant portion of the IP header and the first 8 bytes of the payload. Because it is not possible to store the digest of every packet forwarded on the router, SPIE uses a Bloom filter. This filter computes k distinct digests using k distinct hash functions. The hash function has to be uniform. The n -bit results get indexed to a

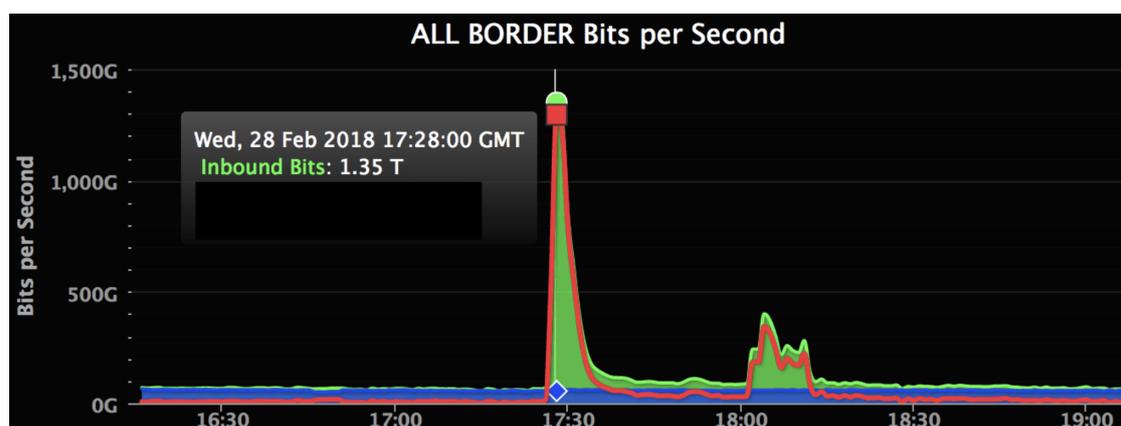


Figure 11.6: This graph provided by Akamai shows inbound traffic in bits per second that reached their edge.[25]

bit array of the size $2n$. If the copy of the file to trace back gets indexed to only 1-bits, with high possibility, the packet got forwarded by this router.[62]

e. Packet Filtering using Traffic Quota

The DDoS Defense System proposed by Xu and Lee (2003) assumes that the firewall is the bottleneck during a DDoS attack. The Server and Client do not need to be changed, and the firewall issues the filtering operation. The goal is to produce a line of defense in the local network with a set of routers, typically belonging to a local ISP. Those so-called perimeter routers should distinguish between DDoS and legitimate traffic. They distinguish two types of DDoS attacks, the first where the attacker uses IP spoofing and the second where the legitimate IP addresses are used.[74]

In the first case, if the first packet of a client arrives, the client gets redirected to a pseudo-IP address and port number pair. This new destination IP address contains a Message Authentication Code (MAC) that is encrypted with a symmetric key shared between the firewall and the perimeter routers. If the sender's IP address is spoofed, the HTTP redirect message will never reach him. To avoid the collecting of valid MAC's from legitimate clients, the Key changes over time and gets a small timestamp or version number.[74]

in the second case, where the attacker uses the real IP addresses, they cannot be distinguished from legitimate customers. The firewall has the job to allocate the available bandwidth fairly between all clients. Using a Deficit Round Robin algorithm, every client gets a traffic quota, and all excess packets are dropped. The firewall tracks the amount of dropped packets and blacklists clients overreaching a particular threshold. To defend against a continuous attack that stays in the traffic quota, a "no loitering" law is enforced. If the total amount of packets sent by a client is more significant than a given quota, its traffic quota gets reduced to a fraction.[74]

The main problems are that an IP spoofed DDoS attack can make it hard for legitimate clients to get their first package through to get redirected and if all attackers use their fair share using the genuine IP address, the service can suffer a response time too big for real customers to deal with.[74]

f. Hop-count Filing

Jin et al. (2003) propose a hop-count-based approach to filter out spoofed IP packets. They use a victim-based approach that does not need the support of the ISP's. The only information needed is contained in the package header, assuming that in most cases of IP spoofing, the hop-count of the package does not correspond to the expected hop-count of the IP address. The hop-count can be inferred from the TTL field of the incoming packets. To avoid dropping packets of legitimate clients because the hop-count information is not correct, the filtering only takes place if an attack gets detected.[22]

11.2.2.2 Economic DDoS Defense

Defending against DDoS attacks with technical solutions results in an arms race between attacker and defender. It has become more common to not only look at technical solutions against direct DDoS attacks but to target the economic aspects of botnets.

Ford and Gordon (2006) describe their Multihost Adware Revenue Killer (MARK), to attack the revenue streams of botnets used for unwanted advertisement and software installs on the victim's machine (Adware and Spyware). This use of botnets is thought to be the most profitable for botmasters. MARK consists of Virtual Machines (VMs), which actively try to get infected by Ad- and Spyware, so-called honeypots. Because no one sees the ads shown on those VMs and there is no personal information to find, the botmaster's revenue increases first, but the advertisement or spyware distributor loses money. This destabilization can lead to a massive decrease in payment (i.e., per click or download), and therefore a massive decrease in revenue for the botmaster when the VMs get cleared and the honeypots are no longer generating revenue.[20]

Li et al. (2009) propose a similar strategy, intending to make it more troublesome for botmasters to maintain their botnets. They focus more on DDoS attacks and propose to use honeypots to introduce uncertainties to the optimizing problem of botmasters and clients. The honeypots do not participate in DDoS attacks; for this reason, the required number of bots for a successful attack increases, and the value of a single bot decreases by consistent maintenance costs. According to their model the profits of attackers and botmasters can decrease dramatically.[39]

11.3 Economics

As a subset of the cybercrime economy, the botnet economy plays out on underground black markets. Over time there has been a shift in incentives, where the attackers are no longer motivated by self-fulfillment and proof of skill, but rather the financial gain attributed to such attacks. As part of this shift towards purely economic incentives, cybercriminals have developed business models that shall guarantee the profitability of their botnets.[39]

The many victims affected by botnets are faced with a severe economic threat. Successful DDoS attacks on companies may lead to both direct economic costs, as well as indirect losses, which are much harder to quantify. Direct costs occur due to lost revenue during the server's downtime. Indirect losses measure the long-term effect loss of customer's trust, and market share has on revenue. The value destroyed by these attacks also inflicts costs on society at large. The damages, although difficult to quantify, are estimated to be in the tens of billions of dollars every year.[7] This chapter shall give an overview of the different stakeholders involved, the business models used to operate successful botnets, as well as an attempt to model the different types of economic damage a botnet attack can inflict.

11.3.1 Stakeholders

In order to gain a better perspective on incentives and the botnet economy as a whole, we first have to define the relevant partners effected by botnets. We will analyze both the parties involved in the attacker and the victim's sides.

11.3.1.1 Attackers

The following is a list of key partners involved on the side of the attacker. We based this list on papers written by C.G.J. Putman et al.: "Business Model of a Botnet" and G. Bottazzi et al.: "The Botnet Revenue Model" [60, 9].

a. Malware Developers

Malware Developers are mainly IT professionals involved in the research and development stage of the botnet life cycle. They identify vulnerabilities in software, develop new malicious software to exploit these vulnerabilities, and maintain existing software. No illegal activities take place in this step of R&D. Only the use of the software is considered illegal. [60, 9]

b. Money Handlers

The goal of money handlers is to offer payment services with a high degree of anonymity. They allow people buying malicious software to remain anonymous. The malware developers from the first stage also have an interest in remaining unidentifiable, as to not be charged as an accessory to any crimes committed by the buyer. The supply of anonymous payment services is also not considered illegal and can be achieved through the use of cryptocurrencies.[60, 9]

c. Command and Control Bulletproof Hosting Providers

Bulletproof Hosting allows the botnet end-user to store any stolen information such as passwords, banking accounts, or other credentials. The control center for the botnet, which can also be provided through the hosting service, consists of one or multiple computers.[60, 9]

d. PPI (Pay Per Install) Distributors

The distribution and multiplication of malicious software are continuously evolving, and many developers lack the necessary resources to spread their malware effectively. For this purpose, developers use a PPI model, which essentially allows them to spread their software by commissioning third parties with the infection and paying them a certain fee per infected device. According to research by Brian Krebs proposed in his paper "Most Malware Tied to Pay-Per-Install Market," PPI costs are estimated to vary from \$7 to \$180 per 1000 installations.[60, 9]

e. Botnet Owners

Botnet owners, also known as botmasters, are the ones who perform the attacks [60, 9]. They either perform these malicious activities for their benefit or on behalf of third-party users for financial gain. The latter is becoming increasingly important.[39]

f. Botnet Users

Botnet users are the final piece in the botnet supply chain. They are customers willing to pay money to make use of the botnet and perform attacks. Users are mainly charged by the pay-per-use model.[60, 9]

Figure 2.7 provides an overview of different stages of the botnet life cycle and the parties involved.

11.3.1.2 Victims and Defenders

The following is a list of victims affected by botnet attacks and third parties involved in defensive measures, as well as their incentives to protect themselves from these attacks.

a. Victims of DDoS and Spamming

The first and probably most obvious victims are the ones that suffer from DDoS attacks and spamming carried out by botnets. The victims of such attacks suffer far-reaching financial consequences from such attacks. Frequently, the targets are businesses that suffer direct losses in the form of lost revenue due to the unavailability of their online services. Nevertheless, they also suffer indirect impacts, such as a loss of customer trust and market

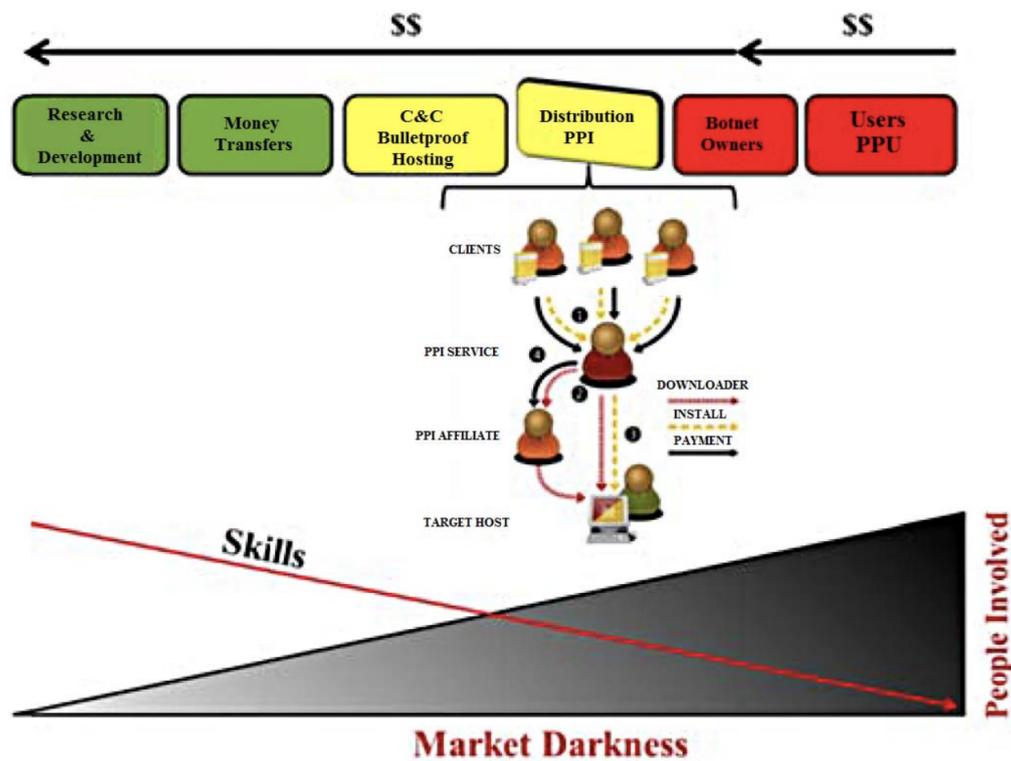


Figure 11.7: The Botnet Revenue Model [9]

value. These damages are usually far more severe and harder to quantify due to the difficulty of attributing these losses directly to botnet attacks [4]. As a result, businesses with specific exposure to malicious attacks and a firm reliance on online presence, have powerful incentives to protect themselves from botnet attacks. Additionally, the loss of customer trust and market shares that these attacks entail, leading companies to try and keep information about suffered attacks out of the public eye.

b. End-users of Infected Devices

Another prominent direct victim is the end-users of the infected devices. In the case of DDoS and spamming attacks, the users of infected devices are not directly affected by any adverse consequences, as they are used as a means to an end. Therefore, they may lack specific incentives to protect themselves adequately from malware and not become part of a larger botnet [7]. Why should a device owner pay for security measures, when he does not suffer any direct economic consequences from being infected? Another explanation might be that many end-users are not aware of the danger malware entails and therefore lack the knowledge required to protect themselves from such attacks adequately.

Other forms of botnet attacks, such as identity and information theft, bank fraud, and phishing, do, however, pose a direct threat to owners of infected devices. According to an estimation by Anderson et al. (2013), consumers globally suffer losses in the range of \$70 million due to online banking fraud [4]. For business, that number rises to around \$300 million. Furthermore, Anderson et al. (2013) estimate that global expenditure on antivirus protection reached \$3.4 billion in 2012 [4]. While the annually indirect cost for malware removal is estimated to be around \$10 billion globally and \$500 million in the UK alone.

c. Antivirus Software Vendors

As mentioned in the section above, security vendors have powerful financial incentives to protect malicious software, as the global expenditure is estimated to be around \$3.4 billion, and a vast majority of consumers have some antivirus software. Additionally, global expenditure on clean-up is estimated at \$10 billion and therefore provides even more financial motivation for security vendors.[4]

At this point, however, it has to state that this includes cost for protection against malware in general and not only related to botnet attacks. Vendors of such software are also faced with extensive development and maintenance costs, as cybercriminals continuously try to find new ways of circumventing existing protection measures.

d. Internet Service Providers

Internet Service Providers (ISPs) play a significant role in botnet mitigation, as they can perform measures on a vast scale and, therefore, more economically than single end-users. They do, however, lack accountability, as they are not the cause of botnets, and economic motivators, as such countermeasure can be very costly, and ISPs do not suffer any direct consequences by users being affected by cybercrime. Advanced detection and follow up actions can also cause privacy concerns of customer data.[58]

Still internet service providers do play a role in botnet mitigation due to organizational and institutional incentives. "Relevant organizational factors for ISPs to address botnet mitigation include the size of their customer base, the internal organization of their abuse desk, and the cost spend on various security measures." (Pijpker et al. (2016)) [58]. Institutional incentives, on the other hand, are not imposed internally but rather by policymakers and market conditions. They include for example the regulatory context and market settings.[58]

Research shows that ISPs currently focus more on prevention and notification of customers and that there is still quite a lot of room for improvement due to a lack of incentives for botnet detection, remediation, and recovery on behalf of internet service providers.[58]

e. Law Enforcement

As costs for cybercrime mitigation can be very high, additional investments in law enforcement can be an effective countermeasure. The cost of law enforcement is generated by the time required to find and prosecute cyber criminals [4]. Anderson et al.[4] estimate that the global expenditure on law enforcement for cybercrime defense is around \$400 million, while that of the UK is about \$15 million in 2010 [4]. When compared to the cost of cybercriminal infrastructure of private corporations, this modest number may be one reason for the low success rate in finding and punishing cybercriminals.

11.3.2 Revenue Streams and Incentives for Botnet Owners

Even though botmasters are not always solely financially motivated, there are several potential revenue streams for botnet owners. The following chapter shall provide a non-extensive list of possible revenue streams for botnet owners with some estimations on the revenue they generate. The following figure provides an overview of all the different revenue streams mentioned in this chapter.

11.3.2.1 DDoS Attacks

DDoS attacks are one of the most common malicious activities performed with the use of botnets that can be highly lucrative. The goal of a DDoS attack is to force a computer system into denial of service for legitimate customers by sending enough requests to the computer of a victim that it can no longer process all of them [52]. Two possible revenue streams can be implemented for DDoS attacks.

The first of which is extortion [8]. Owners of botnets can generate profits by merely threatening to perform a cyber-attack. Often large companies give in to such demands due to the immense costs associated with a successful DDoS attack, both direct and indirect [52].

Additionally, botnet owners can offer their services in underground forums in the form of a DDoS-for-hire service that operates on a subscription or pays per use basis. Prices

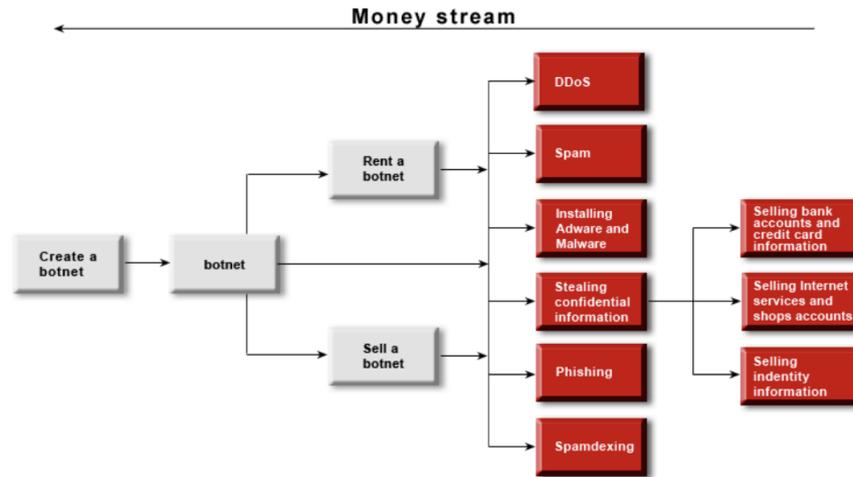


Figure 11.8: Botnet Revenue Streams [52]

for single DDoS attacks vary mainly on the number of bots needed to perform the attack. They can range from only about \$50 for an attack on a small unprotected online store to several thousand dollars for attacks on large international companies who have well-protected web sites. This makes sense, as the cost of operating a botnet increases immensely with a higher number of devices controlled by the botnet [52]. Subscriptions allow users to manually perform an unlimited or limited number of attacks any time they would like through front-end web servers. Prices for these subscriptions vary from just \$5 up to \$300 dollars and allow customers to use the service for one to three months, depending on the tier of the subscription [8].

One example of such a booter service is VDoS. VDoS generated revenue in the region of \$25,985 per month over two years of operation, with their lowest-performing month at \$9,956 and their best performing month at \$42,924 [8]. As they are almost impossible to quantify, there is no information on profits generated from extortion [52].

11.3.2.2 Information Theft

Another source of revenue for botnet owners is the sale and misuse of personal information such as bank accounts and credit card information, online credentials, or identity information. This type of data is directly extracted from the infected bots [9]. Cybercriminals can decide whether they want to use the gathered information for their use or sell them to third parties. The price for a bank account can range between \$1 and \$1,500. The low value of this information can be attributed to intense competition from other vendors. In order to generate a lucrative and sustainable revenue stream, botmasters always need an inflow of new data and a large botnet [52].

One example that shows how lucrative bank account theft can be is the case of Eurograbber. In an attack on European online banking customers through a Zeus based mobile trojan, the perpetrators were able to illicitly transfer funds over 36 million Euros from more than 30,000 bank customers.[23]

11.3.2.3 Phishing

In order to protect phishing sites from being shut down and extend their lifetime, cybercriminals are offering an implementation of fast-flux technology, which allows phishers to change their web site's IP address every few minutes without altering the domain name. This technique of using infected home computers as web servers with phishing content also makes them harder to track down.[52]

In order to acquire a ready-made botnet for their fast-flux solution, phishers pay botnet owners \$1,000 to \$2,000 per month. The average income and damage to consumers is in the millions of dollars per year and comparable to that from the theft of personal information.[52]

11.3.2.4 Spam

One of the most important applications for botnets is sending spam. Spammers use different business models, and some make profits by luring users to infected websites where they might download malware while others are part of phishing campaigns designed to steal user credentials for reasons mentioned above. Botnets provide the required infrastructure necessary for carrying out spamming services.[8]

Prices for spamming services depend on the size of the target audience and can range from just \$70 for a few thousand addresses up to \$1,000 for tens of millions of addresses [52]. In their paper "The Economics of Spam," Justin Rao and David Riley [61] estimate that American firms and consumers experience costs of about \$20 billion per year due to spam messages. Furthermore, they estimate that spammers and spam-advertised merchants make revenues of about \$200 million worldwide annually. This means that the externalities caused by spam messages have a ration of 100:1 [61].

11.3.2.5 Spamdexing

Spamdexing, also known as search engine spam, is a form of deliberate manipulation of search engine indexes to improve a website's position in search engine results. The idea behind this is that websites with a higher position generate more clicks and, therefore, revenue.

There are several criteria search engines use to judge the relevance of a website for given search input. The number of links to that particular website on other websites is one of the main factors in determining which result is displayed at the top. Botnets are used to create many posts and comments containing the link to websites being promoted and related keywords. The price for such services averages at about \$300 per month.[52]

11.3.2.6 Adware and Malware Installation

An additional revenue stream for botnet owners can be the installation of adware on infected computers. Adware is software that generates revenue for its creator by automatically and unwantedly displays advertisements in the user's interface, most often a web browser [46]. Companies that offer such adware often pay between 30 cents and \$1.50 for the installation. With thousands of computers at their fingertips, the installation of adware and other forms of malware can be a lucrative and straightforward way to generate additional profits for botnet owners. A famous example is John Kenneth Schiefer, a hacker convicted in 2007 that used a botnet to distribute adware and managed to generate profits of around \$14'000 in one month.[52]

The pay-per-install business model is also used by cybercriminals when distributing malware. The average price for installation varies depending on the geographical location of the computer, as an infected computer in developed countries is capable of generating a lot more profit because the information is frequently more valuable. The price for 1,000 installs in the United States, for example, can reach up to \$120, while the same number of installs on Chinese computers generates about \$3 in profits.[52]

11.3.2.7 Cyber Fraud

Cyber fraud has many different forms. One widespread form is click fraud, where websites and botnet owners generate profits from clicks on advertisements. This type of click fraud relies on pay-per-click advertising, where companies that place advertisements on other websites, through Google AdSense, for example, pay for these ads based on the number of clicks that they receive. A clickbot helps an attacker commit click fraud by generating fake clicks on ads. This technology can be used in a few different ways. For example, a company can generate many clicks on the ad of a competitor in order to diminish its budget and create more costs. Alternatively, websites that publish the advertisements can also use clickbots to click on the ads placed on their websites, as they receive a share of the profit from ad clicks [16]. Botnets allow attackers to generate thousands of clicks per day, all from different computers as to not raise any suspicions [52]. By analyzing Clickbot.A, N. Dawasi and M. Stoppelman estimated that Google could lose around \$50,000 with a botnet of 100,000 bots. A share of this profit generated by advertisers then goes to the botnet owners [16]. Another click fraud botnet, ZeroAccess, has generated ad losses up to \$900,000 daily with 140 million clicks a day [9].

Another way to commit cyber fraud is by creating fake user ratings and customer reviews on websites or with the rising importance of cryptocurrencies, and botnet owners can abuse the computing power of their botnets to mine cryptocurrencies [60].

11.3.3 Modeling Costs

Costs of cybercrime are hard to model. There are many aspects of the costs to an individual, company, or a whole economy originating in cybercrime. As an approach to model those costs, Anderson et al. propose a framework to model the costs of cybercrime in the paper, measuring the changing cost of cybercrime. They concluded that there are four categories of costs due to cybercrime. Graphical representation of the framework is given in the figure 11.9.[5]

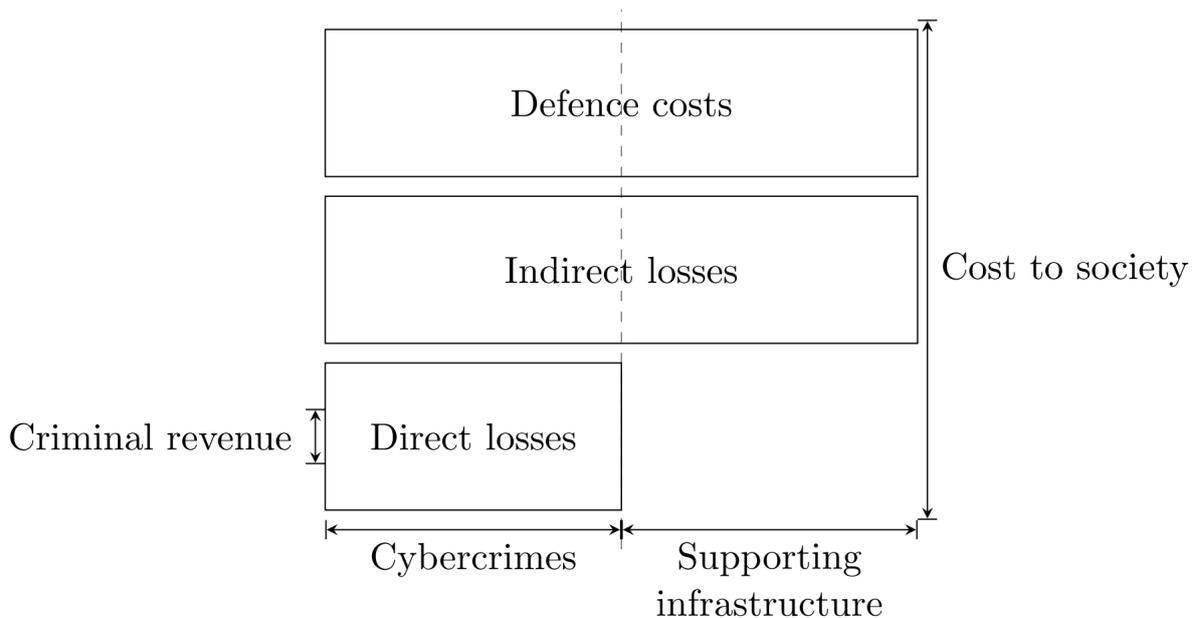


Figure 11.9: Costs of Cybercrime [5]

Criminal Revenue are the direct revenues generated by the criminal. As examples here could be named the money withdrawn from an account via credential theft, ransom

paid in consequence of infecting devices with ransomware or the income generated by a botmaster by sending spam or selling installations on the infected devices.[5]

Direct Loss are the losses and damages on the side of the victims as a direct consequence of the criminal activity. For example, securing the compromised bank account after the withdraw form the first example or setting up the infrastructure after the ransomware has encrypted everything would be direct damages and losses.[5]

Indirect costs are generally speaking the opportunity costs of the cybercriminal activity. They can generally not be attributed to individual victims or perpetrators. Examples for those losses would be the loss of trust in service after it has been hacked which leads to the loss in the income generated by this service or the loss of appointments after a booking platform is taken down via a Distributed denial of service attack.[5]

Defence costs are the costs that occur in prevention efforts. Examples here would be the purchase of security products that aim to prevent cybercrime such as anti-virus software or spam-filters, the costs of law enforcement trying to defend against cybercrime or the opportunity costs of missed appointments by marking them falsely as spam.[5]

To get the overall costs of cybercrime, the four types of costs must be added together to build an approximation of the crime. This leads to the following formula to asses the cost of cybercrime.

$$\text{Cost of cybercrime} = \text{Criminal Revenue} + \text{Direct Losses} + \text{Indirect Losses} + \text{Defence Costs} \quad (11.1)$$

To model the costs for a specific enterprise, the cost of cybercrimes such as DDoS attacks can be approximated by modeling the cost of downtime for this enterprise. Patterson (2002) has stated a formula on in his paper to measure the cost of downtime in an effective and straightforward way [57]. He states that the estimated costs of one hour of downtime are the sum of the losses by the employees that would have worked but could not and the average revenue that is affected by the outage that would have been generated otherwise. This leads to the following equation for the cost per hour of downtime:

$$\text{cost/hour} = \text{Empl. cost/hour} * \frac{\text{Empl affected}}{\text{All empl}} + \text{Rev./hour} * \frac{\text{Rev. affected}}{\text{All Rev.}} \quad (11.2)$$

According to Gartner [37], this cost of downtime can be set to approximately \$ 5600 per minute. This number comes from a survey they did in the year 2014. This adds up to over \$ 300000 per hour. This is just an approximation of the number and an average. Although the \$ 5600 estimated by Gartner is high, this estimate is based on five-year-old numbers. Since there are many shifts in today's economy, with things like industry 4.0, the dependence of enterprises on their IT infrastructure has risen significantly over the last year.

A more recent study by the Ponemon Institute has an even higher estimate. They claim that the cost per minute of the downtime of an enterprise IT network would be as high as \$ 9000 per minute [59].

11.4 Case Study: Mirai

11.4.1 The Botnet

11.4.1.1 Attacks Timeline

Mirai is Japanese and stands for "the future". The Mirai Botnet is one of the most prominent botnets today, and its principal characteristic is the abuse of poorly secured IoT devices as bots. The following is a timeline of the beginning of Mirai [24, 6].

At the end of August 2016, Mirai was identified by MalwareMustDie. "MalwareMustDie, as a white-hat security research workgroup, launched in August 2012, is a non-profit organization media for security professionals and researchers gathered to form the workflow to reduce malware infection on the internet." [42]

Mid-September of 2016, the website of Brian Krebs got attacked as the first prominent target of Mirai. Krebs' is a computer security consultant and journalist, writing articles on his website krebsonsecurity.com. The website was hit with 620 Gbps of traffic, which is much more than would be needed to knock down most websites. [28]

Around the same time, an even bigger attack launched on to the French WebHost and cloud service provider OVH. This attack peaked at 1.1 Tbps.

At the end of September 2016, the creator of Mirai released its source code to the public, going by the nickname "Anna-Senpai." [34] After those Mirai botnets with as many as 400'000 connected devices got offered on the underground market by members of the hacker community. [55]

At the end of October 2016, several high traffic websites went offline for multiple hours resulting in an attack on the service provider Dyn. [51]

On the 31 of October, the Liberian telecom company Lonestar Cell got attacked. [6] The Mirai author got identified in an article by Brian Krebs in early 2017. [34]

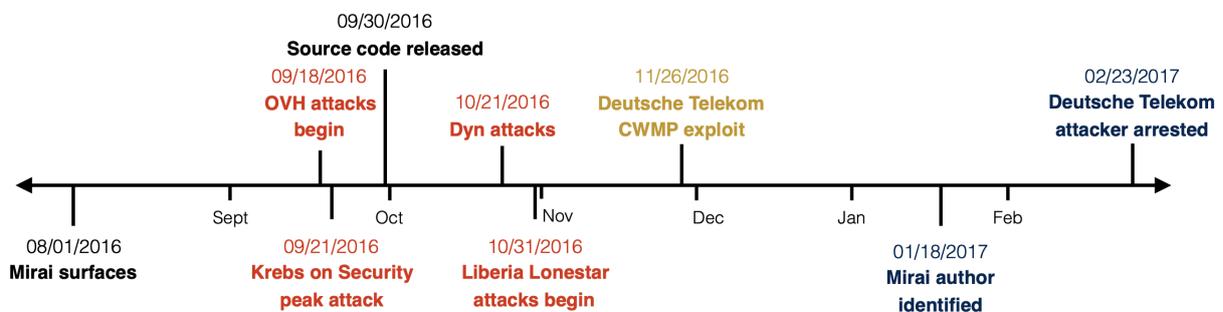


Figure 11.10: Mirai Timeline – Major attacks (red), exploits (yellow), and events (black) related to the Mirai botnet. [6]

11.4.1.2 Operation

Unlike many other botnets, Mirai primarily targets Internet-of-Things (IoT) devices running on ARC processors, such as IP cameras and home routers. In a first step, Mirai enters a rapid scanning stage where it aimlessly sends out TCP SYN probes to pseudo-random IPv4 addresses on Telnet TCP ports 23 and 2323, excluding those in addresses in a hard-coded blacklist [24]. On recognizing a potential victim, Mirai then tries to log into the device with a brute force attack using ten randomly selected username and password combinations out of a list of 62 common factory default usernames and passwords (e. g. 'admin' & 'password'). [6]

After a successful login, various characteristics are sent to a report server through a different port. These reports are used by a C&C server to determine new prospective

victims continually. This server also communicates with the report server to get the current status.[24]

The botmaster then sends an infect command in a separate loader program with the necessary details, such as IP address, system environment, and hardware architecture. A separate loader program then logs into the target device and gets it to download the corresponding binary, usually via GNU Wget or using FTP (File Transfer Protocol).[24] The loader then executes the corresponding binary version of the malware on the device. If the execution of the malware was successful, it is stored in the RAM of the device. Due to this fact, a reboot of the device will remove the malware. At this point, Mirai is ready for further instructions from the botmaster while continually scanning the environment for prospective targets.[24]

Mirai makes use of 10 different attack methods, which were distinguished in the source code within four categories TCP, UDP, GRE, APP. The TCP category includes methods such as SYN, ACK, and STOMP flood attacks. Under UDP, there are UPD, UDPPLAIN (UDP flood with fewer options and optimized for higher packets per second (PPS), VSE (Valve Source Engine specific), and DNS resolver flood attacks. GRE IP and GREETH flood attacks abuse the Generic Routing Encapsulation (GRE) protocol. On the other hand, APP only contains the beneficial HTTP attack [71].

11.4.1.3 Background

In January 2017, a post was released on Krebs on Security, which disclosed the name of the author of the Mirai botnet - Paras Jha, a 22-year-old student at Rutgers University. Jha publicly released the source code of the Mirai botnet on an infamous public internet forum "HackForums" under the name "Anna-Senpai" on September 30. In the post, he stated that he already made his money in the DDoS industry and that releasing the source code was a nice gesture of him.[34]

Following investigation and questioning by the FBI Paras Jha, as well as co-authors Josiah White and Dalton Norman, entered a guilty plea in December 2017 [67]. Jha was fined \$8.6 million and sentenced to 2'500 hours of community service for using Mirai to attack the computer system of his university [30].

Jha admitted to investigators that he launched his attacks on the computer system of Rutgers University not out of financial gain, but for other personal motivations. The first attack he launched was for delaying registration upper-class men registration to an advanced computer science attack he wanted to take, whereas the second attack was to delay an exam. Other attacks on the computer system were launched for his amusement to see outrage.[30]

11.4.1.4 Costs

In the Mirai case study of MalwareTech (2016), it was mentioned that the costs of maintaining desktop botnets exceed the revenue from DDoS attacks for most botnets. Due to the technological advancements in anti-DDoS services, DDoS protection are more affordable than they used to be, which leads to the decline of ransom-driven DDoS attacks.[43] The maintenance of a botnet consisting of IoT devices is more lucrative, which Mirai takes advantage of. Although rebooting would remove the malware, only changing the password or more drastic measures would remove the risk of reinfection. Furthermore, users generally have little control over IoT devices, which makes them only more attractive of a target. These factors lead to reinfection being very likely to succeed for the maintainer of a botnet.[43]

A person controlling the botnet only must maintain a list of the IP addresses, which have been infected in the past, and scan if they are still connected to the botnet.

Putman et al. (2017) estimated the cost for reinfection to be at \$0.0935 per device, and the cost for spreading to be at around \$0,016 per device. Also, they noted that the cost for infection is around \$7 to \$180 per 1'000 installations (\$0.007 to \$0.18 per device). They further estimated the cost of maintaining a botnet to be \$59 per hour based on an estimate of the average salary of a malware developer.[60]

Following the conclusion of another study Putman et al. stated that the initial investment such as acquiring and spreading the malware and the recurring costs are nearly insignificant compared to the revenue it makes. Out of three of four case studies, the set-up costs for a botnet accounted to a maximum of 1,1% of the found monthly revenue.[60]

11.4.1.5 Revenue

Although the \$59 per hour stated in the last section might seem costly, advertising and renting out a botnet costs little to no money. Jha stated that he and the other two co-creators profited from the attack on Krebs on Security where they rented out tens of thousands of infected devices to a customer.[29]

The 30-year old hacker Daniel Kaye, known under the names Bestbuy and Popopret, gave an example of the cost of renting out a modern variation of the Mirai botnet. A botnet attack of 50'000 bots with an attack duration of 1 hour and a 5– 10-minute cooldown goes for roughly 3–4k per two weeks [55]. Kaye stated that his botnet with the codename Mirai #14 supposedly infected over 1,5 million devices by 2019 [12].

Kaye was also responsible for the high-profile attack on Lonestar in 2016, Liberia's market leader in providing Internet and telephone services. According to his testimony, he was compensated with \$10'000 monthly by an unidentified senior official working for Cellcom, a rival internet provider.[26]

In January 2017, over two days, Kaye attacked several well-known banks in the United Kingdom, including Lloyds Bank, Bank of Scotland and Halifax intending to extort them. The Guardian reported that Lloyds was requested to pay approximately £75'000 in bitcoins for the attacks to halt, to which they declined.[14]

Botnets similar to Mirai or early prototypes were already in use as early as 2014 by a group of hackers going by the name of "lelddos," which targeted web servers hosting Minecraft, a popular sandbox video game. The Minecraft server hosting market is very competitive, and a very successful server owner can make over \$50'000 per month from renting space for players to build their world. Lelddos has targeted these lucrative Minecraft servers for years with their attacks and extortion attempts.[34]

Mirai was also used by the creators Jha and Norman to commit advertising fraud, including click fraud as a less obvious way to generate passive income by abusing pay-per-click online advertising systems with the infected devices. In January 2017 Jha admitted that his group made a profit of approximately 200 bitcoins from these activities, which had a value of over \$180'000 at this time.[31]

Putman et al. (2017) concluded about Mirai that the average monthly revenue of a botmaster with regards to four case studies was around \$26'000. Subtracting the estimated costs of \$6000, such as money handler fees, web hosting and advertising results in a monthly profit of about \$20'000.[60]

IBM X-Force researchers found out that some modern variations of the Mirai botnet were also using the computation power of infected devices to mine cryptocurrency. Although this kind of usage is more noticeable in the IoT devices as it could damage by overheating the devices little central processing and graphical processing unit resources.[17]

11.4.2 The Dyn attack

On October 21, 2016, there was a series of DDoS attacks, which exceeded traffic of 1.3 terabytes per second on the popular Domain Service Provider (DNS) Dyn, which disrupted the name resolution for their clients. Services affected by the attack were high traffic sites, such as Amazon, Reddit, Spotify, Tumblr, and Twitter. During the time of the attack, internet users reported that they had trouble accessing the affected websites for several hours [51].

11.4.2.1 Timeline

According to the post-mortem report of Dyn, the first set of attacks started at 11:10 UTC. It was carried out with high-volume floods of TCP and UDP packets, both with destination port 53 from a vast number of source IP addresses. The Engineering and Network Operations teams of Dyn deployed additional mitigation tactics in addition to their automated response techniques, such as traffic-shaping incoming traffic, rebalancing of that traffic by manipulation of anycast policies, application of internal filtering, and deployment of scrubbing services. At 13:20 UTC, these mitigation tactics were entirely in deployment, and the services were restored [18].

Followed by that a more globally distributed set of attacks started at 15:50 UTC. Since this second set of attacks used the same protocol as the first attack, Dyn could mitigate the attack for the most part by 17:00 UTC [18].

In addition to the attacks on Dyn, there were simultaneous attacks on Xbox Live, Microsoft DNS infrastructure, Playstation, Nuclear Fallout game hosting servers, and Valve Steam servers. At 22:17 UTC, the final 10-hour long attack was issued on a set of Dyn and Playstation infrastructure [6]. This suggests that the attacker was targeting gaming infrastructure, and the attacks on Dyn were collateral damage. Cloudflare also came to this conclusion, as the other targets were all related to providing services for video games [11].

Most of the attacks used were SYN floods on the DNS port 53 with few GRE IP attacks. Additionally, the attacks on the gaming services, mostly Playstation Networks infrastructure were carried out via ACK, and GRE IP floods with VSE targeting the Valve Steam Servers.[6]

In the following days, numerous smaller TCP attacks occurred, which were mitigated by the efforts of Dyn and left their customers with no signs of disturbance [18]. Antonakakis et al. (2017) noted that there was a 71% intersection between the 107k IP addresses spotted that attacked Dyn and the Mirai scanning in their network telescope. They stated that while Mirai was involved in the attack, there may have been other hosts involved as well.[6]

11.4.2.2 Motivation

Jha nor his fellow co-authors admitted any involvement in the Dyn attack, and even to this day, it is unknown who the perpetrators behind the cyberattack on Dyn were. A group called "New World Hackers" claimed responsibility shortly after the attack. Other groups such as Anonymous and SpainSquad declared that they were the perpetrators behind the attack, but none of the claims were confirmed.[38]

The fact that the targeted services were associated with gaming and Mirai's past with targetting video game services suggests that there was no personal financial gain as a motive behind the Dyn attack, but served as an act of internet vandalism.

11.4.2.3 Economic Impact

In February 2017, data was published by security services company Bitsight, which showed that Dyn lost a considerable number of customers after the incident. Dyn lost the business of about 8% of their domains, about 14'500 shortly after the attack of their total 178'000, which were analyzed by Bitsight.[68]

The most significant loss was spotted in websites, which used Dyn and another DNS provider. About 8'000 domains were lost, 24% of this category of websites. Websites that used Dyn exclusively as a DNS provider mostly continued to use Dyn, only losing 4% or around 6'000 websites. Bitsight stated though that there is no indication if any of the lost customers went back to Dyn after the attack as it was only a snapshot of the monitored websites right after the attack.[68]

The experienced inaccessibility caused much damage to the involved websites. Estimating the loss for Amazon with their reported profit of almost \$136 billion in 2016 [41] would be around \$260'000 per minute. Assuming the downtime lasted 3 hours, the estimated damage adds up to 46 million during this timeframe. With other internet giants being affected, the damage dealt in the Dyn attack is in the region of multiple hundred million, if not billions.

The Dyn attack also served as a wake-up call for many manufacturers, highlighting the vulnerabilities of IoT devices. Matthew Prince, the CEO of Cloudflare, a popular DDoS protection service, stated that various IoT manufacturers asked the sales team for help to protect against attacks. Especially companies manufacturing IoT devices, where the malfunctioning of the devices could have serious implications, such as in the automobile or healthcare industries, were mainly concerned.[49]

11.4.3 Other major incidents

11.4.3.1 Krebs on Security

On September 20, 2016, a record DDoS attack (measured 620 Gbps) was launched against the popular security blog Krebs on Security ran by American journalist Brian Krebs. According to Akamai, the attack was almost bigger than twice than of previous publicly reported attacks back in 2016 [3]. The attacks were coming from 12'874 IPs, which had an intersection of 96% of observed ips from the Mirai study of Antonakis et al. (2017) [6]. Akamai successfully defended the initial attack, but due to its size, the blog was taken off their network as they were hosting Krebs on Security for free and were bearing the costs [54]. Krebs stated in a later article that his blog was knocked offline for almost four days.[33]

Belonging to the methods used in the incident were SYN, ACK, GET and POST (HTTP) floods as well as most of the attack being generic routing encapsulation traffic (GRE) data packets.[6]

11.4.3.2 Lonestar Cell

During the timeframe of October 2016 and February 2017, Liberia's market leader in providing internet and telecommunication services, Lonestar was continually the victim of attacks by the Mirai botnet. Reports stated that the internet quality of the entire country deteriorated due to these attacks [6]. In a blogpost of Krebs, security architect Kevin Beaumont stated that the attack reached 500 Gbps targeting Liberia's lone underseas large-transit cable, which served as a single point of failure.[27]

The attack vectors in the incident were ACK floods and SYN floods, which made up 87% of the attacks.[6]

According to the UK National Crime Agency, the attack by Kaye on Lonestar was very costly. Revenue loss of tens of millions of US dollars was reported as customers jumped ship after the incident. They also stated that the costs of the remedial measures taken to defend against the attack were approximate \$600'000.[13]

11.4.4 Future

Since Jha released the source code of Mirai back in 2017, there have been many copycats using modern variations of the Mirai botnet even to this day, increasing the number of C&C servers by a large amount. Due to this fact, it has been more challenging to track down the origins of attacks of Mirai-like botnets nowadays.

Mirai experienced a resurgence in 2019 with a focus-shift in targeting enterprise environments due to a more significant number of devices than in households. In an enterprise environment, many of the targets (more than 80%) are in the media/information services or insurance industries.[36]

In a Securelist report, it was estimated that about 21% of IoT device infections in 2018 were carried out by Mirai-clones.[19]

11.4.4.1 IoT Botnet Mitigation

To conclude this case study, we would like to reflect on some of the lessons we can learn from Mirai and its attacks and the implications for IoT security in order to prevent further, similar forms of botnets. One of the main reasons for IoT botnets is that many IoT devices are not held to the same security standards as desktop computers, for example. The lessons we already learned from desktop and web security can help us make the usage of IoT devices safer.

The first step is security hardening [6]. As mentioned before, Mirai attacks IoT devices by storing a repository of default usernames and passwords and using these in a brute force attack on identified victims. In order to limit the expansion of the Mirai botnet or other IoT botnets in general, IoT devices have to be held to the same security standards regarding non-default usernames and passwords as many web services do. Also, ports that are not being used have to be closed.

Another viable measure that focuses less on trying to prevent devices from being infected and more on reducing the number of infected devices is the implementation of automatic updates on IoT devices [6]. Figure 2.11 shows the number of Mirai infected devices with different hosts over time. As we can see, the decline of bots after an initial infection period in November 2016 does show an encouraging development. CWMP TCP/7547 peaked at about 600'000 and declined rapidly to about 6'700 in just one month. This decline is likely due to telecom operators patching the issue and releasing updates for all their devices. So, incorporating automatic updates into IoT devices, while taking into consideration possible resource constraints these devices may have, is critical in botnet mitigation.

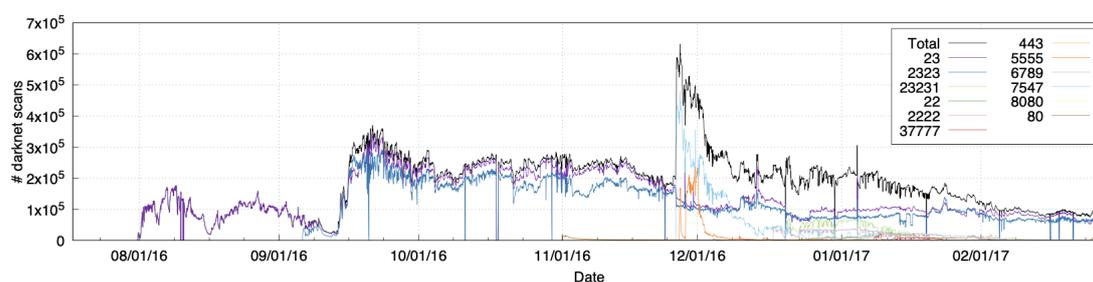


Figure 11.11: Temporal Mirai Infections [6]

The further focus has to be put on correctly identifying infected devices [6]. While studying the Mirai botnet and trying to determine the actual devices infected, M. Antonakakis et al. (2017) [6] found it challenging to attribute devices effectively. From a sample of 55.4 million scanning IP addresses, they managed to extract 1.8 million protocol banners, of which only about a third had any valuable identifying labels, meaning that just over 1% of the original sample could be identified. Without adequate measures for device attribution, it can be near impossible to determine who is responsible for future security problems and how to fix them.

Lastly, there are currently about 30 billion IoT devices connected to the internet, and that number is expected to rise above 75 billion by 2025 [64]. This extreme rise in the number of devices we use, likely means that in the future we will be left with many devices we do not use anymore and are therefore neglected by device manufacturers while they shift their focus on newer products [6]. Nonetheless, they are often still connected to the internet and are therefore susceptible to malware attacks. Further research on this topic could focus on the age of infected devices, however attributing an age to a specific device could prove very difficult.

These mitigation techniques drive up costs for botnet owners as it becomes more difficult to infect devices, and the botnets require a higher rate of reinfection. Higher costs, in turn, means that botmasters have lower incentives to pursue illegal activities.

11.5 Summary

Botnets pose a serious threat to our digital security. As part of this evolving economy with huge revenues and sophisticated actors, botnets play a vital role in the distribution of malware and the execution of cybercriminal acts. While our world and digital community have a shift over to renting and services, the same can be said about botnets. In contrast to a few years, where the users of a botnet were mostly their creators, today anyone can rent a botnet, to use it for a limited time. This renting out of the botnet constitutes one potential revenue stream for a botmaster. Botmasters may not always solely financially motivated but have a variety of potential revenue streams to monetize their botnets, with some of them being quite attractive sources of income. They can either use the botnet under their control to perform criminal activities that result in a financial gain or rent it out.

The most prominent way to monetize a botnet is the distributed denial of service attack, paired with an extortion claim. However, also installing adware on the victims' machines, stealing confidential information such as bank credentials or sending spam can be very lucrative. Although the revenue streams of a botnet can be significant, the damage done by them is also not neglectable. Numbers can only be estimated, damage of \$ 600 billion is very probable. Due to the high threat posed by botnets, not only for the economy as a whole but also for companies and individuals, many possible defense mechanisms are proposed. By fighting botmasters with technical solutions, one only generates an arms-race, between attackers and defenders.

Furthermore, many of the proposed solutions work in theory, but the practical implementation often needs the support of Internet service providers and other parties in the network that lack incentives to act upon the threats of botnets. However, since the creation and usage of botnets have evolved into a very sophisticated market with high revenues, the target of a defense system should be to minimize the possibility of monetizing a botnet. Proposed solutions like the Multihost Adware Revenue Killer try to accomplish precisely that. As shown with the case study of the Mirai botnet and its incidents, new technologies must not only be assessed by the advantages they bring to their users but also the possible dangers they constitute. With predictions of 75 billion IoT devices by

the year 2025, it is crucial to secure them properly in order to minimize the risk of being abused by criminals.

Bibliography

- [1] Accenture: *2017 Cost of Cyber Crime Study*; Report, 2017, https://www.accenture.com/_acnmedia/PDF-62/Accenture-2017CostCybercrime-US-FINAL.pdf?_ga=2.202996689.1777497212.1569861459-710890993.1569861459#zoom=50.
- [2] M. Abu Rajab, J. Zarfoss, F. Monrose, A. Terzis: *A multifaceted approach to understanding the botnet phenomenon*; Proceedings of the 6th ACM SIGCOMM on Internet measurement - IMC '06, Rio de Janeiro, Brazil, 2006, p. 41.
- [3] Akamai: *620+ Gbps Attack - Post Mortem*, <https://blogs.akamai.com/2016/10/620-gbps-attack-post-mortem.html>, October, 2016.
- [4] R. Anderson, C. Barton, R. Böhme, R. Clayton, M. JG Van Eeten, M. Levi, T. Moore, and S. Savage: *Measuring the cost of cybercrime*; In The economics of information security and privacy, Springer, Berlin, Heidelberg, Deutschland, October, 2013, pages 265–300, ISBN: 978-3-642-39498-0, DOI: 10.1007/978-3-642-39498-0_12.
- [5] R. Anderson, C. Barton, R. Böhme, R. Clayton, C. Gana, T. Grasso, M. Levi, T. Moore, M. Vasek: *Measuring the Changing Cost of Cybercrime*; Paper, 2019, https://weis2019.econinfosec.org/wp-content/uploads/sites/6/2019/05/WEIS_2019_paper_25.pdf.
- [6] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas and Y. Zhou: *Understanding the Mirai Botnet*; In 26th USENIX Security Symposium (USENIX Security 17) 2017 (pp. 1093-1110).
- [7] H. Asghari, M. JG van Eeten, and J. M. Bauer: *Economics of fighting botnets: Lessons from a decade of mitigation*; IEEE Security & Privacy, 13(5):16-23, September, 2015.
- [8] M. Bailey, E. Cooke, F. Jahanian, Y. Xu, and M. Karir: *A survey of botnet technology and defenses*; In CATCH'09. Cybersecurity Applications & Technology, Conference For Homeland Security, pages 299–304, Waltham, MA, USA, March 2009. IEEE. DOI:10.1109/CATCH.2009.40.
- [9] G. Bottazzi, G. Me: *The Botnet Revenue Model*; Proceedings of the 7th International Conference on Security of Information and Networks - SIN '14, Glasgow, Scotland, UK, 2014, pp. 459–465.
- [10] CenturyLink: *CentruryLink 2019 Threat Report*; Report, 2019, <https://www.centurylink.com/asset/business/enterprise/report/2019-threat-research-report.pdf>.

- [11] CloudFlare: *Inside the infamous Mirai IoT Botnet: A Retrospective Analysis*; Blog post, 2017, <https://blog.cloudflare.com/inside-mirai-the-infamous-iot-botnet-a-retrospective-analysis/>.
- [12] C. Cimpanu: *After Dodging Prison in Germany, Mirai Hacker "BestBuy" Charged in the UK*; August, 2017, <https://www.bleepingcomputer.com/news/security/after-dodging-prison-in-germany-mirai-hacker-bestbuy-charged-in-the-uk/>.
- [13] C. Cimpanu: *Liberian ISP sues rival for hiring hacker to attack its network*; January, 2019, <https://www.zdnet.com/article/liberian-isp-sues-rival-for-hiring-hacker-to-attack-its-network/>.
- [14] P. Collinson: *Alleged mastermind behind bank cyberattacks extradited to UK*; August, 2017, <https://www.theguardian.com/uk-news/2017/aug/30/alleged-mastermind-daniel-kaye-lloyds-bank-cyber-attacks-extradited-uk>.
- [15] E. Cooke, F. Jahanian, D. McPherson: *The Zombie Roundup: Understanding, Detecting, and Disrupting Botnets*; Jul-2005, https://www.usenix.org/legacy/event/sruti05/tech/full_papers/cooke/cooke_html/ [Accessed: 30-Oct-2019].
- [16] N. Daswani and M. Stoppelman: *The anatomy of a clickbot*; In Proceedings of the first conference on First Workshop on Hot Topics in Understanding Botnets, Hot-Bot'07, page 11, Cambridge, MA, USA, April, 2007, URL: https://www.usenix.org/legacy/event/hotbots07/tech/full_papers/daswani/daswani.pdf.
- [17] C. DeBeck, J. Chung and D. McMillen: *I Can't Believe Mirais: Tracking the Infamous IoT Malware*; July, 2019, <https://securityintelligence.com/posts/i-cant-believe-mirais-tracking-the-infamous-iot-malware-2/>.
- [18] Dyn: *Dyn Analysis Summary Of Friday October 21 Attack*; October, 2016, <https://dyn.com/blog/dyn-analysis-summary-of-friday-october-21-attack/>.
- [19] A. Eremin: *Bots and botnets in 2018*; March, 2018, <https://securelist.com/bots-and-botnets-in-2018/90091/>.
- [20] R. Ford, S. Gordon: *Cent, Five Cent, Ten Cent, Dollar: Hitting Botnets Where It Really Hurts*; Proceedings of the 2006 Workshop on New Security Paradigms, New York, NY, USA, 2007, pp. 3–10.
- [21] G. Gu, J. Zhang, W. Lee: *BotSniffer: Detecting Botnet Command and Control Channels in Network Traffic*; Proceedings of the 15th Annual Network and Distributed System Security Symposium, 2008, <https://corescholar.libraries.wright.edu/cse/7>.
- [22] C. Jin, H. Wang, K. G. Shin: *Hop-Count Filtering: An Effective Defense Against Spoofed DDoS Traffic*; CCS'03, October 27-31, 2003, Washington, DC, USA, p. 12.
- [23] E. Kalige, D. Burkey: *A case study of eurograbber: How 36 million euros was stolen via malware*; Versafe (White paper) 35, December, 2012, URL: <https://www.checkpoint.com/downloads/product-related/whitepapers/eurograbber-malware-bank-customers-millions-stolen.pdf>.
- [24] C. Koliass, G. Kambourakis, A. Stavrou and J. Voas, *"DDoS in the IoT: Mirai and Other Botnets"*; in Computer, vol. 50, no. 7, pp. 80-84, 2017, DOI: 10.1109/MC.2017.201.

- [25] February 28th DDoS Incident Report; <https://github.blog/2018-03-01-ddos-incident-report/>, November, 2019.
- [26] B. Krebs: *Court Hands Down Hard Jail Time For DDoS*; <https://krebsonsecurity.com/2019/01/courts-hand-down-hard-jail-time-for-ddos/>, January, 2019.
- [27] B. Krebs: *Did the Mirai Botnet Really Take Liberia Offline?*; <https://krebsonsecurity.com/2016/11/did-the-mirai-botnet-really-take-liberia-offline/>, November, 2016.
- [28] B. Krebs: *KrebsOnSecurity Hit With Record DDoS*; <https://krebsonsecurity.com/2016/09/krebsonsecurity-hit-with-record-ddos/>, September, 2016.
- [29] B. Krebs: *Mirai Botnet Authors Avoid Jailtime*; <https://krebsonsecurity.com/2018/09/mirai-botnet-authors-avoid-jail-time/>, September, 2018.
- [30] B. Krebs: *Mirai Co-Author Gets 6 Months Confinement, \$8.6M in Fines for Rutgers Attacks*; <https://krebsonsecurity.com/2018/10/mirai-co-author-gets-6-months-confinement-8-6m-in-fines-for-rutgers-attacks/>, October, 2018.
- [31] B. Krebs: *Mirai IoT Botnet Co-Authors Plead Guilty*; <https://krebsonsecurity.com/2017/12/mirai-iot-botnet-co-authors-plead-guilty/>, December, 2017.
- [32] B. Krebs: *Most Malware Tied to 'Pay-Per-Install' Market*; In MIT Technology Review, June, 2011, URL: <https://www.technologyreview.com/s/424241/most-malware-tied-to-pay-per-install-market/>.
- [33] B. Krebs: *'Satori' IoT Botnet Operator Pleads Guilty*; <https://krebsonsecurity.com/2019/09/satori-iot-botnet-operator-pleads-guilty/>, November, 2019.
- [34] B. Krebs: *Who is Anna-Senpai, the Mirai Worm Author?*; <https://krebsonsecurity.com/2017/01/who-is-anna-senpai-the-mirai-worm-author/>, January, 2017.
- [35] O. Kupreev, E. Badovskaya and A. Gutnikov: *DDoS attacks in Q4 2018*; February, 2019, <https://securelist.com/ddos-attacks-in-q4-2018/89565/>.
- [36] R. Lemos: *Mirai Groups Target Business IoT Devices*; July, 2019, <https://www.darkreading.com/mirai-groups-target-business-iot-devices/d/d-id/1335308>.
- [37] A. Lerner: *The Cost of Downtime*; <https://blogs.gartner.com/andrew-lerner/2014/07/16/the-cost-of-downtime/>, July, 2016.
- [38] D. Lewis: *The DDoS Attack Against Dyn One Year Later*; <https://www.forbes.com/sites/davelewis/2017/10/23/the-ddos-attack-against-dyn-one-year-later/>, October, 2017.
- [39] Z. Li, Q. Liao, A. Striegel: *Botnet Economics: Uncertainty Matters*; Managing Information Risk and the Economics of Security, M. E. Johnson, Ed. Boston, MA: Springer US, 2009, pp. 245–267.
- [40] B. Lovelace Jr. and A. José Vielma: *Friday's third cyberattack on Dyn 'has been resolved,' company says*; October, 2016, <https://www.cnn.com/2016/10/21/major-websites-across-east-coast-knocked-out-in-apparent-ddos-attack.html>.

- [41] Macrotrends: *Amazon Revenue 2006-2019 | AMZN* <https://www.macrotrends.net/stocks/charts/AMZN/amazon/revenue> [Accessed: 01-Nov-2019].
- [42] MalwareMustDie: <https://www.malwaremustdie.org>, November, 2019.
- [43] MalwareTech: *Mapping Mirai: A Botnet Case Study*, <https://www.malwaretech.com/2016/10/mapping-mirai-a-botnet-case-study.html>, October, 2016.
- [44] McAfee: *Economic Impact of Cybercrime - No Slowing Down*; Report, February, 2018, <https://www.mcafee.com/enterprise/en-us/assets/reports/restricted/rp-economic-impact-cybercrime.pdf>.
- [45] R. Mahajan, S. M. Bellovin, S. Floyd, J. Ioannidis, V. Paxson, S. Shenker: *Controlling high bandwidth aggregates in the network*; SIGCOMM Comput. Commun. Rev., vol. 32, no. 3, pp. 62–73, Jul. 2002.
- [46] Malwarebytes; <https://www.malwarebytes.com/adware/>, November, 2019.
- [47] M. McGuire: *Into the Web of Profit - Understanding the Growth of the Cyber-crime Economy*; April, 2018, https://www.bromium.com/wp-content/uploads/2018/05/Into-the-Web-of-Profit_Bromium.pdf.
- [48] D. McMillen: *Mirai IoT Botnet: Mining for Bitcoins?*; April, 2017, <https://securityintelligence.com/mirai-iot-botnet-mining-for-bitcoins/>.
- [49] S. Melendez *Cloudflare Shores Up Defenses For Internet Of (Easily Hackable) Things*; April, 2017, <https://www.fastcompany.com/40412529/internet-of-things-security-cloudflare>.
- [50] J. Mirkovic, P. Reiher,: *A taxonomy of DDoS attack and DDoS defense mechanisms*; SIGCOMM Comput. Commun. Rev., vol. 34, no. 2, p. 39, Apr. 2004.
- [51] S. Moss: *Major DDoS attack on Dyn disrupts AWS, Twitter, Spotify and more*; <https://www.datacenterdynamics.com/news/major-ddos-attack-on-dyn-disrupts-aws-twitter-spotify-and-more/>, October, 2016.
- [52] Y. Namestnikov: *The economics of botnets*; Analysis on Viruslist.com, Kaspersky Lab, July, 2009, URL: https://media.kasperskycontenthub.com/wp-content/uploads/sites/43/2009/07/01121538/ynam_botnets_0907_en.pdf.
- [53] Netscout: *Netscout threat intelligence report: powered by Atlas, Findings from 1H 2019; Report, 2019*, <https://www.netscout.com/threatreport>.
- [54] C. Osborne: *Krebs on Security booted off Akamai network after DDoS attack proves pricey*, September, 2016, <https://www.zdnet.com/article/krebs-on-security-booted-off-akamai-network-after-ddos-attack-proves-pricey/>.
- [55] P. Paganini: *Hackers offer a huge Mirai botnet as a DDoS-for-hire-service*; November, 2016, <https://securityaffairs.co/wordpress/53824/cyber-crime/mirai-botnet-ddos-for-hire.html>.
- [56] Kihong Park, Heejo Lee: *On the effectiveness of probabilistic packet marking for IP traceback under denial of service attack*; Proceedings IEEE INFOCOM 2001. Conference on Computer Communications. Twentieth Annual Joint Conference of the IEEE Computer and Communications Society (Cat. No.01CH37213), 2001, vol. 1, pp. 338–347 vol.1.

- [57] D. Patterson: *A simple Way to Estimate the Cost of Downtime*; Paper in Proceedings of LISA '02: Sixteenth Systems Administration Conference, Berkeley, November, 2002, pp. 185 - 188, https://www.usenix.org/legacy/event/lisa02/tech/full_papers/patterson/patterson_html/.
- [58] J. Pijpker, H. Vranken: *The Role of Internet Service Providers in Botnet Mitigation*; 2016 European Intelligence and Security Informatics Conference (EISIC), IEEE, 2016.
- [59] Ponemon Institute: *Cost of Data Center Outages*; Report, January, 2016, https://www.vertiv.com/globalassets/documents/reports/2016-cost-of-data-center-outages-11-11_51190_1.pdf.
- [60] C.G.J. Putman: *Business model of botnets*; 27th Twente Student Conference on IT, July, 2017. URL: <http://referaat.cs.utwente.nl/conference/27/paper/7637/business-model-of-botnets.pdf>.
- [61] J. H. Rao, D. H. Reiley: *The Economics of Spam*; Journal of Economic Perspectives, 26(3), pages 87-110, September, 2012
- [62] Alex C. Snoeren, Craig Partridge, Luis A. Sanchez, Christine E. Jones, Fabrice Tchakountio, Stephen T. Kent, W. Timothy Strayer: *Hash-Based IP Traceback*; SIGCOMM'01, August 27-31, 2001, San Diego, California, USA.
- [63] Spam Data; <http://www.barracudacentral.org/data/spam>, October, 2019
- [64] Statista; <https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/>, August, 2019.
- [65] Symantec: *Internet Security Threat Report*; Report, Volume 24, February, 2019, <https://www.symantec.com/content/dam/symantec/docs/reports/istr-24-2019-en.pdf>.
- [66] Rob Thomas, Jerry Martin. : *The Underground Economy: Priceless*; USENIX ;login, 31(6), 2006, pp. 7-16.
- [67] United States. Department of Justice: *Justice Department Announces Charges And Guilty Pleas In Three Computer Crime Cases Involving Significant Cyber Attacks*; <https://www.justice.gov/usao-nj/pr/justice-department-announces-charges-and-guilty-pleas-three-computer-crime-cases>, December, 2017.
- [68] S. Varghese: *DDoS Attack on Dyn costly for company claim*; February, 2017, <https://www.itwire.com/security/76717-ddos-attack-on-dyn-costly-for-company-claim.html>.
- [69] M. J. G. van Eeten, J. M. Bauer: *Economics of Malware: Security Decisions, Incentives and Externalities*; OECD Science, Technology and Industry Working Papers 2008/01, May 2008.
- [70] P. Wang, S. Sparks, C. C. Zou: *An Advanced Hybrid Peer-to-Peer Botnet*; IEEE Transactions on Dependable and Secure Computing, vol. 7, no. 2, pp. 113-127, Apr. 2010.
- [71] R. Winward: *IoT Attack Handbook: A Field Guide To Understanding IoT Attacks from the Mirai Botnet to its Modern Variants*, https://www.datacom.cz/userfiles/miraihandbookebook_final.pdf [Accessed: 10-Nov-2019].

- [72] GDP; https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=US&most_recent_value_desc=false , October, 2019.
- [73] D. K. Y. Yau, J. C. S. Lui, Feng Liang, Yeung Yam: *Defending against distributed denial-of-service attacks with max-min fair server-centric router throttles*; IEEE/ACM Transactions on Networking, vol. 13, no. 1, pp. 29–42, Feb. 2005.
- [74] J. Xu, W. Lee: *Sustaining availability of Web services under distributed denial of service attacks*; IEEE Transactions on Computers, vol. 52, no. 2, February, 2003, pp. 195–208.
- [75] H. R. Zeidanloo, A. A. Manaf: *Botnet Command and Control Mechanisms*; 2009 Second International Conference on Computer and Electrical Engineering, 2009, vol. 1, pp. 564–568.

Chapter 12

The Renaissance of DAOs: Recurring Trends Toward Decentralized Autonomous Organizations

Francesca Monzeglio, Raphael Beckmann, Benjamin Jeffrey and Roberto Baumann

Decentralized Autonomous Organizations (DAO) first appeared in 2014 and shortly disappeared again in 2016 after the crash of “TheDAO”, a project that was designed to offer a democratic financial institution in which investors had decisional power through voting. However, in recent years, efforts to create and implement DAOs are reappearing. Generally speaking, the goal and purposes of DAOs are comparable to those of their real-world counterpart, but they are pseudonymously governed over the Internet. First, this paper introduces the concept of a DAO based on the Ethereum public blockchain, followed by an analysis of “TheDAO”, detailing its purpose and potential reasons for its demise. Then, more recent DAO projects will be discussed, comparing them on a technical, practical and governmental level.

Contents

12.1 Introduction	125
12.2 Blockchain	125
12.2.1 Ethereum	126
12.2.2 Smart Contracts	128
12.3 The Concept of Decentralized Autonomous Organizations	131
12.3.1 Governance Definition	131
12.3.2 Voting Mechanism	132
12.3.3 Evaluation and Security Issues	132
12.4 The First Rise and Fall: TheDAO	133
12.4.1 The Inspiration for TheDAO	134
12.4.2 TheDAO Creation	134
12.4.3 The Underlying Vulnerabilities	134
12.4.4 The Attack	134
12.4.5 The Solution by the Community: the Ethereum Hard Fork	137
12.4.6 Consequences of the Attack	138
12.5 Current DAO Projects	138
12.5.1 Aragon	139
12.5.2 Dai Stablecoin - MakerDAO	141
12.5.3 Moloch DAO	146
12.6 Comparison	148
12.7 Conclusion	149

12.1 Introduction

The rising popularity of blockchain technology is enabling blockchain-based platforms to become more widespread. A key innovation compared to previous public blockchains, is the capability of Ethereum to distribute the execution of arbitrary pieces of bytecode on-chain. Usually, this piece of bytecode is referred to as smart contracts, which for instance, may encode the terms of an agreement and thus facilitate and automate the way of doing business [19]. The governance concept of a traditional organization can be digitally democratized, and under certain perspective improved, by leveraging a public, permissionless blockchain as the underlying layer, in addition to using smart contracts to digitally enforce the protocols of the company [24]. This concept is known as a *Decentralized Autonomous Organization (DAO)*. Many potential benefits are associated with this approach, but the technical deployment is challenging and is prone to security vulnerabilities. A negative real-world example of a DAO is the infamous hack of TheDAO in 2016. The project was initially very successful, raking in 150 million US\$ in the crowdfunding period, only to completely fail shortly after that, when a recursive call exploit in the contract to create a child DAO according to the *split procedure* led to the loss of approximately 3.6 million Ether [30].

Three years after the hack of TheDAO, new efforts towards decentralized organizations are made. Multiple Ethereum-based projects are in active development and in pursuit of refining DAO concepts. This paper presents an overview and comparison of current DAO projects, with their on-chain governance models. These are preceded by a brief technical introduction on blockchain, smart contracts and the general concept of a DAO.

12.2 Blockchain

Traditionally, a blockchain is defined as a distributed digital ledger composed of *nodes*. Generally all nodes store the most updated copy of that ledger, i.e. of the entire blockchain [3]. The underlying technology consists of a peer-to-peer network that eliminates the need of a central authority responsible for decision making. A decentralized consensus mechanism is implemented instead, to ensure agreement on the current state of the blockchain at any given moment. Nodes update the blockchain by either contributing directly to it or listening to other nodes updating it [3].

Each update to the ledger is called a *block*, which is added to the chain and mainly consists of three parts [4]:

1. The block stores the reference to its predecessor in the chain.
2. A list of transactions executed, starting from the previous block is stored. This also includes a timestamp, the digital *signature* of the participants and other details for each transaction.
3. A hash code and the timestamp, uniquely identifying the block, are also stored.

The security of a blockchain is ensured by a distributed consensus algorithm executed among the participating nodes. The most popular consensus algorithm is called *Proof of Work* [5]. Periodically, the exact time period varies with various factors, a *miner* that “proved” to have done the “work” by trying to compute a specific complex hash, gets to add one new block [4]. In other words, the miner publishes the next update to the blockchain on the network. A transaction broadcasted to the entire network is verified before being added to a block. The more blocks get added to the blockchain after a transaction, the more secure that transaction gets. This is because to change the block

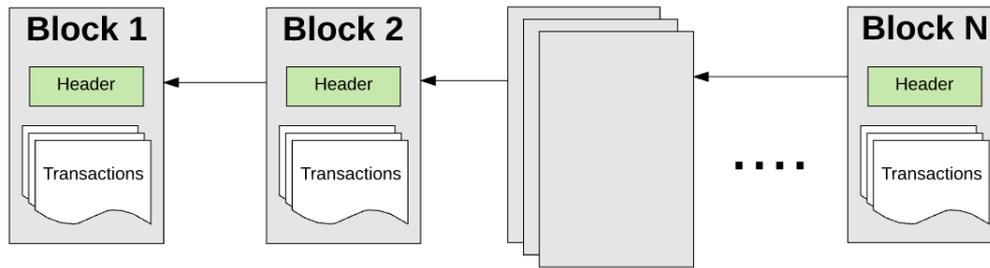


Figure 12.1: A blockchain consists of *blocks* securely linked to its predecessor [13]

that held the transaction, the proof of work for all subsequent blocks would have to be redone as well [4].

Given its above-mentioned qualities, the blockchain has gained popularity as tamper-proof decentralized source of trust for all parties involved [4].

12.2.1 Ethereum

Ethereum is a globally accessible, open-source, blockchain-based platform for decentralized applications launched in 2015 by the Ethereum foundation [7]. The native cryptocurrency of Ethereum is called Ether (ETH), which on October 19th, 2019 was worth 174.38 US\$ [24]. Unlike other blockchains, Ethereum is programmable, meaning that developers can use it to build applications, exploiting the benefit of cryptocurrency and blockchain technology and expecting them to always run as programmed [7]. These applications can be decentralized, namely not controlled by any single entity or person, and include examples like cryptocurrency wallets, financial applications, decentralized markets, games and more [7]. Some of such applications are represented in Figure 12.2.



Figure 12.2: Ethereum serve as a blockchain-based platform for many applications built on top of it such as Aragon, Golem, Gnosis, MakerDAO or Moloch DAO. Own image based on [16]

Therefore, many centralized systems in use today could be built on Ethereum in a decentralized way, thus cutting out the middlemen, which implies lower costs for the end user, and eliminating single points of failure or control [11]. Furthermore, removing third

parties avoids storing sensitive data on central servers of big companies, that are consequently easier to be hacked, sold and disrupted. This allows any user from anywhere to participate on any decentralized application that also cannot be removed or censored [10]. The vision of Ethereum’s development team is indeed to achieve “an unstoppable censorship-resistant self-sustaining decentralised world computer that can perform calculations, store data, and allow communications” [12].

In Ethereum, the concept of a blockchain as a *distributed ledger*, that validates, stores, and replicates transaction data on many computers around the world, has been extended to a distributed data storage, including computations that are replicated and equivalently processed on all computers in the network, without a central coordinator [12]. One of the major applications of Ethereum is actually to run smart contracts, namely digital contracts which consist of a piece of code running on the *Ethereum Virtual Machine* (EVM), that “can control valuable things like ETH or other digital assets” [8]. The EVM automatically runs on the computer of every participant on the Ethereum network so that all the nodes behave equivalently following the peer-to-peer network model [12].

It is everyday understanding that a chain is a linear structure. In the particular case of blockchain this means that every new node added to the chain must follow a set of rules before it can be accept as valid by all the participants on the network. It can happen however, that a subset of miners would like to modify the initial condition and adapt the blockchain. This can be done by creating a copy of the blockchain and starting a new path diverging from the original chain; in technical terms a *fork* has occurred [13]. The “original” chain is then distinguished by Ethereum’s use of the *Ghost protocol*, which simply determines which path has had the most computation done upon it, i.e. the block that needed the highest mining effort to be reached, represents the leaf of the *canonical* version of the blockchain [13].

12.2.1.1 Soft-fork vs. Hard-fork

As the previous section has presented, a fork is a copy of the current state of the blockchain. But since it determines some change to the rules of consensus [14], it can be viewed as a protocol upgrade mechanism [15]. Figure 12.3 offers a visual reference when forking a blockchain, that can be implemented with two methods: soft fork and hard fork. As the later sections will discuss, the difference between these two approaches will be relevant to fully understand the discussion about the possible solutions to handle the failure of TheDAO, a famous crowdfunding project built on Ethereum.

A *soft fork* must be accepted by the majority of the miners before it can be adopted and it strictly reduces the set of transactions that are valid, either by introducing a new rule or by making the existing rules more restrictive [15]. Since the not upgraded nodes will still recognize new blocks¹, a soft fork is said to be *backwards compatible*. However, if the upgrade is not accomplished, some misinterpretation of new transactions may occur [14]. A *hard fork* removes or relaxes consensus rules, allowing previously invalid transactions and blocks to become valid [15]. Once a hard fork has been activated and new blocks have been added according to the new relaxed rules, all not upgraded nodes will automatically reject these new blocks, meaning that a participant on the network must upgrade in order to stay on the hard-forked chain [14]. If some miners and nodes decide to stay on the old software, while others switch to the new, a chain split occurs [14].

Both methods have both advantages and disadvantages. Soft forks are less likely to result in a chain split and are usually considered more convenient for the users, since they are not forced to upgrade their software in order to stay on the chain. On the other hand, hard forks allow the developers much more flexibility as they do not have to ensure the coexistence of old and new rules. Nonetheless, hard forks are considered “coercive”, since

¹The new, more restrictive set of rules is a subset of the old set of rules.

a user preferring the old rules is forced to adapt to the new rules to stay on the network despite personal disapproval [15].

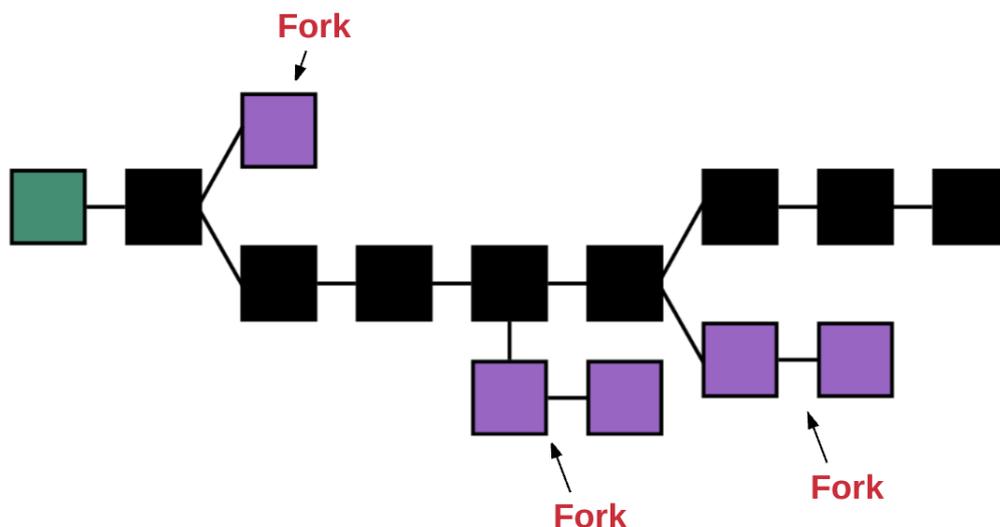


Figure 12.3: It is possible to change the rules of consensus of a blockchain by creating a copy of the current version through a fork. The canonical blockchain is represented by the sequence of black blocks [13]

12.2.2 Smart Contracts

As seen in the previous section, the Ethereum blockchain allows the possibility to store and distributively run small computer programs on the network. These pieces of code are usually referred to as *smart contracts*, as their function is to automatically facilitate, execute and enforce the negotiation or performance of an agreement between two anonymous parties, willing to trade or do business with each other, without the need for a middleman [19]. Smart contracts implement a set of conditions, agreed by the users, that when met, will result in the automatic realization of the terms of the agreement [17]. By way of comparison, a blockchain provides trustworthy storage, while smart contracts deliver distributed trustworthy calculations [18].

On another perspective, they can be viewed as highly programmable digital money since they allow an automatic transfer of money from one person to another under certain conditions and with no third parties involved [11].

12.2.2.1 Development of Smart Contracts

In the Ethereum context, smart contracts are immutable² and run deterministically³ as part of the Ethereum network protocol [20]. Smart contracts are typically written in high-level language such as the ones mentioned below.

Solidity. A procedural programming language syntactically similar to Javascript, which is currently the most popular and functional language for Ethereum smart contracts [12].

Serpent. Another procedural language but similar to Python, that was popular in the early history of Ethereum [12].

LLL (Lisp Like Language). As deducible from the name itself, this is a Lisp-like syntax programming language. It has been the first high-level language to write Ethereum smart contracts; however, it is rarely used today [20].

²Once deployed, the only way to change the code of a smart contract is to deploy another instance [20].

³Under the same initial environmental conditions, the outcome of a smart contract is the same for everyone running it [20].



Figure 12.4: Smart contracts are self-executing contractual agreements running on Ethereum blockchain [23]

Vyper. More recently developed language, similar to Serpent, but intended to get closer to a pure-functional Python-like language [20].

12.2.2.2 Use of Smart Contracts

Once they are written using a high-level language, smart contracts must first be compiled to low-level bytecode running in the EVM before they can be deployed on the Ethereum platform [20]. At this point, the smart contracts are encrypted and sent to all the miners on the network as part of a transaction. The “winning” miner, i.e. the node that has mined the block, publishes and distributes the block throughout the rest of the network. The nodes in the network then individually validate the new block before adding it to their own blockchains, thus updating the current state of the Ethereum blockchain [12]. Contracts never run in the background, and even though a contract can call another contract, which may call another one, thus forming a chain of called-contracts by repeating that pattern, the execution of the first contract in such chain is always ultimately triggered by a special *contract creation* transaction [20].

Deployed smart contracts can only run if they are funded with ETH [12]. In greater detail, when a user⁴ wants to run a particular contract, a transaction containing a payment to the “contract” itself is performed and if needed, other information useful to the execution of the contract can be supplied [12]. This amount of ETH is used to pay the computation on the EVM, which is given as a reward to the miner adding the new block [13].

The reward results from the *Gas Limit* multiplied by the *Gas Price* and corresponds to the maximum transaction fee the user is willing to spend on the contract [13]. *Gas* is the unity used to measure the computational effort to execute a smart contract: each contract requires a different amount of resources, thus the amount of gas needed to perform the transaction is proportional to its complexity, i.e. the number and type of computational steps, the memory usage, etc. [12]. The Gas Limit is decided by the user each time a specific contract is requested and usually depends on the kind of “performance” the user needs. For instance, if the user wants the miners to give high priority to the requested contract and having it executed as soon as possible, then a high Gas Limit will be set [12]. As depicted in Figure 12.5, a transaction is fully executed only if the provided amount of gas is sufficient. In case of success, any remaining gas is refunded to the sender, but if the contract exhausts the gas, the transaction is considered invalid and the state is reverted [13].

⁴The user of a smart contract is identified as the *sender*, since the contract is triggered by a message sent by the user requesting its execution.

Given their full functionality⁵, smart contracts are often described as *Turing complete* [12]. As a consequence, they are susceptible to the halting problem⁶, but the above-mentioned fee mechanism also contributes in protecting the network against infinite loop transactions [13]. Nevertheless, in Section 12.4 it will be shown that this safety measure alone is not always enough.

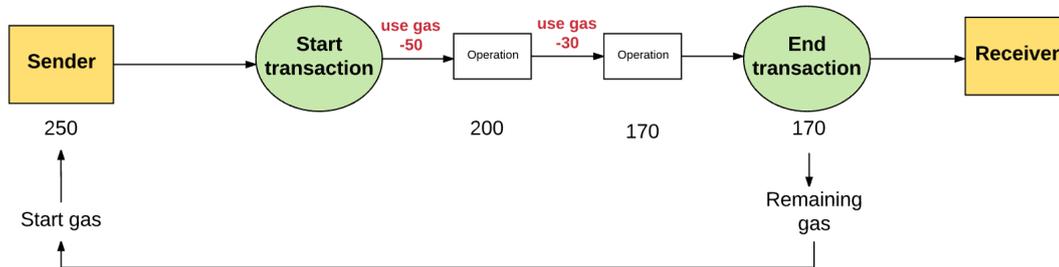


Figure 12.5: The sender must provide enough gas to have the transaction in the smart contract executed and What remains after the transaction is refunded to the sender [13]

12.2.2.3 Attacks and Vulnerabilities

Even though decentralization/replication on a distributed network prevents single party manipulation [20], smart contracts are still subject to vulnerabilities of various kinds, some of which are briefly reported in this section [21].

Re-Entrancy. This vulnerability was used to exploit TheDAO, more details about it will be provided in Section 12.4. It occurs when a contract tries to send ETH before having updated its internal state. Otherwise stated, a call to another contract invoking a re-entrant callback to the calling contract is necessary, so that the function to request ETH is recursively called.

Unhandled Exceptions. Some low-level operations may not throw an exception on failure, but rather return a simple boolean variable, that if not checked, would not stop the execution of the contract and will eventually produce inconsistencies.

Locked Ether. Funds may be locked in a contract, for example when a contract relies on an external one, that has been deleted from the blockchain.

Transaction Order Dependency. Since a block on Ethereum can contain multiple transactions, the state of a contract can be updated several times. An attacker could exploit the effect of two transactions calling the same smart contract within the same block, whose final outcome depends on the execution order of these transactions. To be successful this vulnerability also needs contract's storage manipulation.

Integer Overflow. This common bug has serious consequences in the context of Ethereum smart contracts. An infinite loop caused by a counter overflow will completely freeze the funds of a contract.

In conclusion, it is worth to point out that bugs in smart contracts cost money and therefore it is indispensable to avoid undesired side effects in the logic during the coding phase [20].

⁵They can perform any computation possible with any programming language [12].

⁶Given a running program it is not possible to determine if it will eventually stop or run infinitely [13].

12.3 The Concept of Decentralized Autonomous Organizations

A *Decentralized Autonomous Organization* takes the same idea of a canonical human company but applies the concept of decentralization to it [25]. The main difference is that a DAO autonomously exists on the internet [25] and operates by digitally enforcing its rules through a set of smart contracts [24]. These encode financial transactions and protocols on a blockchain and function as a governance mechanism [33], thus “effectively removing the need for a central governing authority” [30]. The comparison of a DAO structure with a traditional one can be seen in Figure 12.6.

To a certain extent a DAO can be seen as an entity making decisions by itself [25] and more technically as a “cryptographically guaranteed democracy, where stakeholders can vote on adding new rules, changing the rules, or ousting a member” just to mention a few examples [24]. Therefore, this mechanism represents a good endorsement to security since a single person is not able to modify a deployed smart contract and consequently to change the DAO rules. Unfortunately, the immutability of blockchains can also be a disadvantage, since developers would not be able to easily fix a bug on a running DAO in case it should be exploited [24].

Despite automation, the human component still plays a central role within the company. Sometimes they are actually hired to supply for automation limitations and usually interaction in person is not required, since it is performed according to the protocols encoded on the blockchain [25].

Smart contracts are programmed to perform various types of tasks, not only related to assets, such as handling the funds of the company or responding to what has been agreed by a vote [24]. Even though some additional legal support might be needed, smart contracts provide many benefits when used to implement the (protocol) rules of a company. Traditional contracts usually present issues like being under-specified and lacking in details about how a transaction is actually processed, two aspects that can consequently lead to frictions between the contracting parties. Smart contracts on the other hand, provide a solution to these issues by clearly stating what happens when certain conditions are met. Hence, the demanding task of finding consensus among the parties can be simplified when terms and conditions are precisely coded into a smart contract [22].

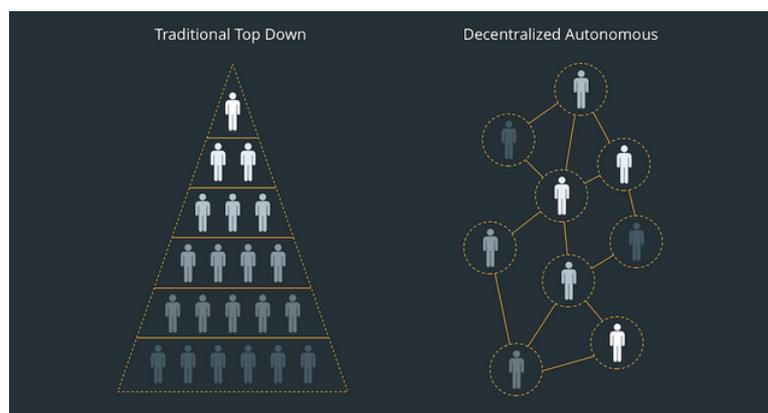


Figure 12.6: In opposition to the traditional top down organization, a DAO presents a decentralized structure without a central governing authority [40]

12.3.1 Governance Definition

The innovative aspect of a DAO principally emerges from the application of a governance mechanism to a new type of organization [26]. According to the Cambridge Dictionary,

governance can be generally defined as “the way that organizations or countries are managed at the highest level, and the systems for doing this” [27]. In the context of DAOs, governance is principally achieved by means of smart contracts, that are defined by a group of people to run the organization [34]. Typically a DAO establishment begins with an initial funding period called Initial Coin Offering (ICO), where stakeholders add the necessary funds to the DAO by buying *ownership* tokens [34]. These tokens give their owners exclusive rights to vote on proposals about the company and furthermore, they differ from traditional equity shares as they do not represent ownership of the DAO itself, which is a property of nobody. Once the ICO is over, the DAO is operating and proposals to it can be submitted. The latter usually regard investment of the available funds [34]. Historically, Bitcoin is the very first example of a DAO, where the consensus among its core team and its mining network produces the governance. Since then, all other DAOs have been deployed on the Ethereum platform [34].

12.3.2 Voting Mechanism

In order to be put in place, governance not only requires written rules defined by smart contracts, but also a mechanism to make decisions. In absence of a central authority, this responsibility has been spread among the members of the company, who express their opinion by means of voting. The voting mechanism works similarly to the one applied in a democratic political system: the majority decides. But who gets the right to vote? And how do we avoid low participation or collusion? As an example on how to address the above-mentioned issues, we can take a look at the voting mechanism adopted by Dash, a payments-focused cryptocurrency managed as a DAO, whose governance mechanism can be seen in Figure 12.7. This voting mechanism is based on *masternode operators*, that cast votes recorded on the blockchain. To be eligible as a masternode, one must prove the ownership of 1000 dash coins. This is to ensure that masternodes will not vote maliciously in order to preserve their own interests, which also corresponds to the success of Dash [5]. Their participation on the vote is also promoted by the funding method adopted by Dash. The mining reward for a block is split into three parts: 45% each is taken by the masternodes and the miner, while the remaining 10% serves as treasury, going into the decentralized governance budget [28]. This budget is used to create value within Dash by supporting the development and expansion of the ecosystem or by promoting the coin, in agreement with the vote of masternodes in the various proposals, that are posted on a public portal [28]. In exchange for voting rights, masternodes must perform some basic tasks, other than voting on the proposals: They must keep an updated copy of the blockchain at all times and they must provide the network with some other functionalities distinctive to Dash [5]. This model additionally guarantees the survival of the network by itself through a kind of “recycling” mechanism. If early masternodes quit Dash and sell their coins, the new owner can set up a masternode and replace the old voter by continuing its duties in the decision-making process on budget and projects [28].

12.3.3 Evaluation and Security Issues

The core goal of DAOs is to remove the central authority from the management scheme of a business corporation, thus reducing the costs and providing more control directly to the investors [30]. Besides democratising the traditional idea of company governance, one positive aspect conferred to DAOs is transparency [26]. As demonstrated in Section 12.2, all changes and transactions are publicly available to everyone having access to a copy of the blockchain, but the identity and the personal data can still be kept private by means of encryption [26].

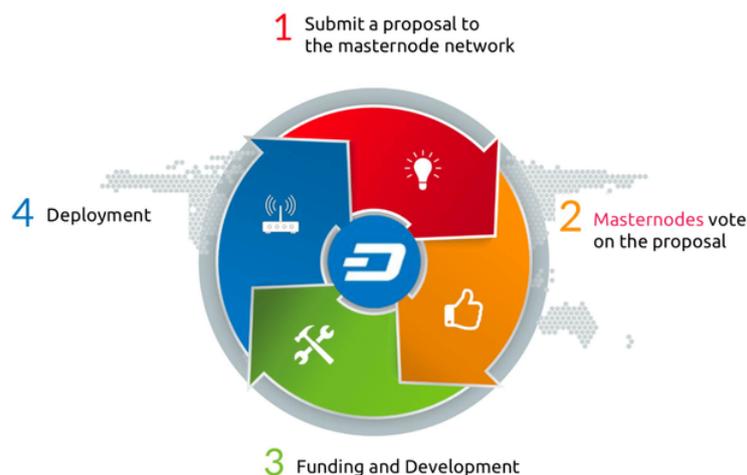


Figure 12.7: The governance mechanism adopted by Dash is very similar to the one implemented by TheDAO. The only difference is the missing role of masternodes in the voting system of TheDAO [29]

A number of issues still raise some concerns about the governance implementation in the context of DAOs [26]. Primarily, the legal status of DAOs remains unclear and is yet to be defined. A second consideration emerges from the procedural nature of voting, which can suffer from lack of voter engagement as it happened with the BitShares exchange. In this case, the vote participation was decreased by the amount of labor required to examine each proposal [26].

The controversial nature of the code immutability was already mentioned earlier. On one side, it enforces security by preventing malicious alterations. On the other side, it also exposes the running system to any sort of attack able to exploit the (existing) bugs. Unluckily, a proper bug fix would require an update and deployment of new code, plus the agreement to migrate all the funds to the new system. Chohan refers to this problem as a *dichotomy between transparency and security*: “Although the code is visible to everyone, it is difficult to repair” [26]. Last, analogously to what may happen in a democracy, the risk of voter manipulation must be taken into account. When a significant percentage, or worse majority of some type of members conspire to direct the DAO activity on their own, we speak of *collusion attacks* [25].

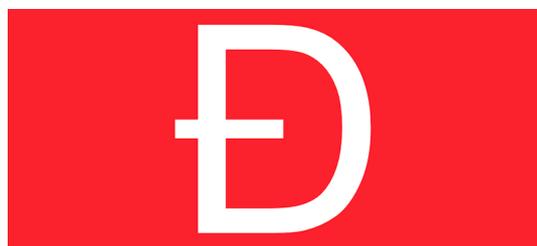


Figure 12.8: Logo of TheDAO [33]

12.4 The First Rise and Fall: TheDAO

Probably, the most famous example of a DAO is TheDAO, a crowd-based venture capital project that suffered from security flaws, which eventually led to a hack and a consequent loss of approximately 50 million US\$ [22]. In short, the hacker exploited a logical vulnerability in the code allowing the large drain of funds from TheDAO. As a response, the Ethereum community decided to rollback using a hard fork [21].

12.4.1 The Inspiration for TheDAO

The vision of the people who created TheDAO was to build a more democratic financial institution in which investors would also have had decisional power over it by means of voting [32]. The latter is conducted on the basis of *one vote per Ether*, meaning that the more substantial the contribution, the higher the vote weight [31]. This new kind of company would therefore, have been characterised by decentralized power without human fallibility. Unfortunately, the incident that later befell TheDAO proved the difficulty of completely removing the human component, even from an automatized system ruled by code [32].

12.4.2 TheDAO Creation

At the beginning of May 2016, an open source coding framework built on Ethereum blockchain, was launched under the name of TheDAO by members of the Ethereum community [30]. The venture fund was raised by selling *DAO tokens* to anyone interested in the project: 100 tokens in exchange for 1 ETH to be sent to a unique wallet address [30]. Surprisingly, during this creation period, almost 13 million ETH were collected, worth approximately 150 million US\$ at the time, but reaching over 250 million US\$ during ETH trading peaks [30].

The platform was open to anyone for casting their idea to the community in search for funds [30]. The owners of TheDAO tokens were then allowed to vote on project proposals and “democratically” decide in which project to invest part of TheDAO funds [24]. In turn, the voters would have received rewards from the possible profits made by the funded projects [30]. Being open to anyone as token holders and allowing them to not only make financial decisions, but also set any kind of rules they voted on, were some examples of how TheDAO wanted to improve the governance of today’s organizations [24].

12.4.3 The Underlying Vulnerabilities

The creators of TheDAO introduced a *split procedure* to protect the minority: by submitting a special form of proposal, the token holders who voted for the proposals that were not approved, were able to split TheDAO in two equivalent DAOs, the old one and the *child DAO*. The “losing” voters were thus able to retrieve their ETH funds by placing them on the child DAO [33]. The split procedure could be initiated anytime by anyone, but according to the hardcoded schedule one must have waited at least 48 days before having the childDAO running on an account completely under the control of the child’s creators. Among these, 27 days composed the *split creation period*, where no further proposals could have been launched on the platform [33].

As we saw in Section 12.3, bug fixing is complicated in the context of smart contracts and DAOs. This impossibility for developers to change the code whenever needed, left TheDAO participants helpless when, on June 17, 2016, a hacker started to slowly drain their funds after a loophole was discovered in the code of the split procedure [30].

12.4.4 The Attack

About 3.6 million ETH were “stolen” from TheDAO and moved to another account before the hacker himself decided to withdraw the attack for unknown reasons [30]. The attacker was following the rules implemented in the smart contracts, so technically nothing wrong was being done [24]. A *recursive call exploit* represented one of the many flaws the code had and was used by the hacker to repeatedly execute a specific smart contract, i.e. receive Ether. This was not only possible due to a neglected recursion but also because of another

logical error in the smart contract, resulting in ETH funds being sent *before* updating its internal token balance [30].

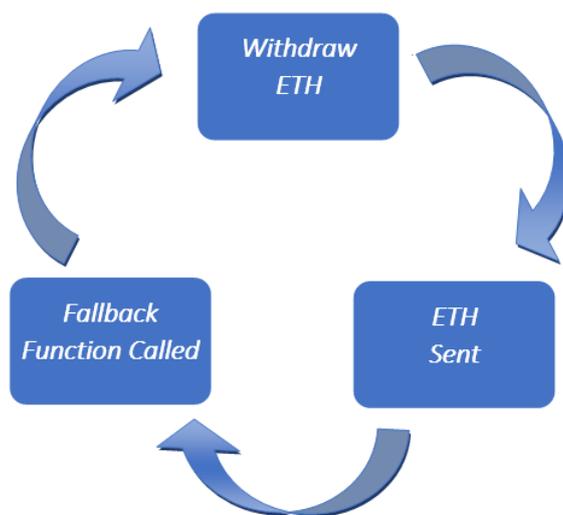


Figure 12.9: The recursive loop of a re-entrancy attack [33]

In order to get a better understanding of what happened, we can use the analogy of a situation familiar to almost everyone. A customer wants to withdraw all the money from his bank account, let's say 50 CHF. At the ATM our customer performs a first transaction requesting 50 CHF. Normally, the machine would first check if the available amount is sufficient, then update the account balance before ejecting the cash and asking if the customer wants to process another transaction. In this case if the customer should request 50 CHF again, the machine will notice that the balance is zero and interrupt the transaction ending the session. But imagine the machine would update the account balance once the entire session is ended. This would allow the customer to keep withdrawing 50 CHF until he stops the session or the machine runs out of resources [32].

This kind of attack to smart contracts falls within the category of *Re-Entrancy attacks*, mentioned in Section 12.2.2 and the basic idea is depicted in Figure 12.9. To deeply dive into the details of the exploit is worth to take a look at the simplified version of the susceptible TheDAO contract, based on [37] and reported in Listing 12.1.

The hacker starts the attack with a donation in ETH to the target contract, so that it can be executed. Then the contract updates the attacker's balance by removing the donation amount. At this point the attacker calls the function `withdraw()` to request the funds back, which are sent back without updating the contract state, i.e. the attacker's balance. This allows the latter to repeatedly call `withdraw()` by means of a fallback function⁷, `function()`, that is automatically triggered anytime the contract sends ETH back. Therefore, the attack enters an unstoppable recursive loop allowing the hacker to continuously get funds back until the smart contract executes the update to the attacker's balance, which can never be done since the smart contract is stuck calling `function()`.

```

1 contract childDAO {
2
3     /* Assign key/value pair so we can look up
4     credit integers of current users using an ETH address */
5     mapping (address => uint256) public credit;
6
7     /* A function for funds to be added to the contract,

```

⁷A fallback function is defined as a contract's function with no name that is automatically executed every time the contract receives only ETH without any data [37].

```

8     sender will be credited the amount sent */
9     function donate(address to) payable {
10        credit[msg.sender] += msg.value;
11    }
12
13    /* Show ether credited to address */
14    function assignedCredit(address) returns (uint) {
15        return credit[msg.sender];
16    }
17
18    /* Withdrawal ether from contract */
19    function withdraw(uint amount) {
20        if (credit[msg.sender] >= amount) {
21            msg.sender.call.value(amount)();
22            credit[msg.sender] -= amount;
23        }
24    }
25 }

```

Listing 12.1: Simplified version of the vulnerable DAO contract including comments for better understanding [37]

Since fallback functions are provided by Solidity, the attacked smart contract does not detect anything wrong [37]. At the end of the attack, the function `drainFunds()` is called to send funds collected by the malicious contract to the attacker's personal address [37]. An example of such a malicious contract that can be used to implement a re-entrancy attack can be found in Listing 12.2.

```

1  import 'browser/childDAO.sol';
2  contract ThisIsAHodlUp {
3
4      /* Assign childDAO contract as "dao" */
5      childDAO public dao = childDAO(0x2ae...);
6      address owner;
7
8      /* Assign contract creator as owner */
9      constructor(ThisIsAHodlUp) public {
10         owner = msg.sender;
11     }
12
13     /* Fallback function, withdraws funds from childDAO */
14     function() public {
15         dao.withdraw(dao.assignedCredit(this));
16     }
17
18     /* Send drained funds to attacker's address */
19     function drainFunds() payable public{
20         owner.transfer(address(this).balance);
21     }
22 }

```

Listing 12.2: Example of a malicious contract an attacker could use to exploit a re-entrancy attack [37]

Two main problems can be detected in the code of `childDAO`: first, as already mentioned, the contract updates the state of `credit[msg.sender]` *after* the funds have been sent to the attacker; second, using `address.call.value()` to transfer funds to the `msg.sender`, a.k.a. the attacker who is sending a message to the smart contract, is an incorrect choice [37].

Both have a rather simple solution. The first one is straightforward: since the update of the balance serves as a confirmation for a transaction to have been executed successfully, it must be done *before* actually sending the funds. The second issue can be solved by using either `address.send()` or `address.transfer()`, which both avoid multiple external calls by providing a gas fee just enough to log one event, corresponding to 2'300 gas [37]. The modifications to `childDAO` are shown in Listing 12.3.

```

1  contract childDAO{      ....   function withdraw(uint amount) {
2      if (credit[msg.sender] >= amount) {
3          credit[msg.sender] -= amount; /* Updates balance first */
4          msg.sender.send(amount()); /* Send funds properly */
5      }
6  }

```

Listing 12.3: Updating the balance before sending the amount of ETH using the proper transfer function will remove the recursive call vulnerability [37]

12.4.5 The Solution by the Community: the Ethereum Hard Fork

The unusual huge transfer from TheDAO fund was noticed by the Ethereum community, who had a quick reaction and immediately proposed multiple approaches to face the attack [30]. Additionally, an open letter was published by someone claiming to be the attacker. The letter asserted the legitimacy of the attacker’s actions being allowed by the code and warned the community about the consequences of any countermeasure they could have undertaken: “Such fork would permanently and irrevocably ruin all confidence in not only Ethereum but also in the field of smart contracts and blockchain technology. [...] Make no mistake: any fork, soft or hard, will further damage Ethereum and destroy its reputation and appeal” [36].

During the split creation period, the community discussed three options: (1) do nothing, (2) soft fork or (3) hard fork [33]. The first one respected the core philosophy of Ethereum blockchain of being trustworthy due to code immutability [33]. The second option consisted of a one-time fix as it would have declared any transaction related to the malicious child DAO invalid, thus “freezing the assets” and allowing the respective owners to retrieve their initial investment in a later moment [34]. Among many other objections, this option was discarded after discovering that the soft fork would have exposed the entire Ethereum network to Denial of Service (DoS) attacks [31]. Basically, an attacker could have flooded the network with computational expensive transactions followed by an invalid action on TheDAO’s contract, so that miners running the soft fork would have spent a lot of resources without collecting any fees [38]. See Listing 12.4 for an example of such malicious transaction.

The controversial solution adopted by TheDAO lead developers and approved by the majority of the community, consisted in bringing the blockchain to a previous state by reversing the transaction history through a hard fork. Hence, the stolen funds could have been entirely returned to the respective owners at the same rate as the initial ETH-DAO tokens offering scale [24]. In the hard-forked version of the Ethereum blockchain, the hacked funds were sent to an account available to the original owners. The refund was also possible since the hacker was following the split procedure and thus had to move the ETH funds to an account subject to the 27-days holding period. This gave the community enough time to take the necessary action to mitigate the attack [30].

```

1  for(uint32 i=0; i < 1000000; i++) {
2      sha3('some data'); // costly computation
3  }
4  DarkDAO.splitDAO(...); // render the transaction invalid

```

Listing 12.4: A malicious transaction consisting of many heavy computational operation followed by an invalid one that would cause the discard of the “expensive” contract. These kind of transactions are source of DoS [38]

12.4.6 Consequences of the Attack

In spite of being adopted as the solution, the hard fork remained contested as it would destroy the integrity and immutability of the Ethereum blockchain. Ironically, it was the characteristic of being not-absolutely immutable that allowed the recovery of the funds [33]. The hard fork was performed on July 20, 2016 [33] and, unsurprisingly, it led to a rift within the Ethereum community [24]. The participants who voted against the hard fork, continued on the old main branch that was called *Ethereum Classic*, to differentiate it from the hard-forked version which undid TheDAO and became “the actual” *Ethereum* [33]. With all the negative aspects TheDAO hack brought with it, it also had the very positive consequence of raising awareness about smart contracts and the technical challenges related to them, as they will probably gain importance in transaction execution as digital techniques evolve [32].

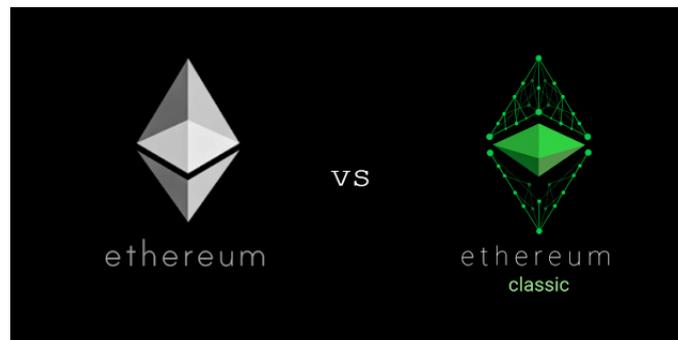


Figure 12.10: The hard fork on the Ethereum blockchain implemented to solve TheDAO hack produced the chain that “restored” *Ethereum*, while the old version continued as *Ethereum Classic* [39]

It especially emphasizes the needed enforcement of security in blockchain platforms [30] and also demonstrates that, despite TheDAO efforts, the human component will always be present: a human error made the attack possible, humans were divided by discussions on the possible solutions and then humans worked together to fix the problem [32]. Currently, the recursive call exploit of TheDAO is still used as reference for practices to avoid [30].

12.5 Current DAO Projects

In the following section, three current DAO projects will be presented.



Figure 12.11: Aragon Logo [57]

12.5.1 Aragon

Aragon is an open-source project with various components, teams and apps aimed to simplify the creation and government of DAOs. Aragon supplies several tools to DAOs which provide governance functionalities in order to organise completely decentralized and on a global scale [41]. The motivation is based on companies today being weighed down by bureaucracy and unnecessary interactions with third parties like governments and other oversight organizations. Aragon's goal is to remove intermediaries and third-parties and allow organizations to focus on their core task of creating value [42].

The core of the Aragon project is the *Aragon Client*. It is a decentralized application (DApp) that acts as a graphical user interface to create and manage DAOs as well as providing governance functionalities to their members. Functionalities such as controlling funds, executing payments to members, fundraising, voting, accounting and setting the permissions of members can be managed through the client [41]. Most functionality is provided by apps that can be installed if necessary, so that the organization is only running the functionalities it needs. APIs and a development portal are provided for creating custom apps for the platform [43].

There is a network of these DAOs that is called the *Aragon Network*. This network is functionally a DAO itself, where its participant organizations can interact and shape its future through proposals and votes. This process is handled through Aragon Governance Proposals (AGPs) [44]. The Network is governed by the Aragon Network Token (ANT). ANT is used to stimulate contribution to the network and funding the development of Aragon. ANT was created and sold in March 2017 during an ICO that raised around 25 million US\$ [45].

These funds are managed by the *Aragon Association*, a non-profit organization based in Zug. It is responsible for allocating funds to various development teams to drive the development of Aragon forward [46]. It acts as the legal entity for Aragon until the transition to a fully decentralized organization is viable. Currently, it also controls Aragon's assets, trademarks, intellectual property and repositories [47].

Aragon One is the core development team for the Aragon project and the Aragon Network. It is primarily funded through grants from the Aragon Association. It strives to function as a DAO but from a legal perspective, is currently a for-profit company also based in Zug [48].

12.5.1.1 Implementation & Governance

The Aragon Network provides a built-in governance system that allows its organizations to decide on services provided by the network and their cost. The Aragon Governance Proposal Process is outlined in AGP-1, the second AGP after AGP-0 (the Aragon Manifesto) [49]. All AGPs must follow AGP-0 and the process described in AGP-1 which incorporates the following steps [50]:

- Stage I: Select AGP Track
- Stage II: Pre-proposal
- Stage III: Draft Proposal
- Stage IV: Community Review
- Stage V: Final Proposal
- Stage VI: Aragon Network Vote

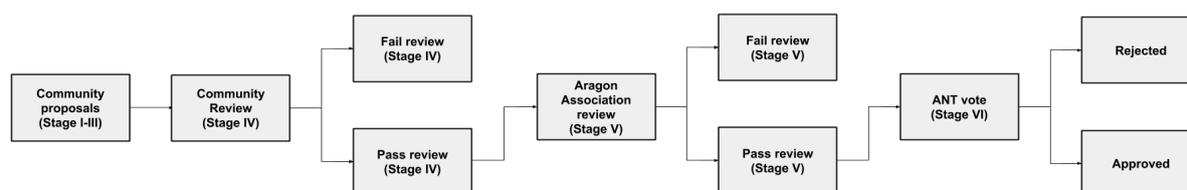


Figure 12.12: The Aragon Governance Proposal Process [50]

Each Proposal is therefore reviewed twice. First by AGP editors in Stage IV and then again by the Aragon Association Board of Directors in Stage V. It is then put up to vote in one of the quarterly Aragon Network votes. In general cases the majority of all votes cast wins. For changes that affect AGP-0 or AGP-1 a two-thirds majority is required however. Votes are token weighted, where the organizations with higher stakes in the network get more voting power [50].

AGPs are human readable documents and therefore can be subjective. To handle disputes of this kind the Aragon Network offers the *Aragon Court* service. If one of the members of the network wants to raise a dispute, he needs to pay a dispute fee as well as deposit collateral. He can also supply evidence to back up his case. The opposition can then either accept the dispute or take the case to the court. In that case they also have to pay the dispute fee and provide evidence to support their position [44].

When a case goes to the Aragon Court a jury is drafted from the other members in the network. As a reward they will receive part of the fees when the dispute is settled, but they also have to deposit some ANTs to ensure their stake in the jury. The jury will then pass a ruling based on a majority vote. The losing party can appeal the decision of the jury at which point both parties will have to increase their collateral. A second ruling will then be reached with an increased number of jurors. There is a maximum number of appeal rounds allowed, at which point the last ruling will be considered final. When a final decision is reached, the fees and collateral are distributed among the jurors and the winner of the case [44].

12.5.1.2 Outlook

The latest update for Aragon was released in September [51]. The one before that in April [52]. The project is therefore committed to continuous development and frequent releases as well as status updates. The Aragon project's GitHub repositories are also for the most part reasonably active. As for the matter of funding, the proceeds from the ANT sale should be able to fund development projects for a few years in the future. Aragon's development road-map is published on their website [53]. It includes new apps for the client like rewards, payroll and budgeting, as well as tools and frameworks to make it easier for developers to contribute and create their own applications. On the Aragon Network side, most efforts will be put into tweaking and optimising the proposal and court systems, as well as some new ideas like a fundraising platform to discover and invest in Aragon organizations.

12.5.1.3 Threats and Weaknesses

Aragon has a couple of weaknesses to overcome before it can be truly considered a DAO. Looking at the governance system of the Aragon Network and Aragon's organizational structure for instance, there are several points where the vision of a truly decentralized and unstoppable organization falls apart.

First consider the AGP process. AGPs can be shut down twice before ever reaching a vote on the network. First by selected AGP editors, which according to AGP-1 (the AGP Process) [50] are currently Luke Duncan and John Light. They are both high ranking employees at Aragon One and therefore report directly to the two co-founders Luis Cuende (CEO, Aragon One) and Jorge Izquierdo (CTO, Aragon One). The co-founders are also on the Aragon Association's board of directors, the second body that can stop AGPs before being voted on.

Luis is also the official author of AGP-0 (the Aragon Manifesto) [49], that all projects and efforts for the Aragon project follow. Furthermore, the Aragon Association is the legal holder of all assets and is the deciding body over funding for development projects.

There was a recent thread on the Aragon forum that weighs in on the problem of too much governing control by the co-founders [54]. Examples that showed the governance system does not work in a decentralized way were put forward and projects were shut down because they did not align with the views of the board. It is currently really hard to create a truly decentralized organization, both technically and legally. However there is justifiable reason to doubt the efforts being put into making this a reality by the Aragon executives.

Neither can the conflict of interest, of being on the board of directors of the non-profit association and heading the for-profit core development team, be easily overseen. For a project that writes "We are committed to decentralizing power in order to dismantle unjustified power - which usually springs from centralization" [49] in their manifesto, the founders seem to be having a hard time giving up their executive power.

Another threat to the decentralization of the network is the associated token. The Aragon Association currently holds substantial amounts of ANTs compared to the other stakeholders. The association therefore has significant influence over the network and could theoretically swing most votes by participating. The Aragon Association says it is committed to never use their ANT reserves for governance purposes, there is however no legal binding to combat such a scenario [55].

As to potential security threats there are two main possibilities. First, the whole system is based on smart contracts. If these have a flawed implementation anything from the governance system to the financial system could be compromised. Second, is the reliance on Ethereum. Being built entirely on smart contracts running on the EVM, Aragon could not survive a collapse of Ethereum.



Figure 12.13: MakerDAO logo [72]

12.5.2 Dai Stablecoin - MakerDAO

Traditional cryptocurrencies have shown high rates of volatility in the past. This can especially be seen well with the rapid rise and fall of the value of Bitcoin (BTC) and ETH in late 2017 [65]. By observing this, it can be concluded that it is unimaginable for traditional cryptocurrencies to replace fiat money anytime soon, considering that the price of commodities could fluctuate by more than 50% within just one month [66].

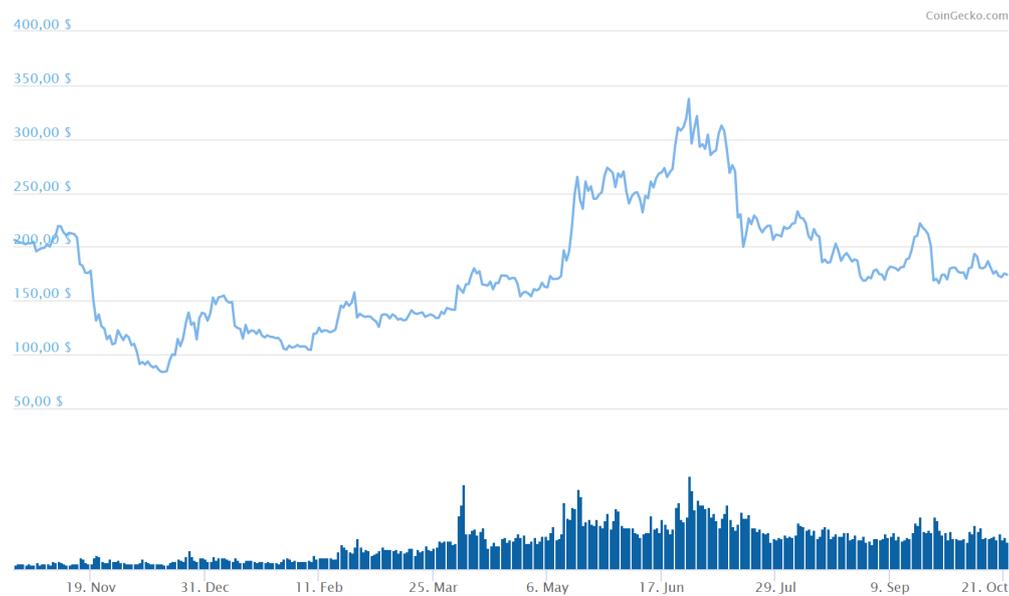


Figure 12.14: Fluctuation of the ETH price, data over the last year [73]

As it can be seen from Figure 2.2, although the fluctuations are not as drastic as they used to be, the ETH price was still highly volatile over the last year. For this reason, the idea of a stable coin came along. A stable coin is a non-volatile cryptocurrency relative to a fiat asset like the US Dollar or gold. If it is possible to redeem a stable coin for currencies or assets, it is said to be backed [67].

MakerDAO’s Dai is an attempt to create a stable coin. That means, that other than traditional cryptocurrencies, the Dai Stablecoin was designed to have a low volatility by trying to fix it to the US Dollar [1]. The problem with most traditional stablecoins is that they are not really decentralized, which goes against the basic principles of blockchains and cryptocurrencies. In most cases, this means that stablecoins are fiat collateralized, which in turn centralizes the organization again by having a centralized fund to collateralize the coins [67].

The Dai stablecoin takes a different approach: instead of taking fiat collateralization, it is collateralized with another cryptocurrency, namely ETH. This allows the collateralization to be organized in a decentralized way. Since ETH is historically highly volatile, it was decided to ask for a collateralization of at least 150% of the value of the Dai produced. This means, that whenever a user wants to take out Dai worth 100 US\$, he has to collateralize this loan with ETH worth at least 150 US\$. This is achieved by creating Wrapped ETH (WETH) out of normal ETH tokens, which can then be exchanged for Pooled ETH (PETH). This effectively locks in the user’s ETH in a smart contract and gives back a receipt in the form of PETH which “can be viewed as a percentage share of the total pool of collateral assets” [69]. The collateral inside the system, in this case WETH, is called Collateralized Debt Position (CDP) [69].

If the ETH collateralization for the Dai falls below the aforementioned 150%, the user will be punished with a mechanism that will not be introduced as it exceeds the scope of this paper [1]. By doing this, MakerDAO tries to isolate the volatility of ETH on the Dai.

12.5.2.1 Implementation & Governance

MakerDAO was launched on December 17, 2017 on the Ethereum mainnet and has gained an incredible amount of success since then [63]. For the governance aspect of MakerDAO, however, it is important to understand that its system consists of two separate coins: Dai and MKR. The Dai has been introduced in the previous chapter. MKR on the other hand serves as the governance coin and for paying Dai’s stability fees. The stability fee is a

yearly interest that is calculated on the amount of Dai that the user takes out on his CDP. It is due as soon as the debt is repaid and can only be paid in MKR. The MKR obtained in the process gets destroyed, making MKR a deflationary currency, because no new MKR are minted. Therefore, this makes MKR more valuable in the long run, increasing the desirability of being a MKR holder. MakerDAO is the biggest holder of MKR tokens and partially pays its employees with MKR [59].

The MakerDAO team describes the exact governance process in their whitepaper: firstly, every Ethereum node can create a proposal in a smart contract. MKR holders can then choose which one of all the proposals they want to have as the active proposal by casting votes with MKR tokens. These tokens will then be staked on top of that proposal until they are withdrawn. The smart contract inside the active proposal will then be executed [1]. This is a continuous process: as soon as a proposal has more MKR staked on it, the previously elected one will itself become the active proposal and be executed. In practice, however, the decision of which contract should be activated is taken outside this mechanism, namely through soft votings (which do not have an immediate impact on the actual implementation) or through discussion in forums [75].

Another aspect of the decentralized mechanism of MakerDAO is the use of oracles. The MKR holders can choose a set of oracle nodes which will then be in charge of checking the market value of Dai compared to US Dollar as well as checking the value of the collaterals [1]. MKR holders also have to choose a set of emergency oracles which have the power to “unilaterally shut down the system”, in case they suspect something bad going on in the MakerDAO network. If even the emergency oracles are compromised, the MKR holders can shut down the system themselves. In case of a shutdown, all Dai holders shall receive the equivalent of collateral for the Dai they had [1].

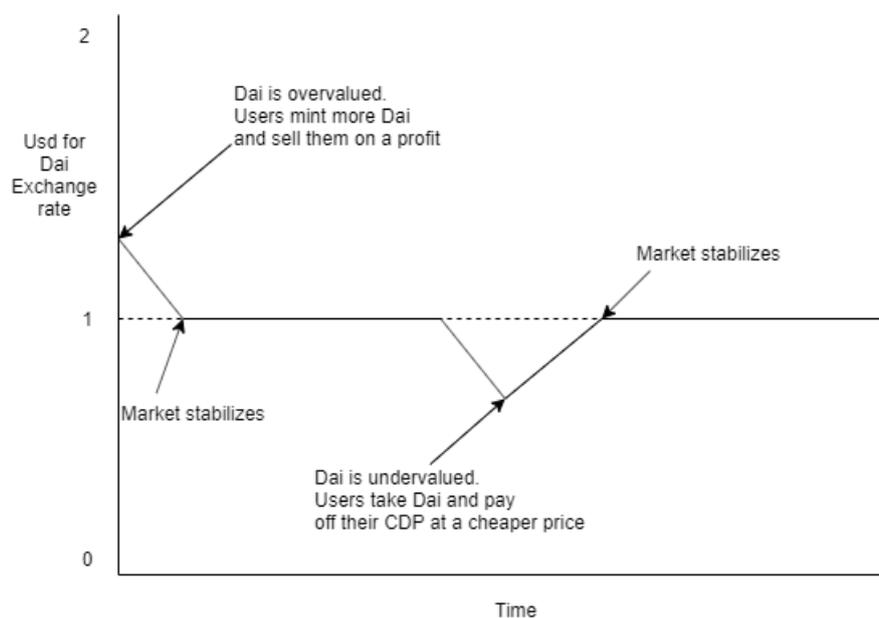


Figure 12.15: Integrated mechanism to stabilize the Dai, own graphic with information from medium.io [63]

The core element of MakerDAO however, is its mechanism to fix the Dai to the US Dollar. Since there is no central element doing the job, the process works fully decentralized. This is, at least in theory, achieved by exploiting simple game theory mechanisms, and visualized in Figure 12.15: in case the Dai is overvalued, meaning that it is worth more than the US Dollar, people are incentivized to create more Dai and then sell it off for a profit, therefore flooding the market and reducing the price of Dai. Conversely, if the Dai is undervalued, people with a CDP can pay back their debt for a cheaper price [63]. However, this view has been challenged by the blockchain experts Su Zhu and Hasu: they

argue that, while an undervalued Dai is easily fixable by just buying more Dai with a potential profit later, or paying back the debt of the CDP, thus creating more demand or less supply, an overvalued Dai is a bigger problem which is not solvable within the decentralized arbitrage system of Dai. This is due to the forced overcollateralization and therefore the inefficient use of capital, and the fees that have to be paid when burning Dai again which overshadows the positive aspect of potential profits from lending out Dai [61]. Another mechanism that MakerDAO can use is the adjustment of the stability fee. The higher the stability fee is, the less attractive it is to have a CDP and take out Dai. Lowering the stability fee means that taking out Dai becomes more attractive again.

12.5.2.2 Purpose

The true purpose of the Dai may not be obvious at first, and even after extensive research seems hard to grasp. In fact, the whitepaper lists a multitude of markets to be penetrated and features to be leveraged. First and foremost, the Dai is a stable cryptocurrency, making it a real option for (anonymous) business as it is easy and cost-efficient to send Dai around, or as a low-risk investment for holding currency due to its stability. Also, by creating CDPs, users can leverage the value of their ETH tokens without having to sell them. On the same note, the mechanism to mint coins along with creating CDPs makes MakerDAO a platform which can be used for margin trading too [1]. This is useful when the ETH holder speculates that the ETH might rise in price later. With this practice, the borrowed Dai can even be used to buy more ETH, increasing potential profits, but also increasing the risk on the user's assets.

Consider Figure 12.16: the user has a fixed amount of ETH, colored in orange, symbolizing that it is a liquid asset, but the user does not want to bargain it away, and thinks that its price will rise in the future. He can use his existing ETH and turn it into a CDP, marked red here meaning it is not a fluid asset. Using the CDP, the user can take out a certain amount of Dai, green meaning it is a fluid asset, which can then be used to buy more ETH. By doing this, the user can increase his exposure to the market without investing more money, i.e. he can leverage his invested capital. This increased exposure results in higher profits or higher losses respectively, depending on the actual market developments [63].

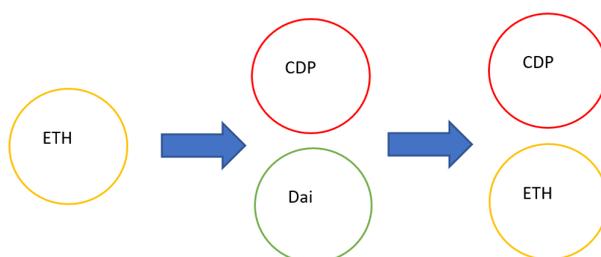


Figure 12.16: The logic behind margin trading, own graphic with information from Medium.io [63]

Its structure as a decentralized system prohibits Dai from becoming as scalable as stablecoins that are backed with fiat money, as Su Zhu and Hasu point out: small market incentives to create a CDP and take out a loan are not enough and connected with too much uncertainty [61].

To conclude this section, it can be seen that the Dai might have many viable use cases, but they all build up on the assumption that one of those use cases will be used extensively enough to keep the demand of Dai in line with the supply. This specific use case is MakerDAO as a lending platform. If no one decides to take out loans, which have to be overcollateralized and paid interest for, all the other mentioned use cases cannot be

leveraged. In the end, it will be interesting to see if it can succeed or if it tries to be too many things at once, offering an overall package that just is not attractive and practical enough for the modern market.

12.5.2.3 Outlook

For the near future, MakerDAO has some big projects due. They can be checked on their official website [1]. By the end of this semester, it is planned to accept more means of collateralization than only ETH. MKR holders will be able to vote as to what these means of collateralization will be [1]. This concept of multicollateralization will have a profound effect on the inner workings and mechanisms of MakerDAO which will not be discussed further in this paper. In short, the project will become even more complex and difficult to understand.

12.5.2.4 Threats and Weaknesses

Looking at MakerDAO in its current state, there are several weaknesses, most of which depend on the use case that Dai should fulfill in the end [62]. For example, on top of the critics by Su Zhu and Hasu about the lacking ability of scalability, MakerDAO got criticized for having a too complicated system [62]. However, the inner working of MakerDAO is not a thing that the everyday user of the Dai needs to be aware of, as Seibel points out in his article for Medium: in an ideal world, the Dai will just be a secure, cheap and fast way to make national and international payments over the Internet, on the foundation of a stable currency [63].

Arguably the biggest one of the weaknesses, which does not depend on the use case of Dai, is the reliance on Ethereum. As Preston Byrne points out, the system will completely collapse if Ethereum collapses first [68]. While being unlikely, the highly volatile market of cryptocurrencies has shown us in the past that such a case is not completely impossible. Furthermore, he reasons that a falling Ethereum course will also doom the Dai as no one wants to lock up their currency that is losing value [68].

Still, there are more weaknesses than just the ones on the economic side. For example, the supply of MKR is highly concentrated on a small amount of holders [70] which might enable a malicious takeover. Even though the biggest holder of MKR, which is MakerDAO itself, committed not to participate in governance affairs [62], it has even been reported that one single MKR holder managed to change the stability fee because most MKR holders do not participate in governance [71]. Also, the development is handled by MakerDAO and mistakes in code might end up in bad situations [62], as we have seen with TheDao in Section 12.4.



Figure 12.17: Moloch DAO logo [79]

12.5.3 Moloch DAO

Moloch DAO was primarily developed to tackle the problems, that have been occurring within the development of Ethereum 2.0, such as funding and incentives. These issues occur, since when developing a project that could positively impact a community, only a few people have to carry those costs. Therefore, their direct gain is significantly lower compared to their commitment, as the total benefit of the project is distributed between every stakeholder of it [76].

Moloch DAO tries to overcome this disparity by pooling funds into a guild bank and letting the members of the guild vote on how these funds should be allocated [76]. This way parties with similar interest can be gathered, which increases the capital extending the organisation's possibilities.

Moloch DAO has been online since February 14th, 2019 [77] and has been able to attract the interest of a lot people. As of October 27th, 2019, the DAO has been able to [78]:

- expand their guild to 71 members (49 with more than 100 shares, which are called elders)
- achieve a guild bank value of more than a million dollars
- complete 102 proposals

Additionally, as of August 16th, Moloch DAO has contributed to 17 projects with over 120'000 US\$ [79].

12.5.3.1 Implementation & Governance

As seen in the overview, in order to vote on and submit new proposals an interested party has to join the Moloch first. In order to avoid proposal spamming, this step cannot be executed from outside the guild, but rather has to be done by asking an existing member to submit a proposal. After finding and convincing a member, there are two ways to enter the guild: the first one is through a membership proposal, which includes a tribute of WETH (currency used in Moloch DAO - exchangeable 1:1 using ETH) that contributes to the Guild Bank [77]. The second one is called a grant proposal which instead of a tribute consists of contributing by doing some work. In both cases the applicant defines the amount of non-transferable voting shares wanted in return. Next, the guild member submits the proposal by depositing 10 WETH, of which 0.1 WETH are used as incentive for the process; the remaining 9.9 WETH are returned, once the process is over. With this the Voting Period has started. However, as a protection mechanism, the applicant is able to abort the proposal (within the first day), if, for example, the member requests fewer shares than expected. If the proposal has been accepted, the applicant receives its shares from the Guild Bank [76].

Looking at the core element of Moloch DAO, voting on proposals, one can see that each proposal lasts for 7 days, and per day a maximum of 5 proposals can be submitted. In order to submit a proposal the member has to deposit 10 WETH. Furthermore, each member can only vote once using all his shares, there is no minimum amount of votes required and whether a proposal fails or passes is decided by the majority that voted on it. After the voting period, the votes are counted and a seven-day long grace period starts. During this period, to avoid bearing the costs of the proposal, members that voted "No" on the proposal can ragequit [76]. Additionally a voter cannot ragequit, if the last proposal he voted "Yes" on still has not been processed [82]. By ragequitting the disagreeing member liquidates its shares and leaves the guild (making him unaffected by other proposals he voted on as well). As a final step, a function to process the proposal has to be called rewarding the member calling it with 0.1 WETH [76].

To prevent ragequitting from being used as a way to harm other members a dilution bound has been specified. This mechanism makes a proposal automatically fail if more than two thirds of the members were ragequitting at the same time. This would then only result in a reduction of the guild bank's balance [76].

Aside from becoming a guild member, one can also contribute to the Moloch DAO by donating WETH to the Moloch Pool [79], whose current value lies at approximately 600 US\$ [78]. Even though this pool also contains shares, they are non-voting. Additionally, the pool's funds are given in the same relative amount to a grant beneficiary as the guild bank donates [83].

12.5.3.2 Threats and Weaknesses

Moloch DAO has a weakness in dealing with free riders. As the DAO is mainly contributing to the Ethereum community, every Ethereum stakeholder will benefit from the DAO irrespective of their contribution [76].

A further flaw lies within the non-existing upgrade-mechanism of Moloch DAO. As a result, the only way of upgrading the contract consists of every member ragequitting and then putting the money into the new contract.

Even though this case should be covered by the abort period, an applicant's inability of submitting a proposal can be seen as a risk, with the worst-case scenario being, that the applicant is donating his WETH, since a guild member can simply set the shares requested to 0 [76].

Another weakness lies in the reliance on Ethereum. This does not only make the guild bank value volatile, but an end to Ethereum would result in the balance being worthless. During a conference in Berlin, Ameen Soleimani, the summoner of Moloch DAO, mentioned some more flaws. One of them being the fact, that it is not possible to kick someone out of the guild. As a result the guild members are more reluctant to let applicants enter the guild and if a member turns out to be damaging to the DAO there is no way of removing him.

Additionally, the lack of possibility to bundle proposals that are connected to each other has been mentioned. This could cause a problem if more than one person is working on a project, since two different proposals would have to be submitted. With both of them having the possibility to fail and thus having the chance to pass or fail independently.

Furthermore, the problem of vote delegation has been mentioned. The joining of the Ethereum Foundation was used as an example, where 1000 ETH were shared between 10 members. As a result of this missing feature, the entry process was a lot more complicated than it could have been.

Even though unlikely, as there is no limitation to how many proposals a single member can submit (besides the maximum number of proposals being 5 per day), a guild member could, at a rather unprofitable cost, decide to submit the maximum amount of proposals per day making it impossible for any other proposals to be submitted [80].

12.5.3.3 Outlook

According to Ameen Soleimani a new version of Moloch DAO will only fix broken things rather than the contract being implemented completely anew. This second version of Moloch DAO, to fix some of the mentioned issues, is already in development. However, how, when and if the new contract is being deployed is still unknown [81].

12.6 Comparison

While all of the three projects presented above serve different purposes, they were intended to be built and operated as DAOs. In what follows, differences and similarities in their goals and implementation will be discussed, along with a final conclusion at the end of the paper.

One major similarity between the three described DAOs is the availability of a voting system. However, whereas it is Moloch DAO's core function, MakerDAO and the Aragon Network use voting in order to decide in which direction the respective DAO should be going, e.g. what type of functionalities should be implemented.

When it comes to who is entitled to vote there are a few differences. Whereas within Moloch DAO every party that is a part of the respective section is entitled to vote, in MakerDAO and Aragon only holders of MKR or ANT tokens respectively, can cast a vote. Furthermore, in MakerDAO every Ethereum node can submit a proposal, whilst in the Aragon Network and Moloch DAO only the members can submit a proposal.

A difference in Aragon's voting system is that the proposals are reviewed before the network can vote on it [50]. Additionally MakerDAO allows you to cast/withdraw a vote at any given time [84], whereas in Moloch DAO and Aragon once a vote has been cast, there is no way to change this decision. Another difference between the DAO's voting systems is that in Moloch DAO and Aragon all active proposals can be voted on simultaneously. In MakerDAO however, MKR holders vote on all proposals as well by staking MKR tokens on them, but only one of them can be activated at a time. As opposed to the other projects, the proposal with the most votes becomes the active proposal and can execute its smart contract. The next active proposal will only be given control when it surpasses the total amount of MKR that are staked on the currently active proposal [75]. To prevent any harmful changes, after the activation there is a period where the implementation can still be stopped by the community [84].

When it comes to the actual business models beyond the DAO organization, the three DAOs that we previously introduced do not have much in common. Aragon, for example, is a DAO which was created to support the creation of new DAOs. They fund development from the Aragon Network Token, which holds its value as long as investors believe in their cause [45]. Only the development companies are for-profit organizations while the Aragon Association behind the actual network is non-profit [46]. Just like Aragon, MakerDAO earns money from its governance token, namely the MKR. Its supply was limited to a certain amount from the beginning [70] and some of it has to be burned whenever a CDP is paid back [1], which makes MKR a deflationary currency, increasing in value as the MakerDAO platform coin increases in popularity. The owners of MakerDAO can keep selling off their MKR tokens that they generated in the beginning, with the remaining tokens increasing in value. MolochDAO on the other hand takes a different approach. Instead of enabling businesses to create their own DAOs, they created a platform for funding Ethereum development projects. Its very nature as a funding platform for Ethereum, which is an open source project, therefore makes Moloch a non-profit organisation.

While Moloch DAO is truly governed as a DAO on and off the chain, the other two projects that have been shown do not represent fully decentralized organizations, with Aragon having a legal entity in Zug [48] and MakerDAO having several offices in North America and Europe [74].

Both MakerDAO and Aragon currently hold majority stakes in their governance token. Although they both promised not to use those tokens to participate in the voting process, this gives them significant influence and centralizes power, which is the opposite of what DAOs hope to achieve. Aragon even goes one step further where they curate proposals before allowing a vote.

Other than the power distribution problem, the DAOs reviewed are all bound to the success of Ethereum and would not survive a collapse. Furthermore, they are built on smart contracts that, unless implemented flawlessly, open them up to exploitation from third parties.

12.7 Conclusion

Recently, blockchain projects have seen more pushes for the implementation and deployment of DAOs [86] and have gained relevance in many ethereum gatherings, such as the “Berlin Blockchain Week” [85]. However, real-world deployments of DAOs still face challenges, for instance, the dependence of certain blockchains (such as Ethereum). Additionally, it still remains very challenging to completely decentralize decision-making as seen in the examples of Aragon and MakerDAO, where a few people can have a major impact on the entire organisation [62][54]. Thus, past efforts to deploy DAOs have faced serious challenges.

Even though these challenges highlighted the importance of Smart Contract Security [30], should a vulnerability be found, an organisation would still not be able to react to a malicious attack in a timely manner [24]. This concern is also reflected in the Moloch DAO, which has a similar approach as TheDAO but has significantly less money in its funds than the latter one [30][78]. The unclear legal status should be taken into consideration as well since its implementation can have a major impact on the viability of DAOs [26]. In our opinion DAOs have the potential to disrupt the way organizations are governed in the future. At this point in time though, we do not believe that DAOs have progressed much further than 3 years ago when TheDAO incident occurred, as the same vulnerabilities still exist and have not been specifically addressed. The technology is still in its infancy and only time will tell how successful DAOs will become in the future.

Bibliography

- [1] MakerDAO; <https://makerdao.com/en/>. Last accessed October 4, 2019
- [2] Aragon; <https://aragon.org>. Last accessed October 4, 2019
- [3] Burkhard STILLER. *Overview on Blockchain Work Performed at the CSG@IfI of the UZH*. 2019.
- [4] Andreas M. ANTONOPOULOS: *Mastering Bitcoin: unlocking digital cryptocurrencies*; O'Reilly Media Inc., 2014, pp. xix, 25-30, 162-164, 221-224
- [5] *Dash School* Video Series; <https://docs.dash.org/en/stable/introduction/about.html#dash-school>. Last accessed October 18, 2019
- [6] Dash (cryptocurrency); Article, Wikipedia; [https://en.wikipedia.org/wiki/Dash_\(cryptocurrency\)](https://en.wikipedia.org/wiki/Dash_(cryptocurrency)). Last accessed December 14, 2019
- [7] Ethereum; <https://www.ethereum.org>. Last accessed October 19, 2019
- [8] *Learn about Ethereum*; <https://www.ethereum.org/learn/#how-ethereum-works>. Last accessed October 19, 2019
- [9] Coindesk, *Ethereum Price*; <https://www.coindesk.com/price/ethereum>. Last accessed October 19, 2019
- [10] District0x, *What Is Ethereum?*; <https://education.district0x.io/general-topics/understanding-ethereum/what-is-ethereum/>. Last accessed October 19, 2019
- [11] Linda XIE: *A beginner's guide to Ethereum*; Article, Medium, February 23, 2017; <https://blog.coinbase.com/a-beginners-guide-to-ethereum-46dd486ceecf>. Last accessed October 19, 2019
- [12] *A gentle introduction to Ethereum*; October 2, 2016; <https://bitsonblocks.net/2016/10/02/gentle-introduction-ethereum/>. Last accessed October 19, 2019
- [13] Preethi KASIREDDY: *How does Ethereum work, anyway?*; Article, Medium, September 27, 2017; <https://medium.com/@preethikasireddy/how-does-ethereum-work-anyway-22d1df506369>. Last accessed October 19, 2019
- [14] Richard RED: *Soft forks, hard forks, chain splits and free coins! In A user's perspective and introduction to blockchain governance*; Article, Medium, April 11, 2018; <https://medium.com/@richardred/a-users-perspective-and-introduction-to-blockchain-governance-80fbe2050222>. Last accessed October 5, 2019

- [15] Vitalik BUTERIN: *Hard Forks, Soft Forks, Defaults and Coercion*; March 14, 2017; https://vitalik.ca/general/2017/03/14/forks_and_markets.html. Last accessed October 20, 2019
- [16] *Ethereum Platform - What You Should Know*; Article, Medium, July 6, 2018; <https://cryptodigestnews.com/ethereum-platform-what-you-should-know-900ff0b5b5ec>. Last accessed October 31, 2019
- [17] Nick SZABO: *Smart Contracts: Building Blocks for Digital Markets*; EXTROPY: The Journal of Transhumanist Thought, (16), 1996, 18: 2
- [18] *A gentle introduction to smart contracts*; February 1, 2016; <https://bitsonblocks.net/2016/02/01/gentle-introduction-smart-contracts/>. Last accessed October 21, 2019
- [19] *The Ultimate Guide to Understanding Smart Contracts*; <https://www.blockchaintechnologies.com/smart-contracts/>. Last accessed October 21, 2019
- [20] *Smart Contracts and Solidity*; <https://github.com/ethereumbook/ethereumbook/blob/develop/07smart-contracts-solidity.asciidoc#what-is-a-smart-contract>. Last accessed October 21, 2019
- [21] Daniel PEREZ, Benjamin LIVSHITS: *Smart Contract Vulnerabilities: Does Anyone Care?*; Article, arXiv preprint arXiv:1902.06710, May 17, 2019.
- [22] Alex NORTA: Designing a smart-contract application layer for transacting decentralized autonomous organizations. In *International Conference on Advances in Computing and Data Sciences*; Springer, Singapore, 2016, p. 595-604.
- [23] *Bedrijven worden overgenomen door Smart Contracts ? 7 voorbeelden*; August 23, 2018; <https://coingids.nl/smart-contracts/smart-contract-uitleg-bedrijven/>. Last accessed November 1, 2019
- [24] Coindesk, *What is a DAO?*; <https://www.coindesk.com/information/what-is-a-dao-ethereum>. Last accessed October 4, 2019
- [25] Vitalik BUTERIN: *DAOs, DACs, DAs and More: An Incomplete Terminology Guide*; Post, Ethereum Blog, May 6, 2014; <https://blog.ethereum.org/2014/05/06/daos-dacs-das-and-more-an-incomplete-terminology-guide/>. Last accessed October 5, 2019
- [26] Usman W. CHOCHAN: *The decentralized autonomous organization and governance issues*; Article, 2017; Available at SSRN 3082055
- [27] Cambridge University Press: *Cambridge Dictionary; What is a DAO?*; 2019; <https://dictionary.cambridge.org/dictionary/english/governance>. Last accessed October 25, 2019
- [28] Dash Core Group: *Understanding Dash Governance*; 2018; <https://docs.dash.org/en/stable/governance/understanding.html>. Last accessed October 25, 2019
- [29] *Investing in DASH - Don't miss the Train on this one*; <https://steemit.com/dash/@kondor1030/investing-in-dash-don-t-miss-the-train-on-this-one>. Last accessed November 5, 2019

- [30] Samuel FALKON: *The Story of the DAO - It's History and Consequences*; Article, Medium, December 24, 2017; <https://medium.com/swlh/the-story-of-the-dao-its-history-and-consequences-71e6a8a551ee>. Last accessed October 4, 2019
- [31] Richard RED: Ethereum and the DAO hard fork. In *A user's perspective and introduction to blockchain governance*; Article, Medium, April 11, 2018; <https://medium.com/@richardred/a-users-perspective-and-introduction-to-blockchain-governance-80fbe2050222>. Last accessed October 5, 2019
- [32] Klint FINLEY: *A \$50 Million Hack Just Showed That the DAO Was All Too Human*; Article, Wired, June 18, 2016; <https://www.wired.com/2016/06/50-million-hack-just-showed-dao-human/>. Last accessed October 5, 2019
- [33] Osman Gazi GÜÇLÜTÜRK: *The DAO Hack Explained: Unfortunate Take-off of Smart Contracts*; Article, Medium, August 1, 2018; <https://medium.com/@ogucluturk/the-dao-hack-explained-unfortunate-take-off-of-smart-contracts-2bd8c8db3562>. Last accessed October 5, 2019
- [34] David SIEGEL: *Understanding The DAO Attack*; Article, Coindesk, June 25, 2016; <https://www.coindesk.com/understanding-dao-hack-journalists>. Last accessed October 5, 2019
- [35] Quinn DUPONT: Experiments in algorithmic governance: A history and ethnography of The DAO, a failed decentralized autonomous organization. In *Bitcoin and Beyond (Open Access)*. Routledge, 2017. p. 157-177.
- [36] *An open letter*; June 18, 2016; <https://pastebin.com/CcGUBgDG>. Last accessed October 28, 2019
- [37] Pete HUMISTON: *Smart Contract Attacks [Part 1] - 3 Attacks We Should All Learn From The DAO*; Article, Hackernoon, May 13, 2018; <https://hackernoon.com/smart-contract-attacks-part-1-3-attacks-we-should-all-learn-from-the-dao-909ae44>. Last accessed October 5, 2019
- [38] Tjaden HESS, River KEEFER, Gün SIRER: *Ethereum's DAO Wars Soft Fork is a Potential DoS Vector*; Article, Hacking, Distributed, June 28, 2016; <http://hackingdistributed.com/2016/06/28/ethereum-soft-fork-dos-vector/>. Last accessed October 29, 2019
- [39] *Noob's Guide to Ethereum vs Ethereum Classic*; <https://www.numoney.my/my/post/noob-guide-to-eth-vs-etc>. Last accessed November 2, 2019
- [40] *Decentralized autonomous organization (DAOstacks)*; <https://steemit.com/crypto/@steemclown/decentralized-autonomous-organization-daostacks>. Last accessed November 2, 2019
- [41] Aragon Wiki; <https://wiki.aragon.org>. Last accessed October 26, 2019
- [42] Luis CUENDE: *Introducing Aragon: Unstoppable companies*; Article, February 10, 2017; <https://blog.aragon.org/introducing-aragon-unstoppable-companies-58c1fd2d00ce/>. Last accessed October 26, 2019

- [43] Aragon Developer Portal; <https://hack.aragon.org>. Last accessed October 26, 2019
- [44] *Aragon Network*; Whitepaper, May 10, 2019; <https://github.com/aragon/whitepaper>. Last accessed October 25, 2019
- [45] *Aragon Network Token*; https://wiki.aragon.org/network/aragon_network_token/. Last accessed October 26, 2019
- [46] *About the Aragon Association*; <https://wiki.aragon.org/association/overview/>. Last accessed October 26, 2019
- [47] Luis CUENDE: *Decentralizing Aragon's development II: Minimum Viable Foundation*; Article, May 2, 2018; <https://blog.aragon.org/decentralizing-aragons-development-ii-minimum-viable-foundation-8ec1f9a13ebc/>. Last accessed October 26, 2019
- [48] Luis CUENDE: *Introducing Aragon One*; Article, May 16, 2018; <https://blog.aragon.org/introducing-aragon-one-b14dd804c5ce/>. Last accessed October 26, 2019
- [49] Luis CUENDE: *AGP-0: The Aragon Manifesto*; Article, May 8, 2018; <https://github.com/aragon/AGPs/blob/master/AGPs/AGP-0.md>. Last accessed October 27, 2019
- [50] John LIGHT: *AGP-1: The Aragon Governance Proposal Process*; Article, October 12, 2018; <https://github.com/aragon/AGPs/blob/master/AGPs/AGP-1.md>. Last accessed October 27, 2019
- [51] Luis CUENDE: *Introducing Aragon 0.8 Camino*; Article, September 9, 2019; <https://blog.aragon.org/aragon-0-8-camino/>. Last accessed October 28, 2019
- [52] Luis CUENDE: *Aragon 0.7 Bella is here*; Article, April 17, 2019; <https://blog.aragon.org/aragon-0-7-bella-is-here/>. Last accessed October 28, 2019
- [53] Aragon Roadmap; <https://aragon.org/project/roadmap/>. Last accessed October 28, 2019
- [54] *Is Aragon falling into a Centralized Culture Framework?*; <https://forum.aragon.org/t/is-aragon-falling-into-a-centralized-culture-framework/1349>. Last accessed October 28, 2019
- [55] *Aragon Association ANT Policy*; https://wiki.aragon.org/association/association_ant/. Last accessed October 28, 2019
- [56] Etherscan ANT Token; <https://etherscan.io/token/0x960b236A07cf122663c4303350609A66A7B288C0#balances>. Last accessed November 5, 2019
- [57] Aragon logo; <https://wiki.aragon.org/design/logo/>. Last accessed November 13, 2019
- [58] Aragon Court Process Diagram; <https://blog.aragon.org/aragon-network-jurisdiction-part-1-decentralized-court-c8ab2a675e82/>. Last accessed November 13, 2019

- [59] Ben MUNSTER: *How to loan yourself money with MakerDAO*; Article, Decrypt, Feb 25, 2019; <https://decrypt.co/5344/be-your-own-bank-with-makerdao>. Last accessed October 16, 2019
- [60] Marc-Andre DUMAS: *MakerDAO Q1 2019 Revenue Analysis*; Article, Medium, Apr 9, 2019; <https://medium.com/@marcandrudas/makerdao-q1-2019-revenue-analysis-9afe82af3372>. Last accessed October 16, 2019
- [61] HASU, Su ZHU: *Maker Dai: Stable, but not scalable*; Article, Medium, Jan 8, 2019; <https://medium.com/@hasufly/maker-dai-stable-but-not-scalable-3107ba730484>. Last accessed October 16, 2019
- [62] Matteo LEIBOWITZ: *Addressing popular MakerDAO criticisms*; Article, The Block, Sept 12, 2019; <https://www.theblockcrypto.com/post/39595/addressing-preston-byrnes-makerdao-criticisms>. Last accessed October 16, 2019
- [63] James SEIBEL: *The Dai Stablecoin is a Game Changer for Ethereum and the Entire Cryptocurrency Ecosystem*; Article, Medium, Apr 12, 2018; https://medium.com/@james_3093/the-dai-stablecoin-is-a-game-changer-for-ethereum-and-the-entire-cryptocurrency- Last accessed October 16, 2019
- [64] Jack PURDY: *Maker (MKR) Investment Thesis*; Article, Medium, Dec 18, 2018; <https://medium.com/coinmonks/cryptoasset-research-maker-mkr-a0e89fccb985>. Last accessed October 17, 2019
- [65] Coin Desk: *Ethereum Price (ETH)*; <https://www.coindesk.com/price/ethereum>. Last accessed October 17, 2019
- [66] Fintech News Switzerland: *The 5 Biggest All-Time Monthly Losses in Bitcoin and Ethereum*; Article, Fintech News Switzerland, Dec 18, 2018; https://fintechnews.ch/blockchain_bitcoin/bitcoin-and-ethereum-all-time-biggest-losses/24430/. Last accessed October 17, 2019
- [67] Bilal MEMON: *Guide to stablecoin: Types of stablecoins & its importance*; Article, Master The Crypto; <https://masterthecrypto.com/guide-to-stablecoin-types-of-stablecoins/>. Last accessed October 17, 2019
- [68] Preston BYRNE: *Stablecoins are doomed to fail*; Article, prestonbyrne.com, Dec 17, 2017; <https://prestonbyrne.com/2017/12/10/stablecoins-are-doomed-to-fail/>. Last accessed October 17, 2019
- [69] MakerDAO: *What is PETH*; <https://cdp.makerdao.com/help/what-is-peth>. Last accessed October 21, 2019
- [70] Etherscan MKR Token; <https://etherscan.io/token/0x9f8f72aa9304c8b593d555f12ef6589cc3a579a2#balances>. Last accessed November 3, 2019

- [71] Patrick CLEATH: *One Maker voter changed the interest rate for DAI*; Article, cryptocult, Oct 29, 2019; <https://cryptocult.co/2019/10/29/one-maker-voter-changed-the-interest-rate-for-dai/>. Last accessed November 3, 2019
- [72] MakerDAO logo; <https://blockchainwelt.de/wp-content/uploads/2019/03/makerdao-dai-stablecoin-logo.jpg>. Last accessed November 11, 2019
- [73] Value of Ethereum; <https://www.coingecko.com/de/munze/ethereum>. Last accessed November 12, 2019
- [74] MakerDAO Company Overview; <https://craft.co/makerdao>. Last accessed November 12, 2019
- [75] MakerDAO Governance: Core Foundation principles; <https://community-development.makerdao.com/governance/core-principles>. Last accessed December 11, 2019
- [76] The Moloch DAO, Whitepaper; <https://github.com/MolochVentures/Whitepaper/blob/master/Whitepaper.pdf>. Last accessed October 27, 2019
- [77] Moloch Summoning Guide; <https://medium.com/molochdao/moloch-summoning-guide-12a2a288e0ff>. Last accessed October 27, 2019
- [78] Moloch DAO; <https://molochdao.com/>. Last accessed October 27, 2019
- [79] @MolochDAO; <https://twitter.com/MolochDAO>. Last accessed October 27, 2019
- [80] DAPPCON 2019: MolochDAO has risen, now what? - Ameen Soleimani (MolochDAO); <https://www.youtube.com/watch?v=10RnoWdhja8>. Last accessed October 27, 2019
- [81] Moloch v2 Draft Code Walkthrough; <https://www.youtube.com/watch?v=A6MWCBFN1Kw>. Last accessed October 31, 2019
- [82] GitHub, Moloch.sol; <https://github.com/MolochVentures/moloch/blob/master/contracts/Moloch.sol>. Last accessed November 3, 2019
- [83] William M. Peaster: *MolochDAO Looks Back on Its Rising Role in Ethereum Ecosystem*; <https://blockonomi.com/molochdao-rising-role-ethereum/>. Last accessed November 5, 2019
- [84] What is MKR; <https://medium.com/makerdao/what-is-mkr-e6915d5ca1b3>. Last accessed November 3, 2019
- [85] Christine Kim, *New Interest in DAOs Prompts Old Question: Are They Legal?*; <https://www.coindesk.com/new-interest-in-daos-prompts-old-question-are-they-legal>. Last accessed November 5, 2019
- [86] Eric Gorski, *2019 is the Year of the DAO*; <https://blog.gnosis.pm/2019-is-the-year-of-the-dao-5a428f90fb55>. Last accessed November 12, 2019

